ITU KALEIDOSCOPE 2016

ICTs for a Sustainable World

The 8th ITU Kaleidoscope academic conference

Bangkok, Thailand, 14-16 November 2016

Technically co-sponsored by:

IEEE

IEEE ComSoc IEEE Communications Society

Co-located with:

ITU TELECOM WORLD '16

Bangkok 14–17 November

In partnership with:

ROYAL HOLLOWAY UNIVERSITY OF LONDON

UNESCO United Nations Educational, Scientific and Cultural Organization

UNESCO Chair in ICT for Development Royal Holloway, University of London

早稲田大学 WASEDA University

I.I.E.E.J

EURAS

eic

COMSYS

RWTH AACHEN UNIVERSITY

Universidad del País Vasco Euskal Herriko Unibertsitatea

UNIVERSITÉ DE GENÈVE

1956 2016 CCITT / ITU-T

Organized by:

ITU

PROCEEDINGS OF THE 2016

ITU KALEIDOSCOPE

ACADEMIC CONFERENCE

ICTs for a Sustainable World

BANGKOK, THAILAND, 14-16 NOVEMBER 2016

Technically co-sponsored by:

IEEE

IEEE ComSoc
IEEE Communications Society

Co-located with:

ITU TELECOM WORLD '16
Bangkok 14-17 November

In partnership with:

ROYAL HOLLOWAY UNIVERSITY OF LONDON

UNESCO Chair in ICT for Development, Royal Holloway, University of London

早稲田大学
WASEDA University

IEEJ

EURAS

eic

COM SYS

RWTH AACHEN UNIVERSITY

Universidad del País Vasco
Euskal Herriko Unibertsitatea

UNIVERSITÉ DE GENÈVE

1956 2016
CCITT / ITU-T

Organized by:

ITU

IEEE Catalogue Number: CFP1668P-ART

ISBN:
978-92-61-20431-0 (paper version)
978-92-61-20441-9 (electronic version)
978-92-61-20451-8 (CD-ROM)

Disclaimer

The opinions expressed in these Proceedings are those of the paper authors and do not necessarily reflect the views of the International telecommunication Union or of its membership.

Foreword

Chaesub Lee
Director
ITU Telecommunication Standardization Bureau

Innovation is fundamental for the development of new technologies, which will help to achieve social, economic and environmental sustainability. Academia's long-term approach to research is a major contributor to innovation. It also has actively participated in the success of information and communication technologies (ICTs). These are some of the reasons why ITU seeks to strengthen its relationship with academia, the source of so much crucial research and so many innovative ideas.

The Kaleidoscope conference is ITU's flagship academic event. Now in its eighth edition, the conference has matured into one of the highlights of ITU's calendar of events. This year, for the first time, the conference was held in conjunction with ITU Telecom World in Bangkok, Thailand. These peer-reviewed academic conferences increase dialogue between academics and experts working on the standardization of ICTs, uncovering emerging trends to assist the diffusion of research findings through the development of internationally recognized ITU standards.

Kaleidoscope 2016: ICTs for a Sustainable World called for research into ICT technical developments, innovative ICT applications, and policy and regulatory considerations relevant to the pursuit of the United Nations' Sustainable Development Goals (SDGs). The SDGs call for every industry sector to innovate in the interests of sustainable development. Innovations will be plentiful, multifaceted and tailored to context, but all innovators are looking to ICTs to form part of their portfolio of sustainability measures. Participants in Kaleidoscope 2016 highlighted research into ICT developments capable of supporting the broad spectrum of innovation required to achieve the SDGs. They emphasized the role of international ICT standards in providing the platform for this innovation to achieve its goals on a global scale.

Once again I would like to extend my gratitude toward all Kaleidoscope's participants for their support in propelling the series' success. Academia has been contributing to ITU's work for decades. A natural step was the introduction of the ITU academic membership category in January 2011. Academic and research institutes can now join all three sectors of ITU for a single fee: More than 120 academia members are participating in ITU's expert groups together with industry-leading engineers, policymakers and business strategists.

On behalf of ITU, I thank our technical co-sponsors, the Institute of Electrical and Electronics Engineers (IEEE), and the IEEE Communications Society; our supportive partners, the Institute of Electronics, Information and Communication Engineers of Japan (IEICE), Waseda University, the Institute of Image Electronics Engineers of Japan (I.I.E.E.J.), the European Academy for Standardization (EURAS), the University of the Basque Country, the Chair of Communication and Distributed Systems at RWTH Aachen University, Chulalongkorn University (CU), University of Geneva, Royal Holloway - University of London, and UNESCO Chair in ICT4D; our dedicated Steering Committee and Technical Programme Committee members; and, of course, our distinguished Chairman, Bundhit Eua-arporn, President of Chulalongkorn University, Thailand.

Chaesub Lee
Director
ITU Telecommunication Standardization Bureau

Chair's Message

Bundhit Eua-arporn
General Chair

ITU initiated its Kaleidoscope series of conferences in 2008 to provide an exchange platform for researchers and experts on the standardization of information and communication technologies (ICTs). The ITU academia membership category, established in 2010, gave further impetus to the Kaleidoscope series. I would like to express my appreciation to ITU for selecting Chulalongkorn University as this year's local partner.

It has been a privilege to chair Kaleidoscope 2016: *ICTs for a Sustainable World*. The conference's theme was a very topical one. ICTs are omnipresent. They are applied as enabling technologies in the business processes of virtually all sectors of industry and society, and it is clear that they will play an essential role as enabling technologies in achieving social, environmental and economic sustainability.

The Kaleidoscope 2016 Technical Programme Committee, chaired by Kai Jakobs of RWTH Aachen University in Germany, selected 25 papers from the 83 submissions received from 23 countries. The committee selected papers on the basis of double-blind reviews with the help of over 100 international experts, and also took on the challenging task of identifying candidate papers for awards. I offer my sincere thanks to all reviewers and members of the Technical Programme Committee for their generous contribution of time and expertise.

A side-event held the day prior to the Kaleidoscope conference, the ITU Secretary-General's Academia Consultation, offered to academia representatives, as well as interested industry members and governments a unique opportunity to meet with the ITU's Secretary-General to discuss ways to enhance collaboration between ITU and academia, regarding three areas in particular, including creating: a new ITU Journal; an Advisory Board of Academia to the Secretary-General; and, a platform/consultation mechanism to strengthen cooperation with the academic world.

Kaleidoscope 2016 featured three distinguished keynote speakers in Thomas Wiegand, Executive Director, Fraunhofer Heinrich Hertz Institute, and Chair, Image Communication, TU Berlin; Hossein Moiin, Executive Vice President and Chief Technology Officer of Nokia Networks, Finland, who delivered a talk on decoupling economic growth from carbon emission growth; and Simon Tuff, Principal Technologist, British Broadcasting Corporation (BBC), United Kingdom, and chair of the European Broadcasting Union (EBU) group on "Sustainable Technology in Broadcasting", who presented BBC experience on environmentally sustainable broadcasting.

In addition to selected papers, Kaleidoscope 2016 hosted two invited papers.

The first invited paper – authored by Nathalie Devillier (Grenoble Ecole de Management, France) – focused on ageing, well-being and technology, showing a French perspective on the relationship between quality of life improvement and digital rights management.

The second – authored by Luca Chiaraviglio; Nicola Blefari-Melazzi (CNIT/University of Rome Tor Vergata, Italy); William Liu; Jairo A. Gutierrez (Auckland University of Technology, New Zealand); Jaap Van De Beek (Lulea University of Technology, Sweden); Robert Birke; Lydia Chen (IBM Research, Switzerland); Filip Idzikowski (Poznan University of Technology, Poland); Daniel Kilper (The University of Arizona, USA); Paolo Monti (KTH Royal Institute of Technology, Sweden); and Jinsong Wu (University of Chile, Chile) – analysed the main challenges to deploy 5G networks in rural and low-income zones. The paper also defines the main pillars to follow in order to deploy 5G networks in such zones, as well as a proposal of a future network architecture.

Considering the state of today's technology and the extraordinary possibilities appearing on the horizon, Jules Verne's corner (JVc) at this year's Kaleidoscope conference asked futurists to imagine the potentially limitless Artificial Intelligence (AI) applications that can help address environmental, economic, and societal challenges concerning sustainable development and a sustainable future. JVc 2016: *Artificial Intelligence for a sustainable future: friendly companion or threatening conqueror?* featured three speakers from industry and academic circles: Sauvik Banerjjee, Global Chief Technology Architect and Innovation Lead, SAP, Germany; Malavika Jayaram, Executive Director, Digital Asia Hub, Hong Kong, China; and Prabhas Chongstitvatana, Professor, Department of Computer Engineering, Chulalongkorn University, Thailand.

Thanks to an ITU agreement with IEEE Communications Society, selected papers from each year's Kaleidoscope conference are considered for publication in a special-feature section of *IEEE Communications Magazine*. In addition, special issues of the International Journal of Technology Marketing, the International Journal of Standardization Research, and the Journal of ICT Standardization, are interested in publishing extended versions of Kaleidoscope papers.

All accepted and presented papers have been submitted for publication in the IEEE *Xplore* Digital Library. The Conference Proceedings from 2009 onwards can be downloaded free of charge from http://itu-kaleidoscope.org.

In closing, I would like to thank our technical co-sponsors, our supportive partners, and Alessia Magliarditi and her team from ITU for playing the leading role in the year-on-year progression of the Kaleidoscope series.

Bundhit Eua-arporn
President of Chulalongkorn University (CU), Thailand

TABLE OF CONTENTS

COMMITTEES

Steering Committee

- General Chairman: Bundhit Eua-arporn (President of Chulalongkorn University (CU), Thailand)

- Christoph Dosch (ITU-R Study Group 6 Vice-Chairman; IRT GmbH, Germany)
- Kai Jakobs (RWTH Aachen University, Germany)
- Takuro Sato (Waseda University, Japan)
- Mostafa Hashem Sherif (AT&T, USA)

Secretariat

- Alessia Magliarditi, Kaleidoscope Coordinator

- Erica Campilongo, Collaborator and Coordinator of Jules Verne's corner and Exhibit
- Martin Adolph, Technical Adviser
- Pascal Borde, Promotional support

Technical Programme Committee

- Chairman: Kai Jakobs (RWTH Aachen University, Germany)

- Mohammad Aazam (Carleton University, Canada)
- Marcelo Francisco Abbade (Universidade Estadual Paulista, Brazil)
- Sasan Adibi (Deakin University, Australia)
- Eyhab Al-Masri (Wilfrid Laurier University, USA)
- Ayesha Haider Ali (Lahore College for Women University, Pakistan)
- Costantinos Marios Angelopoulos (Bournemouth University, United Kingdom)
- Chaodit Aswakul (Chulanlongkorn University, Thailand)
- Luigi Atzori (University of Cagliari, Italy)
- Antoine Bagula (University of the Western Cape, South Africa)
- Ahmad Zaki Bin A Bakar (Universiti Teknikal Malaysia Melaka, Malaysia)
- Bartosz Balis (AGH University of Science and Technology, Poland)
- Paolo Bellavista (University of Bologna, Italy)
- Fernando Beltrán (University of Auckland, New Zealand)
- Michael Bove (Massachusetts Institute of Technology, USA)
- Carlos Caicedo Bastidas (Syracuse University, USA)
- Jorge Carapinha (Portugal Telecom, Portugal)
- Marcelo Carvalho (University of Brasilia, Brazil)
- Vicente Casares-Giner (Universidad Politécnica de Valencia, Spain)
- Alexandre De Masi (Université de Genève, Switzerland)
- Ilker Demirkol (Universitat Politecnica de Catalunya, Spain)
- Nathalie Devillier (Grenoble Ecole de Management, France)
- Christoph Dosch (ITU-R Study Group 6 Chairman; IRT GmbH, Germany)
- Tineke Mirjam Egyedi (Delft University of Technology, The Netherlands)
- Gerard Faria (TeamCast Inc, Singapore)
- José Ewerton P. de Farias (Federal University of Campina Grande, Brazil)
- Luca Foschini (University of Bologna, Italy)
- Ivan Gaboli (Italtel Spa, Italy)
- Ivan Ganchev (University of Limerick, Ireland / Plovdiv University "Paisii Hilendarski", Bulgaria)
- Aminata Garba (Carnegie Mellon University, USA)
- Joan Garcia-Haro (Universidad Politécnica de Cartagena, Spain)
- Molka Gharbaoui (Scuola Superiore Sant'Anne, Italy)
- Katja Gilly (Miguel Hernandez University, Spain)
- Ian Graham (University of Edinburgh, United Kingdom)
- Eva Ibarrola (University of the Basque Country, Spain)

- Kai Jakobs (RWTH Aachen University, Germany)
- Oliver Jung (Austrian Institute of Technology, Austria)
- Ved Kafle (National Institute of Information and Communications Technology, Japan)
- Wataru Kameyama (Waseda University, Japan)
- Tim Kelly (World Bank, USA)
- Andrej Kos (University of Ljubljana, Slovenia)
- Katarzyna Kosek-Szott (AGH University of Science and Technology, Poland)
- Andrey Koucheryavy (Bonch-Bruevich Saint-Petersburg State University of Telecommunications, Russian Federation)
- Ken Krechmer (IEEE, USA)
- Dhananjay Kumar (Anna University, India)
- Andreas Kunz (NEC Europe, Germany)
- Gyu Myoung Lee (Liverpool John Moores University, United Kingdom)
- Heejin Lee (Yonsei University, Korea)
- Mark Leeson (University of Warwick, United Kingdom)
- Yang-wen Liang (Samsung Modem Solutions Lab, USA)
- Fidel Liberal (University of the Basque Country-UPV/EHU, Spain)
- Morten Lindeberg (University of Oslo, Norway)
- Luigi Logrippo (Université de Québec en Outaouais, Canada)
- Waslon Terllizzie A. Lopes (Federal University of Campina Grande, Brazil)
- Kalle Lyytinen (Case Western Reserve University, USA)
- Giovani Mancilla (Universidad Distrital Francisco Jose de Caldas, Colombia)
- Maja Matijasevic (University of Zagreb, Croatia)
- Arturas Medeisis (ITU Arab Office, Riyadh station)
- Tetsuya Miki (University of Electro-Communications, Japan)
- Werner Mohr (Nokia Solutions and Networks Management International GmbH, Germany)
- Sean Moore (Centripetal Networks Inc, USA)
- Yoshitoshi Murata (Iwate Prefectural University, Japan)
- Rafael Paiva (Sennheiser Streaming Technologies, Germany)
- David Palma (Norwegian University of Technology and Science, Norway)
- Henrique Pequeno (Federal University of Ceará, Brazil)
- Alberto G. Perotti (Huawei Technologies, Sweden)
- Antonio Puliafito (University of Messina, Italy)
- Mubashir Rehmani (COMSATS Institute of Information Technology, Pakistan)
- Woo Seop Rhee (Hanbat National University, Korea)
- Anna Riccioni (University of Bologna, Italy)
- Cesare Riillo (STATEC, Luxembourg)
- Laurynas Riliskis (Stanford University, USA)

- André Samberg (Sec-Control Innovation, Finland)
- Viliam Sarian (Federal State Unitary Entreprise Radio Research and Development Institute, Russian Federation)
- Hans-Ulrich Schoen (self-employed, Germany)
- DongBack Seo (Chungbuk National University, Korea)
- Mostafa Hashem Sherif (AT&T, USA)
- Duncan Sparrell (s-Fractal Consulting LLC, USA)
- Michael Spring (University of Pittsburgh, USA)
- Ravi Subrahmanyan (Invisage Technologies, USA)
- Kurt Tutschku (Blekinge Institute of Technology, Sweden)
- Hiromi Ueda (Tokyo University of Technology, Japan)
- Manuel Urueña (Universidad de Carlos III de Madrid, Spain)
- Mathias Uslar (Institute for Information Technology (OFFIS), Germany)
- Hrishikesh Venkataraman (Indian Institute of Information Technology, India)
- Vino Vinodrai (McMaster University, Canada)
- Katarzyna Wac (Université de Genève, Switzerland)
- Robert Wojcik (AGH University of Science and Technology, Poland)

KEYNOTE SUMMARY

DECOUPLING ECONOMIC GROWTH FROM CARBON EMISSION GROWTH

Hossein Moiin

Executive Vice President and Chief Technology Officer for Nokia Mobile Networks
Nokia Corporation
Finland

For years there has been an implicit assumption that the economic growth is dependent on emission growth. However, ICT has the potential to reduce energy consumption of almost all other industries. GeSi SMARTer2030 (2015) report[1] shows that ICT has the potential to enable a 20% reduction of global CO2 emissions by 2030, thus holding emissions at 2015 levels, and demonstrating that judicious usage of ICT could effectively decouple economic growth from emissions growt.

In addition, we must do more to ensure that the energy consumption of the networks that carry the traffic is minimized. While we know that the use of ICT will continue to grow we must find ways to manage this growth in a sustainable manner. As an example consider that by 2017 the volume of mobile traffic will be 85 times greater than in 2010[2]. This creates a challenge: how to ensure that the energy footprint of tomorrow's networks doesn't grow?

At Nokia we are committed to combat climate change and to ensure the sustainable use of natural resources. Many telco operators, who are our customers, are now spending more money on electricity to power their networks than they are able to invest in expanding and upgrading networks to meet the ever-increasing demand for mobile data usage. We see it as an economic imperative to enable operators to expand their networks and serve their customers, while minimizing their total cost of ownership of their networks through lower energy consumption and CO2 emissions.

In Nokia we believe that environmental impact must be considered in all product life cycle phases starting from design to end-of-life. For product efficiency, the in-house design of key components is in focus whereas for system efficiency it is necessary to have holistic view of the network to identify opportunities for reduction of power consumption. A simple example is to put cells that are not needed into 'sleep' mode during off peak periods.

Networks evolution fueled by the growing traffic volumes, gives the industry an opportunity for continuous improvement of network level energy efficiency. We are enabling mobile networks to use smaller and smaller cells to reduce the energy consumption. A small cell close to a user, with clean line of sight taking advantage of several antenna elements to improve the quality of the radio channel can achieve much higher spectral efficiency than a large high power cell. For the same energy consumption, far more bits get transmitted.

Today Nokia leads the global development of 5G, for the future mobile technology enabling the industry to continue taking steps in energy efficiency and transform other industries. It is in this domain, the impact to other industries, that the greatest potential for decoupling of economic and emission growths lies. The transformative nature of the ICT industry is the key enabler for this decoupling. However, technology alone cannot ensure that such decoupling takes place. To do that we need the right enabling policies and regulations. We should all join forces: industry, policy makers and international institutions like ITU, unified by the goal to decouple economic growth from emission growth through ICT.

[1] http://gesi.org/portfolio/file/1

[2] Report ITU-R M.2370[2] indicates, based upon data from Nokia

KEYNOTE PAPER

IMPORTANT CONSIDERATIONS FOR ENVIRONMENTALLY SUSTAINABLE BROADCASTING: THE BRITISH BROADCASTING CORPORATION EXPERIENCE

Simon Tuff

British Broadcasting Corporation, London, United Kingdom, simon.tuff@bbc.co.uk

ABSTRACT

Many people recognise the important role broadcasters have in providing people with an understanding of environmental issues is often through programmes and other content but broadcasters themselves also have a significant impact on the environment through the process of content creation and distribution.

This touches on the editorial and creative challenges of informing audiences about the environment and focuses on the sustainability of programme making and distribution, using the UK and particularly the BBC as an example. In the first case using information from Albert, the UK carbon calculator for the television industry, the carbon footprint of different production workflows will be assessed. Then using research undertaken by BBC R&D and partners the paper will look at the impact of transmission and the energy differences between distribution platforms. In conclusion some thoughts on the challenges facing this industry will be shared as well as some suggestions on where efforts might productively be focused.

Keywords— Broadcaster, Programme, Production, Audiences, Transmission, Distribution, Content.

1. INTRODUCTION

The British Broadcasting Company (BBC) was founded in the UK as a radio broadcaster in 1922 [1]. Largely as a result of the General Strike of 1926 the benefits and influence of broadcasting became more apparent to the UK government of the day and the Company was transformed into the current British Broadcasting Corporation by Royal Charter in 1927. Today the BBC is a global broadcasting organisation, still firmly routed in its public service tradition, funded by a license fee and defined by its Royal Charter.

As it approaches the start of a new charter at the end of 2016, the BBC operates a range of UK television and radio services both via traditional transmission [both analogue and increasingly digital] as well as comprehensive set of on-line services including text pages, live streaming and on demand [or catch up] TV & Radio. The BBC also operates internationally as a news gathering organization and as broadcaster via the BBC World Service.

As a sensible starting point in trying to know what sustainable broadcasting might look like it would make sense to understand what proportion of the world's Green House Gases (GHG) are due to broadcasting. Based on IPCC (Intergovernmental Panel on Climate Change) data it is widely accepted that about 2% of the world anthropogenic Carbon Dioxide emissions are the result of the aviation industry [2] [3]. The ITU estimates that the ICT (Information, Technology & Communications) sector (comprising of telecommunications, computing and the Internet, but excluding broadcasting transmitters and receivers) contributes around 2 to 2.5 per cent of GHG emissions, or just under 1 Gigatonne of CO_2 equivalent per annum [4], whereas a study for Ofcom in the UK estimated that broadcasting accounted for around 1.8% of GHG emissions [5], a significance which surprises many as broadcasting may be more significant in GHG terms than most people have appreciated. Either way, all three of these industries are growing in size and thus GHG output but both aviation and ITC industries are making concerted efforts to become more efficient, whereas steps to reduce the impact of broadcasting are less widespread. This might be because of the distributed nature of the footprint but whatever the cause it is rarely considered a significant issue. In fact, as we will see later, the move from traditional transmission platforms to the Internet could be growing the industries footprint at an increasing rate.

For the purposes of this paper we'll focus on the impact of television. This is not because the radio or written output of the BBC isn't important, to the contrary, but because the creation of video, with the necessary addition of cameras, lights, sets, make up, costumes etc. is a far more resource hungry and complex undertaking with a resultant increase in the size of its carbon footprint. This is illustrated by the fact that the BBC's flagship television service, BBC One, consumes nearly a third of the BBC's annual license fee funding [6].

It is in this context that this paper looks at the challenges for sustainable TV broadcasting and explores how the BBC is becoming a more sustainable organization, whilst guarding its independence and maintaining its impartiality.

2. THE BROADCASTING SYSTEM

For the purpose of helping us examine the BBC's sustainability we start by trying to simplify the TV service provided by this complex public service organisation by identifying 3 broad categories.

Product manufacture (or programme making), product distribution (or broadcasting) & consumption (or watching TV).

Firstly, like any modern business, the BBC makes things and in this case its programmes, which are made using a creative, craft and storytelling process normally know as production. We will touch on the nature of those programmes and the editorial choices and constraints that drive their creation shortly.

Secondly the BBC has to distribute its products, or content as it is now frequently called, by broadcasting, although as you might imagine with the advent of the Internet broadcasting is becoming an increasingly ambiguous term.

Lastly BBC audiences watch or consume content on an increasing number of devices (from large TV screens to handheld Smart Phones) and via a range of different distribution mechanisms (including satellite, terrestrial transmission & the internet) known as platforms.

Like any large business or public sector organisation the BBC requires a back office function and in this case it spans the first

Figure 1: A simplified television service.

two components of the model shown in figure 1. The majority of the footprint mostly results from buildings, technology, travel and waste. As all these components are probably very familiar, they'll only be mentioned in passing, and this paper will focus on the more unique and hopefully interesting areas of programme production and distribution. Furthermore the BBC has worked hard to reduce this overhead and the Corporation's annual report [6] claims that "In 2015/16, 94% of the BBC's controllable spend was focused on content and delivery, with just 6% spent on running the organization". The BBC has also reduced emissions of CO2e from it buildings and onsite technology by 50,000 tines since 2008, a reduction of 33%.

3. EDITORIAL CHOICES.

Opinions on the shape of the world differ and it is important for public service broadcasters to navigate a way around this range of views. In fact, being impartial is an obligation for the BBC as set out in its Royal Charter. To aid its journalist and production staff and to clarify to audiences how this is to be achieved, the BBC publishes comprehensive Editorial Guidelines [7] and this states that "impartiality lies at the core of the BBC's commitment to its audiences. We will apply due impartiality to all our subject matter and will reflect a breadth and diversity of opinion across our output as a whole, over an appropriate period, so that no significant strand of thought is knowingly un-reflected or under-represented. We will be fair and open-minded when examining evidence and weighing material facts."

Although this sounds clear, sensible and indeed obvious, it is an area of reporting and programme making which is commonly misunderstood. It does not mean that all views are equal or will be given an equal amount of air time, as the approach should be based on due impartiality, with matters grounded in fact and should not be detached from fundamental social and democratic values. It also needs to be recognised that knowledge and views change over time and this needs to be reflected the way the BBC reports things.

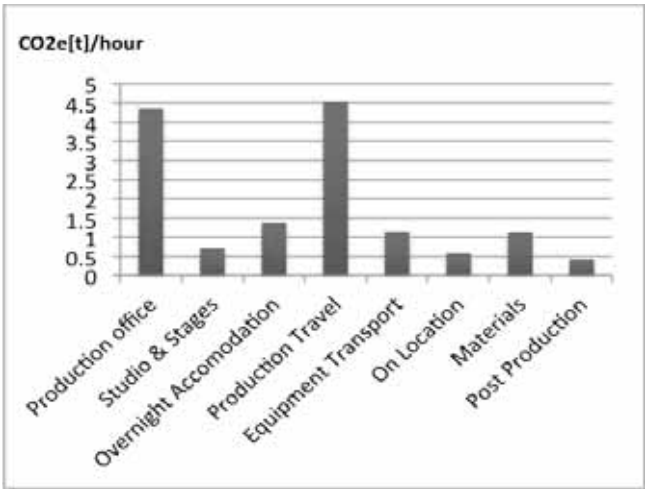


Figure 2: the carbon footprint of a typical hour of TV production totals13.6 CO_2e [t] per hour 2015/16.

The importance of this journalistic challenge for the BBC is explored by Professor Steve Jones (Emeritus Professor of Genetics at University College London) in his report to the BBC Trust in 2011 [8]. Here he highlights the need to consider due weight along side due impartiality. The BBC has a long tradition of making ground breaking and informative programmes on the subject of climate change, although obviously it hasn't always be called that, with early examples like BBC Radio's 1969 Reith lectures by Frank Fraser Darling called "Wilderness and Plenty" [9]. Landmark examples include David Attenborough's 2006 programmes called "Are We Changing Planet Earth?" and "Can We Save Planet Earth?" and Dr Iain Stewart's "Earth: The Climate Wars" three-part series from 2008. Last year Horizon, the leading science series on BBC 2, covered the topic in a programme entitled "Climate Change: A Horizon Guide" whilst BBC Four examined the data in "Climate Change by Numbers". In fact the BBC's news, current affairs and factual programmes regularly examine the topic and the subject features in comedy, drama and entertainment programmes across all platforms as well, as befits its importance in the modern world.

4. ENVIRONMENTALLY SUSTAINABLE PROGRAMME MAKING.

The BBC started its sustainability journey in earnest in 2008 when it set is first targets to reduce energy consumption, waste creation, water use and travel. Then in September 2009, in order to address the challenges of sustainable programme making, work began on a carbon calculator that could give producers and managers an idea of the CO_2e that their productions created. This project, know as Albert, had 4 main motivations. Firstly the BBC had no method for understanding what the impact of one of its main activities was on the environment. Secondly, the targets set the previous year were high level and difficult to engage with, so it was hoped that a carbon calculator would allow programme teams to understand and thus manage the impact of their activity. Thirdly, that waste costs and any publicly funded organisation is obligated to be as cost effective as possible. Lastly, because it was the right thing to do. So for the BBC, whose reputation is its lifeblood, this was and remains a key driving force.

Several things were done to attempt to ensure that the Albert initiative would be a success. First, the data input process was simplified as much as possible, so that ordinary production team members can understand what is required without specialist knowledge or training. The questions it asks are chosen so that the data can be gleaned, with a fair degree of accuracy, from a typical set of production notes and records such as travel expenses and facilities bookings. The process of collecting and imputing the data shouldn't be too arduous and typically takes one to two hours (depending the scale and complexity of the production). Next there has been clear and consistent executive leadership from the last two directors of television, not only championing the use of Albert in all productions but also pushing the sustainability agenda at the BBC as a whole. However the BBC commissions many of its programmes from independent production companies and so early in the calculator's life a relationship was forged with BAFTA (the British Academy for Film and Television Arts) to take on its hosting and to make it available to the whole of the UK TV industry. After about 5 years we have managed to ensure that pretty much every in house TV production logs its footprint in the calculator and 50% of the top 140 independent production companies in the UK are registered albert users, meaning that over 300 organisations have signed up to use it including Sky TV and UKTV.

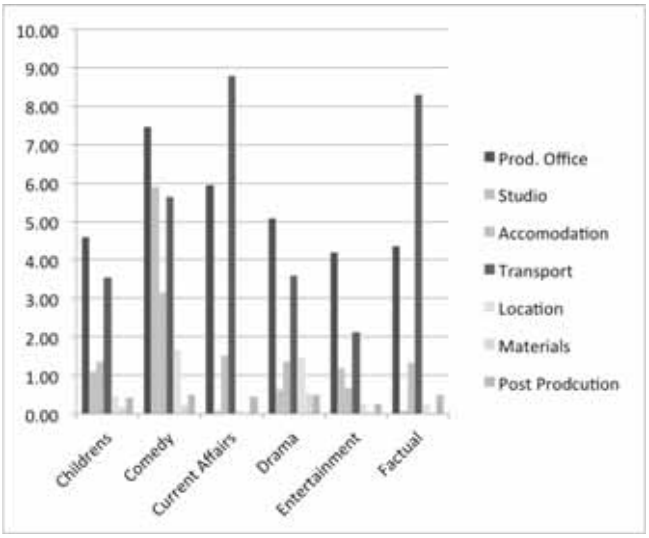


Figure 3: Emissions in CO_2e [t] per hour of content, for each production unit, by TV programme genre.

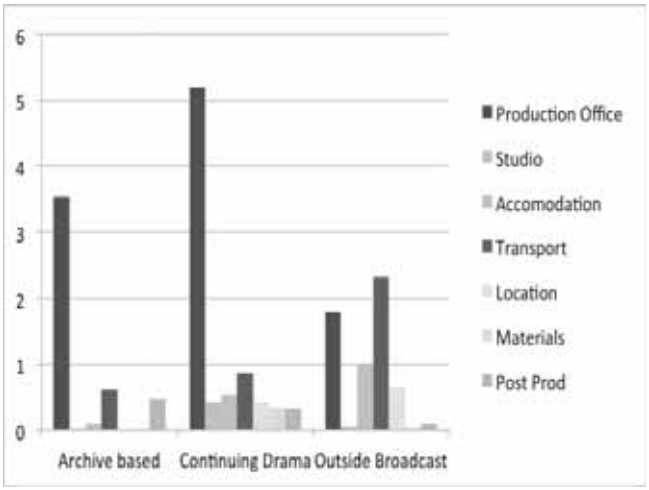


Figure 4: Emissions in CO_2e [t] by production method per hour of content for 3 main production types.

carbon calculator. It is made up from the main production activities including studio time, travel and location working, the production office, materials for sets etc.

If the main components of location working are summed we find that 60% of the typical footprint is due to working away from base. Surely the latest versions of technology, like CGI (Computer Generated Images), would provide a more efficient method of achieving the same outcome by virtually creating locations? This is true in some cases but when the editorial impact is considered using such artifice may be deemed as misrepresentation or untruthful by audiences, with an associated impact on credibility.

Furthermore these components are likely to vary depending on the type or genre of the programme being made. We can extract data to illustrate this as shown in figure 3.

However programme makers themselves are most interested in seeing how the different programme making process compare and this is shown in figure 4.

The ultimate objective of obtaining and reporting on this data is primarily to create a frame work in which productions can understand and reduce their GHG emissions. Production teams are increasingly compelled to record, then input this data and commissioning decisions could be influenced by what it shows. By going through a certification process currently known as "Albert+", productions can achieve a star rating (in a concept similar to those used for domestic appliances). As a result those productions that perform well or show they are effective in reducing their carbon footprint against clearly defined criteria are acknowledged by the BAFTA Albert Consortium (a pan-industry body of broadcasters and "Indies") and are awarded the "Albert+" status, which they can then add to the end of programme credits. In an industry that is inherently competitive and where motivation frequently comes from the recognition of peers and awards from august institutions we are hopeful that this could be quite effective in changing behaviors and developing lower carbon working practices. However it's recognized that this improvement does not just happen when the Albert analysis is produced; it requires leadership from senior members in the programme team or production company, occasional support from specialist sustainability advisers required to choose and deploy the right techniques and training for a range of team members to ensure they understand the importance of sustainability and the impact of their behavior. This approach seems to be working as amongst the range of programmes now displaying Albert+ status as sustainability improvers are two of the UK's most popular soap operas. Although we are yet to see if this has a significant impact in the footprint of broadcasting as a whole, it should still be seen as a serious expression of intent and as an example to the creative industries and their audiences.

4. SUSTAINABLE BROADCASTING.

Two hugely significant elements (as shown in figure 1) of the overall footprint of broadcasting are the distribution and consumption of programmes. In fact most studies on the broadcasting industry thus far have focused on this area.

4.1 Distribution

Broadcasters currently use a range of mechanisms to get their programmes to their audiences. The base of the BBC's TV distribution is a network of Digital Terrestrial Transmitters (DTT), which is part of the FreeView platform in the UK. Then there is Digital Satellite Transmission (DST) called FreeSat in the UK. Add to this distribution over the Internet by live streaming and Video On Demand (VOD) TV services, which include the BBC iPlayer. Such services are often referred to as Over The Top (OTT) TV and make calculating the carbon footprint difficult as they are additional to and circumvent the more established platforms. Audiences can also record programmes off air with PVRs (Personal Video Recorders) to watch later or download files to their mobile devices to watch in the near future using the iPlayer app. This wealth of options is in fact part of the problem, because not only do we have a wide range of platforms and combinations of usage to consider but each platform ends up with the capability to carry virtually all the same services resulting a substantial duplication of functionality and thus energy.

4.3 Terrestrial & Satellite Transmission

In a traditional terrestrial broadcasting system a single or network of transmitters use the RF spectrum to serve a large audience. The topology is often built around a few powerful widely spaced hill top transmitters, complemented by smaller, local, infill transmitters, in order to provide even coverage for the service area

Figure 5: UHF FreeView coverage of Mux A from Crystal Palace.

Figure 6: The footprint of Astra 2E carrying FreeSat TV services

Figure 5 shows the coverage achieved by the Arqiva transmitters for UK FreeView at the Crystal Palace site in south London and the supporting network of infill or relay transmitters. This "one too many" broadcasting, approach is never more true than when using a satellite system. Here, in the case of the UK, the beam of a single geostationary satellite is able to illuminate the whole country with FreeSat. The footprint for Astra's 2E Satellite is shown in figure 6. This has its drawbacks though because if you want to deliver a regional service, this can only be achieved by delivering to the whole footprint or by using more satellites. The BBC has a history of regularly "opting" local TV transmitters to carry local services for short periods. Local news bulletins for example are broadcast at the end of the national news programme by briefly breaking the national network in to a number of smaller networks, each carrying programme focused on events with the smaller service area. This continues to work well on FreeView but on FreeSat the opting approach is problematic and so the 18 regional news programmes can only be delivered by creating 17 extra versions of BBC One, which are broadcast all the time. Not the most bandwidth efficient approach.

The simple systems diagram in figure 4 shows the main power consuming elements that BBC R&D used when modeling the carbon footprint of TV broadcasting. The production element is not part of the distribution chain and so is not included in this part of the footprint but we have already discussed the impact of production earlier. The BBC funded infrastructure components are show in dark grey whilst those that are provided by the viewer are lighter grey. You can see that DTT & DST technologies have several advantages from a sustainability point of view. Firstly the system components can be clearly identified, measured and modeled. Secondly, although there is a significant investment in setting up the infrastructure, they have a long life and a constant operating power consumption regardless of the number of viewers. The UK FreeView infrastructure has already made one major technology step by moving from Standard Definition TV with stereo audio to High Definition TV (or T2) with surround sound and this was achieved without a significant increase in the power consumed. Partially this was possible because of the inherent characteristics and strengths of the technology but also because of the improved performance of Video Coding technology. It should however be noted that at the moment all major services are transmitted in both SD & HD to maintain compatibility with older television receivers and this does mean a doubling up of these channels, with a consequential increase the power consumed by transmitters. The same platform may even eventually carry Ultra High Definition pictures and sound. Additionally efficient systems design has resulted in terrestrial and satellite TV infrastructures sharing several components. For example the coding and multiplexing technology - that is the way that the services are data compressed (or coded) and bundled together into bit streams (or multiplexed) - is the same for both platforms.

There are 1,182 FreeView transmitter sites across the whole of the UK but for the BBC 83 of these are the most significant in terms of the size of audience they serve and the power that they consume to do this. As you might imagine with a system that has gown over a number of years and developed across a number of technologies there are a range of efficiencies but typically the transmitters have an ERP (Effective Radiated Power) of between 20-100kW with a range of efficiencies between 10-30%. At the top end is the Crystal Palace transmitter, which serves London, with an ERP of 200kW. It covers 4.9 million homes or about a fifth of the UK population. In simple terms this makes it a very efficient way of delivering TV to a large number of homes (i.e. 0.04 W/home).

A number of strategies are being considered to improve the energy efficiency of these transmitters and perhaps the most obvious is replacing older equipment with modern, more efficient, amplifiers, modulators and antenna. In some cases transmitter efficiency can be raised to 38% but even then the significant capital outlay and the large number of sites means that it could take over 20 years to produce a financial pay back. A more effective approach is to re-assign efficient, existing equipment to where it can be more effective and swap out the less efficient units to less critical or back up roles. In fact resilience is another area where cost effective improvements in power consumption can be achieved at a moderate increase in risk, for example by running back up systems as lower power or even off when not in use. Most main sites have 2 transmitters, each with a backup in a 2N + 2 configuration, so perhaps a 2N + 1 configuration, where the backup is shared, would be nearly as effective but save 25% of the operating energy. Other techniques being considered include increasing the Forward Error Correction producing a more robust signal, which needs less radiated power to achieve the same coverage (although be it with an increased delay) and a feedback system using receivers across the coverage area to allow the transmit power to be turned down when the atmospheric conditions allow good propagation, only turning up the transmitter power when the reverse is the case. The size of the impact of these

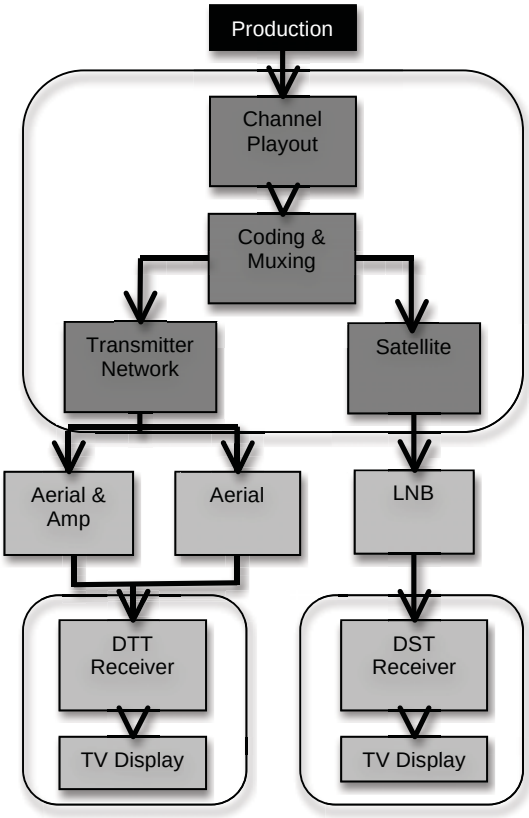


Figure 7: DTT & DST distribution system

measures is not yet known but it is expected to be relatively slight and certainly less than 10%. Additionally we have to be mindful of the impact across the whole infrastructure because if the end result is the addition of more powered head end/aerial amplifiers in thousands of homes across the coverage area to compensate for lower radiated field strengths, then the overall system power saving might be easily negated. When trying to understand the impact of carbon footprint of broadcasting we tend to use the metric of per viewer-hour of TV consumed. This approach not only allows us to include the impact of the audiences equipment being in use but it also distributes the over head of fixed infrastructure over the amount it has used. Figure 8 shows this well because as you will note, BBC One, the BBC's most popular service, looks very efficient broadcast over DTT with audiences frequently in excess of 6 million and emissions of 0.000265 kg CO_2e/viewer-hour whereas the BBC's Parliament service is far less widely viewed and its share of the infrastructure makes DTT look like a far less sustainable choice at 0.196 kg CO_2e/viewer-hour, so much so in fact that it's way off the top of the scale used in figures 8 & 9. In these TX means the proportion resulting from Transmission and C&M from the Coding and Multiplexing.

When we get to the edges of DTT coverage, signal strength has dropped and the viewer needs to use an aerial amplifier to ensure good pictures; we can see the impact of a small piece of technology adding 9.85 x10-3 kg CO_2e/viewer-hour and being replicated across the 25 million homes in the UK as shown in figure 9. In fact the impact for BBC Parliament is that its overall emissions only increase by 5% where as for the much more popular BBC One the factor increases by 37 times.

On the positive side the 2008 decision under the EU Ecodesign Directive has reduced the energy losses in passive, standby and off modes of a broad range of TV products have been in recent years.

On average the BBC R&D white paper WHP 189 [10] concludes that in 2011 the DTT distribution for the UK emitted an average of

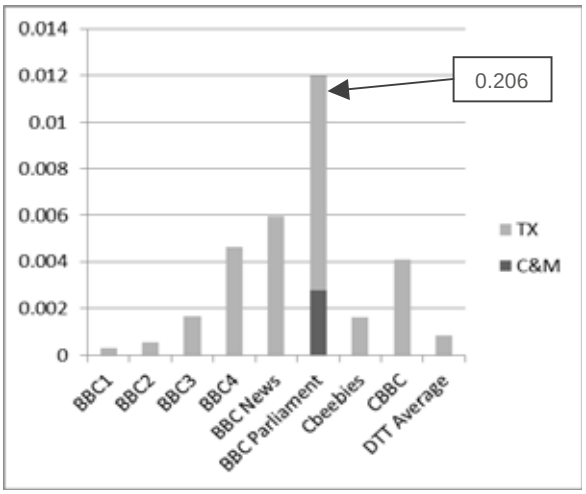


Figure 8: Emissions in kgCO_2e from DTT without an aerial amplifier.

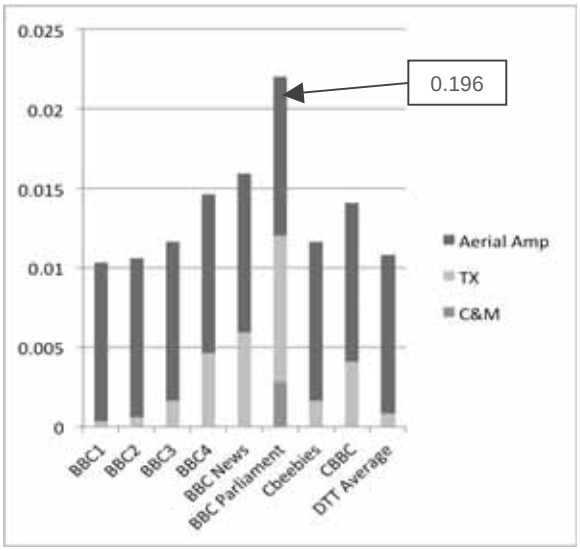


Figure 9: Emissions from DTT with an aerial amplifier.

8.45 x 10-4 kg CO_2e/viewer-hour. The DTT system probably consumes more power than this now, as many of the transmitters were increased in power during the digital switchover from analogue TV (which was completed in 2012). On the positive side, this meant that the whole duplicate, analogue, infrastructure was turned off, probably more than halving the power consumption of terrestrial TV in the UK.

4.3 On-Line Delivery

The outlook for delivery over the Internet is far more complex, as many of the components are much harder to identify and when they are, it is hard to know how much of their energy consumption is due to delivering television pictures and sound.

In figure 10 we see the simplified system diagram for the delivery of video on demand [or VOD] TV. The only component common with Figure 7 is the programme production element and once again this is not included in this part of the analysis as it is not part of the distribution system. The infrastructure elements are once again shown in darker grey than those in the home. The coding

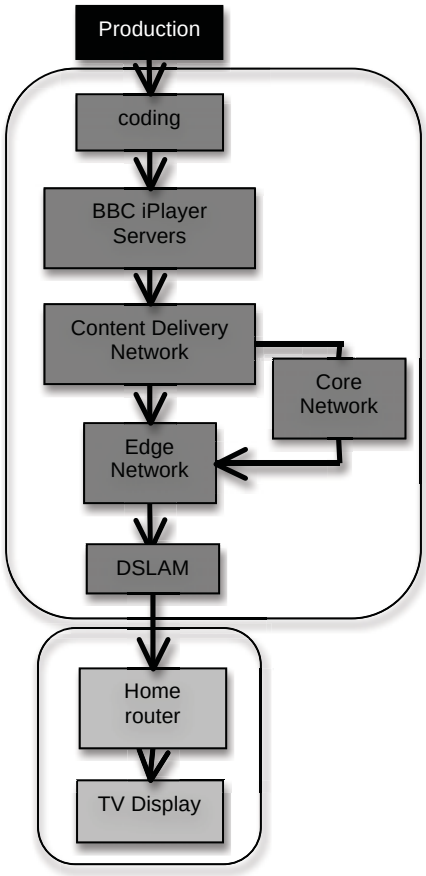


Figure 10: A typical Video on demand distribution chain.

phase is where BBC programmes are encoded especially for on-line distribution and placed on hosting servers, which form part of the BBC's iPlayer and the BBC's wider online services. The Content Delivery Network (CDN) is made up of the servers necessary to meet user requests for content and is partly the BBC's own BIDI [BBC Internet Distribution Infrastructure] and partly those of service providers. These CDN services carry the content of a large number of content providers and are scaled to meet peaks as demand varies across the day (with the peaks usually occurring at midday and late evening). From the CDN content is delivered to the home across the internet, which for these purposes has been divided in to three components: the core network and the edge or metro network sits within the telcos and the last, consisting of digital subscriber line access multiplexer (DSLAM) used to connect via ADSL, is provided by Internet Service Providers (ISP). Lastly all consumers will need to power a home modem/router.

Not only is it difficult to know which components are in use for VOD at any one time, as the whole infrastructure is shared across a range of on-line activities and multiple users simultaneously, but the nature, make up and technology of all these components is changing continuously and relatively rapidly, especially compared to the much more stable transmission infrastructure discussed earlier. This has meant that BBC R&D have had to make a range of assumptions to be able to develop a usable model and these are documented in White Paper WHP 189 [10].

The scale of what is happening with on-line traffic is dramatic and well described in Sandvine's 2015 [11] report that claims that Netflix use consumes 36.5% of all downstream Internet bandwidth during peak periods in North America, compared with just 2.7% for Facebook. This is serving a user base, which is only about 12% of the US population and was calculated before higher quality

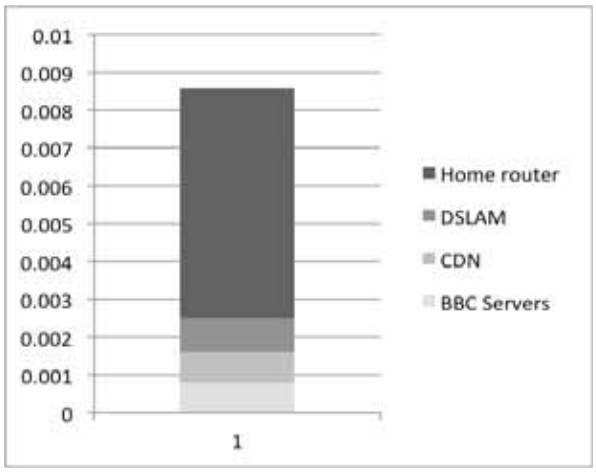


Figure 11: Emissions from video on demand.

video & audio formats (such as 4k or HDR) were in regular use. As audiences increase their use of these on-line platforms and use higher quality formats, then the energy consumption will grow in proportion. This supports the work of Professor Andrew Ellis of Aston University in the UK who has suggested that in 2015 the internet accounted for at least 8% of the UKs power consumption (with up to 3% of the UK power consumption being used by the major Telcos alone) and that this was expected to double over the next four years if current behaviors and technologies remained unchanged. All this presents some real challenges, not only for sustainability but for other prized concepts like Net Neutrality. It also demonstrates the continued importance of video, in a TV type format, for modern audiences. Consumption in these sort of numbers seems to suggest that people are still interested in long form story telling but what is perhaps less well understood is how they find and share this content and the importance of social media in this process. A topic for another paper perhaps?

4.3 Comparing DTT & On-Line

Figure 11 shows us how the carbon footprint of the on-line distribution is made up. Once again we can see the significant impact of a small technology component being replicated across millions of households as the home router is responsible for nearly three quarters of the overall footprint, although once again you could argue that this router would be present and powered to deliver all the other data & telephony services the modern household requires, but even so it probably going to need to be of higher specification, work harder and be switched on for longer if used for serving video to the home.

With a better understanding of the way the broadcast and on-line carbon footprints are made up we would now like to understand how the two platforms compare. If we use the metric of CO_2e/viewer-hour again and start with a DTT implementation with good coverage where the viewer doesn't require additional UHF amplification, then the emissions are shown in figure 12. As we might expect the programmes with large audiences operate very efficiently and, with the exception of BBC Parliament, they produce typically a half to a third less CO_2e/viewer-hour. However if an aerial amplifier is required, as shown in figure 12, then this advantage is removed and the current model shows that IP distribution (including the local router) is still more CO_2e efficient for certain services.

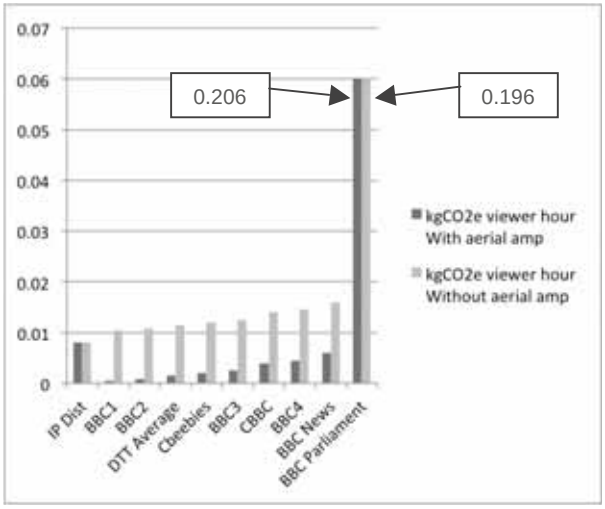


Figure 12: Comparison of the carbon footprint per viewer-hour of DTT broadcast without and without an aerial amplifier and with on-line distribution.

5. CONCLUSIONS

5.1 Editorial

As the global social, political and most importantly scientific consensus is now firm that anthropogenic climate change is happening, broadcasters are expected to increasingly cover this topic in their programme making and reporting, however there will still be editorial challenges to ensure impartiality and thus hard questions to be asked of all parties. It is also a topic that will require creativity to ensure that audiences are informed, educated, entertained and engaged. Talented production teams should be able cover the issues and explain the complexities through a range of programme genres and formats. Although most broadcasters will create long form TV and radio content to do this, they will also need to use social media so that their audiences can find and share this content and "join the conversation", as they say. Linear programming on its own will only meet its potential if these tools and platforms are used effectively.

5.2 Production

As well as the editorial complexity there are three further factors that will enable increasingly suitable production.

Teams wish to make the best programme possible for audiences and in order to obtain good ratings or audience approval, to be re-commissioned or to be in the running for an award. They will invariably want to use the approach that they consider best on screen over the one that is more sustainable. This means that sustainable technology and practices have to be credible. For example as the colour temperature of LED and other low power techniques are now becoming widely acceptable to lighting directors, they are being widely deployed, not only for their power efficiency, but they also have the added advantage of being lighter and safer. On location catering units now routinely issue bottles or mugs for crew members for reuse and this reduces waste, empowers crew members and creates a branding opportunity for production mementos…

As is common across many areas of sustainability production team behavior is key, especially so with an undertaking which is so reliant on the activity of a large number of people, many of whom are freelance or contracted. Making sure that sustainable values, leadership, understanding and motivation are in place and regularly re-enforced is an important part for ensuring continued improvement.

The TV production world is under continued financial pressure. This is exemplified by the way that sports rights continue to eat into the commissioning budgets in the UK. Already 40% of the total production spend in the UK is on sport programming and Lord Puttnam in his report on "a Future for public service television" [12] expects this to rise. Given these pressures it easy for managers and producers to focus on the bottom line and ignore the role of sustainability in production. It is therefore important to exploit the cost saving relationship that sustainability can bring to decision-making. For example as travel cost rise it will make more sense to hire crews and equipment locally rather than to ship the productions own to location but this will only work if unit managers can have the tools to allow them to find the right gear and hire good personnel before travelling. Another example of virtuous synergies that the BBC has sought to exploit is the use of bioethanol-powered fuel cells for powering remote cameras. Not only are these more sustainable than batteries or diesel generators but they are lighter, safer and silent!

At the end of the day a typical hour of TV production produces 13.6 tonnes of CO_2 (fig 2, financial year 2015/16) whilst the average UK home produces 4.5 tonnes per year (from gas and electricity).

5.3 Distribution

As we have seen a well-designed DTT network, with suitable receiving equipment, serving large populations can be more efficient than an IP distribution model for live services as shown in figure 13. In fact BBC R&D investigations reported in WHP 258 [12] indicate that even for catch up viewing, DTT, using a PVR (Personal Video Recorder), can be 60% more efficient than an iPlayer VOD alternative. This however cannot be the full story as one of the main advantages of the iPlayer is that you don't have to know what you want to watch before it is broadcast. Even with better Electronic Programme Guides (EPG) or sophisticated use of social media the PVR will not be programed to record all the progammes the viewer may want to watch and could well end up recording programmes that the owner never watches, so it won't surprise you learn that many organisations, including the BBC, are looking at PVRs that have a capability to learn their owners preferences and therefore predict which programmes they might record in order to close this gap. There are also some interesting options for hybrid platforms combining the best of broadcast and IP distribution. This could mean that during peak times services are delivered by broadcast but then as audiences drop, overnight for example, distribution could switch to online at the point where that platform becomes more efficient.

These approaches are still only research projects and for such complex systems to be widely adopted a combination of technical standards and regulatory frameworks will probably need to be in place. Although the BBC and its service providers are working to improve the efficiency of the distribution chain we have demonstrated how important the consumption of the audiences' own equipment in the home is and there are considerable pressures in this area. Efforts are being made to produce more efficient receivers and the IEEE's de facto standard for green computing covers Televisions in IEEE 1680.3 and IEEE 1680.6 [13] for the environmental assessment of complex set top boxes.

This is all the more important when we realize that home screens over the last 10 years have been getting larger and larger, thus

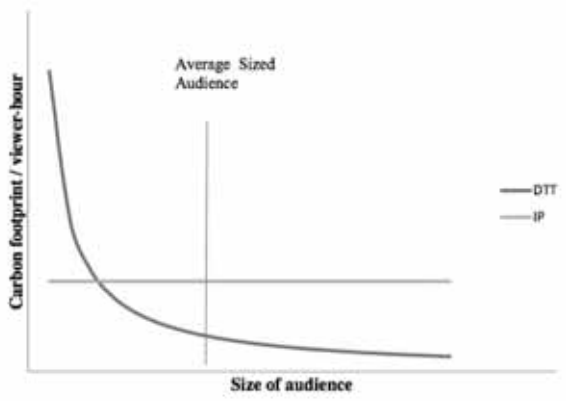


Figure 13: The relationship between carbon footprint per viewer-hour for DTT and video-on-demand and the size of the audience.

requiring more power (although a recent report by IHS shows that this trend may now have slowed or even reversed a little with a typical primary screen in the UK now being 55"). It is however more likely that with large displays there may be occasions when several people watch the same TV (for the purposes in these calculations the BBC assumes that typically a TV is watched by 1.5 people), which offsets the power consumption to some extent.

6. WHAT NEXT?

The technology of television continues to improve. The development of HDR technology (High Dynamic Range or an increase in the range of light to dark that a TV picture can reproduce) will provide better pictures. Displays will also run at higher frame rates, carry more pixels (e.g. 4K & 8K) and show more colours (Wide Colour Gamut). In fact elements of these enhancements are already appearing on some online programming from the likes of Netflix and BT. All of these technologies will likely require more power in production, distribution and reception and in the early days will create duplication as new formats supplement old, as legacy formats continue to be broadcast to serve older TVs e.g. it would be useful to have a better understanding of the energy consumption when distributing TV over Mobile telephone networks. Add to this the range of platforms and the number of connected devices, which are frequently used simultaneously, then there should be some scope to manage this duplication more effectively. Bristol University and the BBC are currently refreshing some of the data in WHP 189 [10] and including analysis for the impact of tablet and hand held devices. Lastly the rate of change itself and its impact on the carbon footprint of broadcasting are not well understood and we have very little understanding of the embodied energy in all these devices. We believe that the main display in typical household is used as such for 5 to 7 years but when it is replaced in the living room it frequently finds further use in another room in the home but standards that help us understand embodied energy and support a more circular or life cycle approach to equipment design and use would be most welcome.

So far there are few standards covering environmental sustainability specifically for use by broadcasters. A best practice for production is beginning to emerge. Some standards from other sectors like ITC can be applied to help deliver better production technology but there is an opportunity for standards to facilitate an increase in the sustainability of distribution platforms and reception equipment. As this last area is where the footprint is largest and the scope for invitation with hybrid IP technology is greatest, it presents an intriguing possibility.

7. REFERENCES

[1] The Timeline of the BBC, Wikipedia
https://en.wikipedia.org/wiki/Timeline_of_the_BBC

[2] Joyce E. Penner, David H. Lister, David J. Griggs, David J. Dokken, Mack McFarland UK, "Aviation and the Global Atmosphere," *IPCC Special Report*, 1999
https://www.ipcc.ch/pdf/special-reports/spm/av-en.pdf

[3] ICAO Secretariat, "Aviation's contribution to Climate Change," *ICAO Environmental report*, 2010
http://www.icao.int/environmental-protection/Documents/EnvironmentReport-2010/ICAO_EnvReport10-Ch1_en.pdf

[4] ITU, "ICTs and Climate Change."
http://www.itu.int/themes/climate/docs/report/02_ICTandClimateChange.html

[5] Forster C., Dickie I., Maile G., Smith H. and Crisp. "Understanding the Environmental Impact of Communication Systems," *Report for Ofcom*, 2009
http://stakeholders.ofcom.org.uk/binaries/research/technology-research/environ.pdf

[6] BBC Annual Report & Accounts 2015/16
http://downloads.bbc.co.uk/aboutthebbc/insidethebbc/reports/pdf/bbc-annualreport-201516.pdf

[7] BBC Editorial Guidelines
http://downloads.bbc.co.uk/guidelines/editorialguidelines/pdfs/Editorial_Guidelines_in_full.pdf

[8] BBC Trust, Steve Jones, content research Imperial College London, "Impartiality and the accuracy of the BBC's coverage of science," *A review by the BBC Trust*, July 2011
http://downloads.bbc.co.uk/bbctrust/assets/files/pdf/our_work/science_impartiality/science_impartiality.pdf

[9] Frank Fraser Darling, "Wilderness & Plenty," *The BBC R4 Reith Lectures*, 1969
http://www.bbc.co.uk/programmes/p00h3xk5/episodes/guide

[10] Jigna Chandaria, Jeff Hunter, Adrian Williams, "A Comparison of a Carbon Footprint of Digital Terrestrial Television with Video on Demand," *BBC R&D White Paper WHP 189*, March 2011
http://downloads.bbc.co.uk/rd/pubs/whp/whp-pdf-files/WHP189.pdf

[11] Sandvine, "Global Internet Phenomena Report," 2015
https://www.sandvine.com/downloads/general/global-internet-phenomena/2015/global-internet-phenomena-report-latin-america-and-north-america.pdf

[12] Lord Puttnam et al. "A future for public service television: Content & Platforms in a Digital World", *A report by GoldSmiths University of London*, June 2016
http://futureoftv.org.uk/wp-content/uploads/2016/06/FOTV-Report-Online-SP.pdf

[13] IEEE P1680.6 – "Standard for Environmental Assessment of Complex Set Top Boxes" IEEE P1680.6
https://standards.ieee.org/develop/project/1680.6.html

All links retrieved on the 25th July 2016

SESSION 1

ROLE OF ICT IN ENVIRONMENTAL SUSTAINABILITY

S1.1 How organisations can assess and improve their green ICT activities in a standard and efficient way.

S1.2 Mobile signal extension in deep sea - towards a safe and sustainable fisheries.

S1.3 Human safety considerations in the emerging ICT environment.

HOW ORGANISATIONS CAN ASSESS AND IMPROVE THEIR GREEN ICT ACTIVITIES IN A STANDARD AND EFFICIENT WAY

Albert Hankel & Patricia Lago

VU University Amsterdam
Department of Computer Science
De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands

ABSTRACT

This study demonstrates how a maturity model on Green ICT can help organisations improve themselves and become more environmental sustainable in a standard and efficient manner. For this we have used the SURF Green ICT Maturity Model and facilitated the use of this model in four organisations. These organisations participated in a maturity scan, evaluation session to discuss the results of the scan and a questionnaire on the use of the model. This field study showed that individual participants were very positive about the use of such a model and that it provided inspiration for improvement, both to reduce the environmental impact of ICT as well as to use ICT as an green solution for other business processes.

Keywords— Green ICT, ICT for Sustainability, Green IS, Maturity Model, Benchmarking

1. INTRODUCTION

In academic research as well as in industry, the environmental impact of ICT is an important topic, spanning across multiple disciplines. ICT is seen as both a relevant contributor to CO_2-emissions due to its increasing carbon footprint [1], and as an enabler for reducing the footprint of other sectors through "smart" systems (e.g. smart buildings, smart grids). During the years many studies tried to evaluate the general impact of ICT on the environment. According to a report of the Global e-Sustainability Initiative, ICT itself is roughly responsible for 2% of global CO_2-emissions, while ICT solutions have the potential to reduce global CO_2-emissions by up to 16% [2].

We define Green ICT as a combination of activities that minimise the negative impact of ICT on the environment and optimise the positive impact ICT can have. Or, in other words, as any activity that considers the direct, indirect and systemic impact of ICT on the environment [3]. Because the relations between ICT and the environment are numerous and often complex, it helps to consider all these effects. As a general-purpose technology information and communication technologies can be used by themselves or as part of other technologies. This is the reason why ICT is viewed as an environmental friendly solution as mentioned above. Others [4] have defined these effects in three orders:

- **First order** or *primary* effects: effects of the physical existence of ICT (environmental impacts of the production, use, recycling and disposal of ICT hardware).
- **Second order** or *secondary* effects: indirect environmental effects of ICT due to its power to change processes (such as production or transport processes), resulting in a modification (decrease or increase) of their environmental impacts.
- **Third order** or *tertiary* effects: environmental effects of the medium- or long-term adaptation of behavior (e.g. consumption patterns) or economic structures due to the stable availability of ICT and the services it provides.

While these effects are widely recognised and can be understood on an abstract or global level, it is often difficult for individuals or organisations to apply them.

In the past decade more and more businesses have realised that their actions have long term effects on the environment and society and are taking responsibility for their actions through several social and environmental initiatives that reduce their impact [5]. Green ICT can contribute significantly and can thus help organisations achieve their sustainability goals. In order to do so, they need to know the answer to at least two questions:

- What are the key environmental impacts arising from ICT?
- How can ICT assist organisations in their efforts to improve their environmental sustainability?

We know that ICT can consume large amounts of energy in datacenters and in (mobile) communication networks for example [6]. It is also important to consider the use of rare materials in the ICT equipment as well as what happens with the equipment at the end of their life cycle, also known as e-waste [7]. On the other hand, ICT solutions can reduce travel, dematerialise paper use and material use, and optimise business processes as a whole. When organisations consider Green ICT, they often focus only on the first part, to reduce the environmental impact of ICT. The second part, using ICT as a solution is much less common.

In order to facilitate considering both aspects, frameworks, maturity models and their likes have been published both in academic and practitioners literature. In addition, there are general impact assessment tools that can also be applied to ICT, such as life cycle analysis or green house gases audits (see Ecofys et al. [8] for an overview of general tools applied to ICT).

Most of the specific models and tools focus on energy efficiency and reducing the negative impacts of ICT, such as those developed by the Green Grid and the OpenDCME model. While these are mostly focused on data centres, others such as those developed by Gartner [9] and Molla et al. [10] capture the entirety of ICT but are still only limited to the direct impacts in scope or are very general/abstract. A few tools have been developed that also include the positive impacts of ICT, such as those by UK HM Government [11], deMonsabert et al. [12] and Donnellan et al. [13]. Still those mostly focus on the negative impacts, too. From a system perspective or the total global footprint of human society this seems strange since the negative impacts are responsible for 2% of that footprint, while the positive impacts have the potential to reduce the global footprint by 16% [2].

As far as we can derive from research literature and practice, there is a lack of publications on assessing the quality of the tools, if and how they are being used and whether they achieve their intended effects. Similarly, there has not been a lot of research on what capabilities companies need in order to green their ICT and how to measure these capabilities [5]. While such questions might be trivial for other assessment topics, this is not the case for the environmental impact of ICT. Environmental sustainability is typically multi-dimensional and prone to local optimisations and it is therefore complex to assess progress.

From the perspective of usability in practice, the published models vary in how abstract they are, their scope and ease of use. By looking at the meta-level, it can be seen that most of them have a very similar structure. Most of them have a general idea of what should be included in Green ICT, and contain a collection of several concepts which consists of several components. Some of the models include extra aspects like approaches or maturity levels. This could provide a standard structure for assessing, evaluating and improving the use of green ICT. If many organisations use the same model or at least the same structure in their models, this could pave the way for standardisation and eventually benchmarking.

In this study we want to demonstrate how organisations can use a maturity model, the SURF Green ICT Maturity Model, to assess and improve the use of Green ICT. As we wanted to know how organisations would respond to such a maturity model and whether it would be effective, we facilitated the use of the maturity model in four different organisations. The following sections describe the maturity model, the method of the field study, the results and the discussion of the results.

2. SURF GREEN ICT MATURITY MODEL (SGIMM)

SURF, the Dutch higher education and research partnership for ICT, decided to develop a maturity model on Green ICT after interviewing a number of Dutch higher education and research institutions. In these interviews the institutions expressed a clear need for some way to know how well they are doing in terms of Green ICT. SURF wanted to develop a maturity model based on expert views and opinions and validate this through a survey spread amongst practitioners. The SGIMM was developed by SURF and a number of Green ICT experts, both from the Dutch higher education and research community as well as outside it. Responsibility for ICT in organisations part of this community is typically delegated to an ICT department. The SGIMM was therefore designed from the ICT department's perspective. Even though the SGIMM is developed for higher education, it can be used easily by other organisations as it covers topics that are equally relevant for any organisation. More information on the model can be found at [14] where it is also available for download to use freely under a creative commons license.

The concept of the maturity model is based on the Capability Maturity Model, representing a framework with five maturity levels for quality and process improvements. The five levels are (1) initial, (2) repeatable, (3) defined, (4) managed and (5) optimising. At the lowest level, the initial level, the organisation does not provide a stable environment for the activity. At this level the process is ad hoc. However, at the highest level, which is the optimising level, the entire organisation is focused on continuous process improvement [15].

The SGIMM conceptually consists of four domains covering negative and positive impacts and aspects of ICT. Each domain consists of attributes that have a definition, factors involved and descriptions of each of the five maturity levels. Three domains and attributes are generally applicable to any organisation, being: 'Green ICT in the Organisation', 'Greening of ICT' and 'Greening of Operations with ICT'. The fourth domain is sector-specific and covers 'Greening of primary processes with ICT'. For instance, for the higher education sector, the primary processes would relate to education and research. The first three domains and attributes are summarised in Table 1.

A few models and frameworks exist that help apply Green ICT principles to primary processes in an organisation. For example, the framework in [13] contains the capability building block "ICT-enabled business processes". They found that involving the ICT department as well as 'business' raises awareness on both sides of the potential of Green ICT. However, the framework seems to remain high-level and offers little practical guidelines to apply Green ICT to business processes. In general, the impression we gather from other work - and this is also what we experience - that it is difficult to apply Green ICT outside the datacenter.

The SGIMM is designed to give organisations insights into the maturity of Green ICT of the organisation. It is set-up as a self-assessment and enables organisations to have an internal dialogue, to gain agreement on the status quo and to

Table 1. Overview of the first three domains of the SURF Green ICT Maturity Model.

Green ICT in the organisation	**Greening of ICT**	**Greening of Operations with ICT**
Green ICT Strategy	Housing	Travel reductions with ICT
Governance of ICT Services	Computing Infrastructure	Area reductions with ICT
Green ICT Procurement	Network Infrastructure	Energy reductions with ICT
E-waste Policy	Storage Infrastructure	Paper reductions with ICT
Green ICT in Information Management and Architecture	End user ICT equipment	Feedback and Decision support
Community Collaboration	Software and ICT services	
Green ICT Supply Chain Management		

define actions for improvement. By letting several individuals within an organisation score the attributes and discussing theses scores with the participants (average, minimum, maximum scores, etc.), an organisation can identify weak and strong Green ICT aspects.

The relevancy and completeness of the SGIMM was evaluated in a survey amongst Dutch practitioners by Hankel et al. [16]. The results from that survey were used to improve the maturity model. A few attributes were removed or adjusted, but in general the model was well received. The version that was released after these adjustments was used in our field study here.

3. METHOD OF FIELD STUDY

The purpose of the field study is to see how the SURF Green ICT Maturity Model is used in practice. We want to demonstrate how organisations can use the model to assess and improve the use of Green ICT and show the effect of the model.

Our theory is that the maturity model helps organisations improve on the use of Green ICT within the organisation. We assume that assessing the level of maturity of Green ICT increases awareness and motivation to improve. At this point at least, it is impossible to test this theory, as this requires large-scale use of the model and access to users of the model versus organisations not using the model. However we do want to see if using the maturity model has an effect on organisations. That is why we are gathering evidence for this theory through the use of case studies.

Each of the cases are facilitated in the same manner. They followed the same approach as a normal self-scan, with a few deviations explained below. For the self-scan, the following steps are recommended:

1. Someone in the organisation should take the initiative. This could be the CIO or an ICT manager with sustainability in his or her portfolio. This person will be the Assessment Manager (AM). It is important that the AM has the influence and the ability to make sure that the self scan is properly done and that the action plan is implemented.

2. The AM starts with composing an assessment team. This is a group of people who represent the organisation in filling out the maturity model. Their scores are used to get an average maturity score for each attribute. After obtaining the averages, they are discussed during an evaluation session at which also a number of possible actions for improvement are defined. To obtain good results, the right combination of team members is crucial.

3. The AM organises a kick-off for the assessment team. During this meeting, the model will be discussed, as well as how to fill out the maturity levels. Finally the whole assessment process is explained. After the meeting the SGIMM spreadsheet is sent to all the participants.

4. All participants individually fill out the spreadsheet and score all the attributes. When everyone has sent their scores back to the AM, he or she will analyse the results and create a summary.

5. The summary of maturity scores and how it relates to the individual scores are discussed in an evaluation session with the assessment team. Based on this discussion, actions for improvement are defined.

6. After carrying out the action plan or after a certain period of time, the AM can evaluate the obtained results. This is also a good time to use the SGIMM again to gain new maturity scores and formulate new actions for improvement. This creates a cycle for constant improvement of Green ICT in the organisation.

Note that the evaluation session is the key part here, and not necessarily the scores themselves. The purpose of using the model is more focused on improvement rather than being absolute correct about the scores. Of course the scores do give a good indication of the maturity of the organisation, especially when supported by the (summary) results of the evaluation session.

Table 2. Example of scores with 9 participants. The deviations of 1 point or more from the average are highlighted.

Greening of Operations with ICT	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**Avg**
Travel reductions with ICT	2	3	2	1	3	2	3	3	2	2,3
Area reductions with ICT	2	2	1	2	3	3	4	2	2	2,3
Energy reductions with ICT	1	2	2	3	2	1	4	2	1	2
Paper reductions with ICT	2	2	3	3	3	2	3	2	2	2,4
Feedback and decision support	1	3	2	1	2	1	3	1	1	1,7

For the case studies, we facilitated the Assessment Manager of each participating organisation by organising the evaluation session, analysing the maturity scores beforehand and

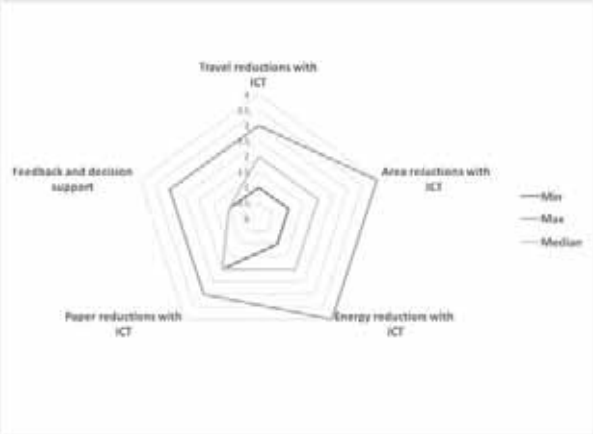


Figure 1. Example of radar diagram with minimum, maximum and median scores.

providing a summary of the evaluation session including possible actions for improvement. The Assessment Manager was responsible for forming the team, getting them to fill out the maturity scores and organising a date for the evaluation session. After the session the participants were asked to fill out a questionnaire on their experience with using the maturity model. As mentioned, the Assessment Manager was provided with a summary of the process that could then be used to create an action plan.

The evaluation sessions were facilitated through presenting the analyses of the individual scores combined. Each of the domains was briefly illustrated by showing a radar diagram with the minimum, maximum, and median scores (an example is included in Figure 1. Then the individual scores were shown as in Table 2, in which the scores that deviated by 1 point or more from the average were highlighted. Both strong deviations and the general averages were good starting points for the discussion. The sessions were recorded on video.

The questionnaire after the evaluation session was simple and brief in nature and consisted of the following questions:

- Did you find it useful to fill out the maturity model? Why (not)?
- Did you gain more insight in the possibilities of Green ICT?
- Were you more inspired to apply Green ICT?
- Will you from now apply more Green ICT in your daily routines?
- Do you have any feedback on the maturity model?

4. RESULTS

Four organisations participated in this study. They all had an assessment team of 4 - 10 members who all filled out the maturity scan individually. The averages for each organisation on each attribute are shown in Table 3.

Table 3. Overview of all the averages scores at each of the participating organisations.

	A	B	C	D
Green ICT in the organisation				
Green ICT Strategy	1,8	1,8	1,6	1,2
Governance of ICT services	2,6	3,8	2,6	2,8
Green ICT Procurement	2,1	2,6	2,1	1,8
E-waste Policy	2,2	3,2	2,3	1,4
Green ICT in Information Management and Architecture	1,7	2,0	1,4	1,4
Community collaboration	2,3	3,0	1,8	1,6
Green ICT Supply Chain Management	2,0	2,6	1,8	1,6
Greening of ICT				
Housing	2,0	2,2	1,8	1,6
Computing infrastructure	1,9	2,2	1,8	1,6
Network infrastructure	1,6	2,2	1,6	1,3
Storage infrastructure	2,0	2,0	1,6	1,3
End user ICT equipment	2,4	2,6	1,8	1,4
Software and ICT services	1,6	1,6	1,4	1,2
Greening of Operations with ICT				
Travel reductions with ICT	3,0	3,0	2,3	1,8
Area reductions with ICT	2,7	2,6	2,3	2,4
Energy reductions with ICT	1,8	2,0	2,0	1,4
Paper reductions with ICT	2,9	2,8	2,4	2,2
Feedback and decision support	1,4	2,0	1,7	1,2

We will qualitatively describe the results of the discussion to give an impression per domain.

Starting with the domain 'Green ICT in the Organisation': for all participants an in depth Green ICT Strategy document did not exist. Sometimes there was a short position paper on sustainability from an ICT perspective that connected to the general sustainability policies but those. Good governance of ICT services is something that is common practice for the participating organisations, even without environmental considerations. For three out of four participants policy documents for procurement and e-waste did not exist, but sustainable principles were sometimes applied. So usually there were some guidelines that were applied by people motivated to do so, but these were not broadly supported and communicated. Finally, three out of four deemed Green ICT in Information Management and Architecture, Community Collaboration (beyond exchanging information or good practices) and Green ICT Supply Chain Management as too advanced topics to grow in for now. Organisation B is the positive exception here with more mature scores that is reflected in their activities as well. They try to set an example for other organisations and talk to their suppliers and users about environmental concerns.

Considering the second domain, the Greening of ICT itself, the scores do not show big differences between A, B and C - the participants in organisation D said that they only just started and did not do much yet to apply energy and material efficient principles in their ICT equipment and use. Common practices amongst the participants were virtualisation

of server hardware, free cooling in data centers (where possible), applying sleep mode in equipment powered over ethernet (e.g. IP telephones), hot-cold-frozen data storage, pc power management and centralising equipment such as printers (shared instead of private use). On the other hand, participants found it difficult to balance reliability and availability (redundancy) with energy efficiency, which is reflected in not so green solutions for key elements in the ICT infrastructure such as network equipment. Monitoring on green KPIs is not very common as of yet, at max some general energy consumption figures were available.

For the third domain 'Greening of operations with ICT' there was much more variation in the scores. One organisation focused more on support for video conference solutions and teleworking, the other offered flexible office space and a third had banned all paper work as much as possible. Often, the financial department is having difficulties to transform to paperless. In terms of using ICT to reduce other energy consumption, this seemed quite difficult for all participants because there is not a strong connection between those responsible for ICT and those responsible for building and facility management. This is also something that is reflected in the final scores on Feedback and Decision Support where it would be beneficial to connect all the information systems that are used, analyse behaviour and return this as feedback to individual users as well as higher management. While the participants recognised the potential, they found it difficult to take action here other than to talk to other departments.

The results for the questionnaire are summarised over all participating organisations as they did not show any significant differences between the organisations. 93% of all individual participants found it useful to use the maturity model. In their comments they commonly wrote that it helped their awareness and that it gave insights in where their organisation stood. This is also reflected in their answers to further questions where again 93% had learned about new possibilities in using Green ICT, 73% was inspired to promote more Green ICT activity and 67% was going to apply Green ICT principles in their daily work. The feedback on the maturity model consisted of detailed comments that were used to improve the model and general comments that were on the whole positive. Some found it difficult to fill out or said that they needed the evaluation session to fully understand the model.

5. DISCUSSION

The goal of the SURF Green ICT Maturity Model is to help organisations learn more about the possibilities of Green ICT in a practical and efficient way. It gives an overview of the most important issues and areas that can be addressed with Green ICT. As demonstrated in the results, organisations gain insight in both the ICT-as-a-problem side as well as the ICT-as-a-solution side in their particular situation.

Looking at the things that were discussed in the evaluation sessions, the issues and solutions that participants talked about were quite similar for the four organisations. Almost always there was someone or a group of people who were giving good examples on how to procure more environmentally friendly or dispose of equipment properly. However, their methods were not formalised in policies and enforced and therefore this was a common suggestion for improvement. Similarly, communicating about what is already there, from policy documents to good practices was also a popular action to follow-up with. Especially in the 'Greening of ICT' domain, reduction possibilities were in place, but not taken advantage of by users. The same goes for ICT facilities such as videoconferencing to reduce traveling. Promotion of and communication about these solutions is a key activity for all participants (to different degrees). Since these are typical signs of bottom-up enthusiasm, the reverse may also be true that the main issue behind the above observations is a lack of top-down support. This was also sometimes mentioned in the discussions, especially related to Green ICT Strategy: with a strong and clear strategy it would become much easier to take the next step.

Looking at the most common improvement suggestions (for example: putting a green paragraph in project documents, implementing total-cost-of-ownership in procurement, virtualisation of servers, promoting teleworking, using power-over-ethernet solutions, improving asset management and monitoring) a pattern seems to emerge that organisations have started picking the low hanging fruits and are slowly moving beyond these. Some areas are clearly deemed to be too advanced to take on for now, such as green supply chain management and applying green principles in software and ICT services. Participants also often stated that any green action should go hand in hand with an improvement in either quality of service or financial sense. There is some room to be lenient (for example in extending the return on investment horizon) but doing something just because it is the right thing to do from an environmental perspective is difficult to achieve.

Finally, the questionnaire gave us insight in what the individual attitude of the participants was regarding Green ICT and the maturity model itself. The results of the questionnaire are quite clear: almost everyone found it a useful (and efficient) way to get an overview on where their organisation stands and quickly generate ideas for improvement. Furthermore, they increased their own awareness, learned more about the possibilities of Green ICT (beyond actions in the datacenter and also apply ICT as an environmental solution in business processes), and two third actually planned to apply Green ICT principles in their daily work. Even though this is a field study with limited participants, the general attitude is highly positive and seems a good indication of the positive effects the use of such a maturity model can have.

Furthermore, the model can be applied in an improvement cycle. Measuring the Green ICT maturity repeatedly in a standard way allows organisations to see whether they grow in maturity. For extra benefits, the maturity model could also be used as a benchmark tool to compare multiple organisations. It could be especially effective to identify best practices in this way and stimulate organisations to exchange

these amongst themselves.

6. CONCLUSION

The work described in this study addresses the issue of making Green ICT more practical for organisations. Often organisations know they need to make their data center more energy efficient by turning up the heat or thinking carefully about how they cool their equipment but they do not look beyond the data center. For Green ICT to be truly effective, organisations need to know how they can both reduce the environmental footprint of their ICT equipment as well as how to use their ICT equipment to help reduce the footprint of their other business activities.

The SURF Green ICT Maturity Model is designed to address exactly this issue. It helps organisations to identify areas for improvement in an efficient way. We illustrated this through a field study that followed four organisations who used to maturity model to gain insight in where they stand as well as to get inspired to take action. By using the model not only to get an overall impression of the maturity model Green ICT, but also to discuss the results with participants in an evaluation session, the value of the maturity model was clear to the participants. Both individuals and organisations as a whole, were very positive on the use of the maturity model. It increased awareness, inspired to take action and was insightful. A model such as the SGIMM can be used to benchmark amongst organisations, and become a standard way of measuring and addressing Green ICT.

Even though the results described here are limited to four organisations and we should be careful to draw general conclusions, the strong positive reception indicates the potential such a model can have for organisations to guide them in using Green ICT. Moreover, the participants were stimulated to consider actions to reduce the environmental impact of ICT as well as use it to enable improvements in other business areas.

For future work, we will consider following the participating organisations over a longer period of time to see if they actually follow-up on the results of their maturity scan. It would be interesting to see whether they would write an action plan and carry out improvement actions. After a year or so, the organisation can then repeat the maturity scan to see where they stand compared to the year before. Such a comparison would give insight in whether the use of the model is actually effective in improving the maturity and therefore the greenness of an organisation.

REFERENCES

[1] S. Murugesan, "Harnessing green IT: Principles and practices," *IT Professional*, vol. 10, no. 1, pp. 24–33, Jan 2008.

[2] "#SMARTer2030, ICT Solutions for 21st Century Challenges," 2015.

[3] Frans Berkhout and Julia Hertin, "Impacts of information and communication technologies on environmental sustainability: Speculations and evidence," *Report to the OECD, Brighton*, vol. 21, 2001.

[4] Andreas Köhler and Lorenz Erdmann, "Expected environmental impacts of pervasive computing," *Human and Ecological Risk Assessment*, vol. 10, no. 5, pp. 831–852, 2004.

[5] Alemayehu Molla, Vanessa Cooper, and S. Pittayachawan, "It and eco-sustainability: Developing and validating a green it readiness model.," in *ICIS 2009 Proceedings*, 2009.

[6] E. Gelenbe and Y. Caseau, "The impact of information technology on energy consumption and carbon emissions.," *Ubiquity*, vol. 1, 2015.

[7] Balakrishnan Ramesh Babu, Anand Kuber Parande, and Chiya Ahmed Basha, "Electrical and electronic waste: a global environmental problem," *Waste Management & Research*, vol. 25, no. 4, pp. 307–318, 2007.

[8] Ecofys, Quantis, and BIO Intelligence Service, "Ict footprint: Pilot testing on methodologies for energy consumption and carbon footprint of the ict-sector," 2013.

[9] Gartner, "Introducing the gartner green and sustainable it infrastructure and operations maturity model," 2013.

[10] Alemayehu Molla and Vanessa Cooper, "Green it readiness a framework and preliminary proof of concept," *Australasian journal of information systems*, vol. 16, no. 2, 2010.

[11] UK HM Government, "Uk hm government green ict maturity model," 2013.

[12] S deMonsabert, Khuloud Odeh, and Jenny Meszaros, "Sustainabits: A framework and rating system for sustainable it," in *Green Computing Conference (IGCC), 2012 International*. IEEE, 2012, pp. 1–9.

[13] Brian Donnellan, Charles Sheridan, and Edward Curry, "A capability maturity framework for sustainable information and communication technology," *IT professional*, vol. 13, no. 1, pp. 33–40, 2011.

[14] Albert Hankel, "Surf green ict maturity model," 2015.

[15] Mark C Paulk, Bill Curtis, Mary B Chrissis, and Charles V Weber, "Capability maturity model, version 1.1," *Software, IEEE*, vol. 10, no. 4, pp. 18–27, 1993.

[16] Albert Hankel, Lisa Oud, Maiko Saan, Patricia Lago, et al., "A maturity model for green ict: The case of the surf green ict maturity model.," in *EnviroInfo*, 2014, pp. 33–40.

MOBILE SIGNAL EXTENSION IN DEEP SEA - TOWARDS A SAFE AND SUSTAINABLE FISHERIES

Dineshkumar Singh[1], Sanjay Kimbahune[1], Veerendra Veer Singh[2]

[1]Innovation Lab – Mumbai, Tata Consultancy Services Ltd.,Yantra Park, Thane (MH), India
[2]Mumbai Research Center of CMFRI, Old CIFE Campus, Versova, Andheri, Mumbai (MH), India

ABSTRACT

Despite having one million active fishermen, India lacks a scalable boat-to-shore communication framework. Small boat operators, which form 80% of entire fishing community, are more vulnerable to the wilder nature of the ocean, due to dynamic sea current, wind speed, direction and wave heights, etc. This is also one of reason for fishermen entering accidently in other country's territory, getting caught and sometimes costing their life. This poses a serious question to fishermen and their family, whether they would return the next time they venture out in to the sea.

Deep sea mobile connectivity and Information, given in advance about the sea conditions such as wind speed and wave height can help fishermen to plan their trip effectively and avoid the danger zone. This infrastructure could provide path of building "knowledge society" to fishermen deprived of real time and usable information. For realizing this, integration of the existing communication technologies like VHF transceivers, satellite communication and extension of mobile communication network is needed.

A consortium of partners like Central Marine Fisheries Research Institute, Tata Consultancy Service and Tata Tele Services piloted a project for mobile signal extension into the sea upto 30kms. This enabled an extended network accessibility, which led to create an Early Warning System over a digital highway of 30 by 120 KM, along the coast.

This paper captures the experiences and recommendations derived from this pilot.

Keywords— fisheries, mobile, deep-sea, communication, EWS, security.

1. INTRODUCTION

India, with a fish production of 9.58 million MT per year, is the second largest producer of fish worldwide and constitutes 5.68% of the world's total fish production [1]. Marine fisheries accounts for approximately 39% of the production. This sector contributes 0.83% to the national Gross Domestic Product (GDP) and ˜4.6% to the agricultural GDP [2], but is a crucial component of the rural coastal economy in 3202 villages across 8,118 km coastline. 0.9 Million fishermen are employed full time, followed by another 3.5 million gaining livelihood in the processing and marketing sector. About 61% of the entire population is below poverty line, with annual earnings less than Rs 25,000 (US$570) [3]. Majority of them are less educated and have education only between zero and upto 10th std. Hence, it is challenging to find an alternative livelihood. The situation becomes more difficult during the lean period or monsoon ban period. These factors lead to dependency on the local moneylenders to make ends meet and for emergency needs [4].

Currently the family structure is slowly moving towards nuclear family, which involves the husband, wife and children. Male person is usually the primary bread-winner. Hence his safety is of utmost importance for well-being of family. Women primarily work in the fish processing industry like sorting, grading and local sell.

In 1999, the International Labor Organization estimated that every year, approximately 24 million non-fatal injuries occur to fishermen [5]. Owing to wind speed, direction and wave height, the sea conditions keep changing unexpectedly and become dangerous which increases the vulnerability for those at sea. They could get caught unaware due to the sudden change in weather conditions which would make controlling the small boat difficult [6]. The situation is dangerous, especially when the boat is laden with the catch leading to capsizing of the boat and drowning the occupants.

During 1891 to 2000, 308 cyclones have hit the coastal population [7], endangering the lives of 370 million within the 100kms off the coast. This necessitates the design of a multi-technology communication framework to facilitate an Early Warning System (EWS) for the fisher folks and the other coastal population.

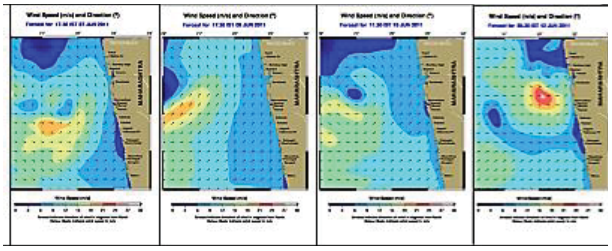


Figure 1. OSF – Wind Speed and Direction. Red indicates High wind speed (danger zone) and its movement over 36 hours

2. BACKGROUND

Indian National Centre for Ocean Information Services (INCOIS) generates Potential Fishing Zone (PFZ) and Ocean State Forecast (OSF) Advisories (fig. 1) for Indian coastal region [8]. It processes the Chlorophyll and TSM images received from Oceansat-2 and Sea Surface Temperature (SST) images from NOAA-18/19 and MetOp-A/B satellites. These images are further processed for geometric correction, filtering and are geo-coded. The PFZ map helps fishermen to plan their fishing locations [4].

OSF services display the temporal and spatial aspects of the sea using Wind Speed and Direction, Swells and Wave height parameters (fig. 1). These images are generated four times a day, five days in advance. These images show the regions of sea in different colors. Each color indicates the intensity of the wind speed or the wave height. Higher the intensity, the more dangerous the zone will be. Hence, availability of this information in advance helps fishermen

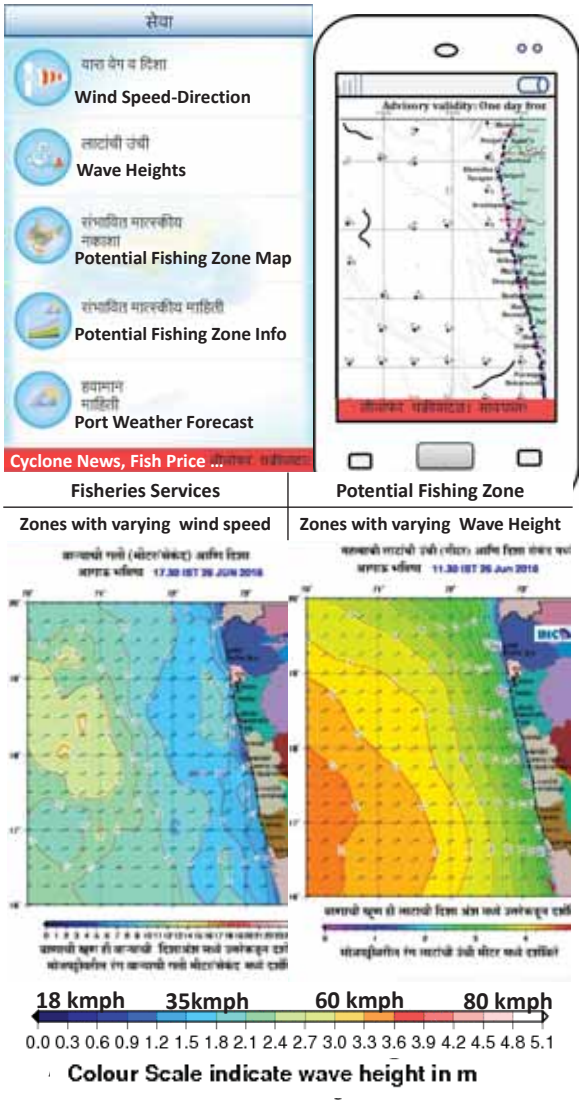


Figure 2. mKRISHI Fisheries service with PFZ, Wind Speed and Wave height services

to get an idea of the state of the ocean at any given point of time and forms the basis of the EWS (Fig. 2).

But there were challenges associated with the distribution of this information to the fisher population. Hence, mKRISHI® Fisheries, a Java (J2ME) and Android mobile based application (Fig. 2) were developed for the fishermen [9]. This mobile app translated the information in the local regional language (Marathi) in Maharashtra State of India. The services were piloted in Raigad district of Maharashtra (India). This app helped spread the services quickly in the region and was accessible to a large population.

Advance information helped fisher to identify the PFZ location, estimate the trip duration, and accordingly budget the fuel, ice and labor requirement. Wind speed image (fig.2) helped them in finding out the areas where sea is calm and where it is turbulent. This brought more discipline in the trip planning and risk assurance, leading to savings in diesel, cost and life. Planning the boat cruise, in the direction of the wind, reduced the drag on the boat and hence improved the diesel efficiency. This further resulted in the saving of fuel. Pre-planning can help save diesel, reduction in CO_2 emission and cost of catch too. This can make the fishing safer and sustainable [9].

During fishermen group discussions (FGDs) the fishermen highlighted that though the information helped them to plan a safe trip and saved diesel, they also complained that the information could not be updated while they were at deep sea. This was especially true for the multi-day trawlers going for up to 10 to 15 days of fishing trip. The mobile service network was available only up to 5-8 kms in the sea and hence, it was difficult for them to get the latest information while they were away in the deep sea. This effectively reduced the utility of the EWS system.

As shown in fig. 1, during the 48 hours period between 10th and 12th June 2011, the wind speed near Raigad coast, increased from 27 kmph to 81kmph. Even at a speed of 20-25 kmph a small four cylinder boat can behave like a toy in the ocean. This caused the sinking of two boats, 20kms from the coast. An investigation in the incident concluded that the boats were loaded with the catch and were enroute to the landing center. Unaware of the sudden development, they were caught in the turmoil at sea. A EWS with the extended mobile network in sea could have helped to replan their route or navigate towards a calmer area in the deep sea. This incident led to the research on the extension of the wireless signal (preferred mobility) in deep sea to support the mKRISHI Fisheries services.

3. MATERIAL AND METHOD

We studied three popular modes of the ICT techniques (fig.3).

1. Wireless (VHF) Transceivers (Walkie-talkie)
2. Satellite based (Satellite phones and Direct To Home (DTH))

3. Mobile phones (cellular technology)

We interacted with the equipment providers, service operators, users (fishermen) and the regulators to understand the science behind it, ease of use, affordability, constraints, range it can be relayed to, cost of the equipment and the operational costs. Brief analysis is as below:

3.1 Wireless set / Ultra or Very High Frequency (U/VHF) Transceivers

A few big fishermen use the VHF/UHF portable two-way radio handsets (also called Walkie – Talkie). It is a type of radio communication in VHF frequency band (operating in 30 MHz to 300 MHz frequency band). It requires license from the Ministry of Communication to own and operate maritime mobile communication and electronic navigational equipment on designated Maritime frequencies on board Indian ships or boats. The walkie-talkie, generally works in a radius of 32-35 km from the communication towers which relay the signals. Actual distance range available is subject to the terrain conditions. The major drawback of this technology is: generally one-way communication (at a time), need for the license, specific equipment, unlicensed operators, can cause signal interference to licensed stations, mainly voice and limited text and no graphics/image.

3.2 Satellite based (Satellite phones and Direct To Home (DTH))

Satellite phones, are the mobile phone that connects to geo-stationary satellites 500 to 1,000 miles above the earth. It needs a direct line of sight to the satellite. Sat phones are rarely affected by violent storms. The calls originated from the sat phone, route through multiple satellites, before connecting to the appropriate ground station, which then transfers the call to the public voice network or Internet.

Major challenges with satellite phones

- High call cost ($2 to $15 per minute).
- License is required under Telegraph Act, IT Act and Wireless Act, which has become tougher to get, post 26/11 terrorist attack in Mumbai (India).

Major challenges with DTH type of communication:

- Satellite based DTH Service, could be an alternative, but transmission is only one way and in narrow band – which is costly.
- In Ku Band Antenna position stabilization at sea is a big challenge.

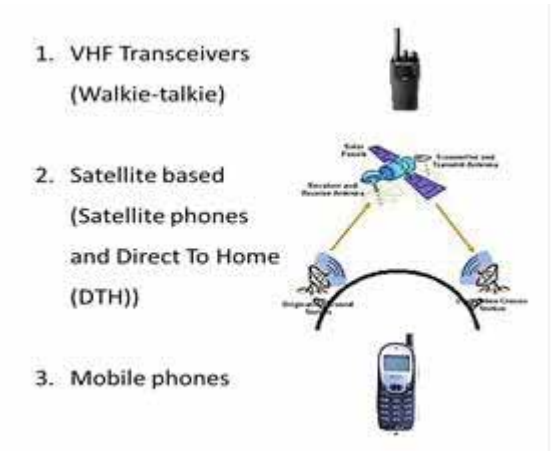


Figure 3. popular modes of the ICT techniques

3.3 Mobile phones (cellular technology)

As per Telecom Regulatory Authority of India (TRAI) report in Jan 2015, India has a tele-density of 76%, with the total subscriber base reaching 957.61 million. About half the population (580 million Indians) will be online by 2018, including people from all age groups, women and the rural population.

Two popular cellular technologies are: Global System for Mobile communications (GSM) and Code Division Multiple Access (CDMA). GSM is the most common networking standard for mobile phones in the world, whereas CDMA typically has the best sound quality and data service. General Packet Radio Service (GPRS) is an enhancement to GSM networks and gives data rates between 30 Kbps and 90 Kbps.

Mobile phone services are matured and multiple telecom service providers provide the services at an affordable price. Equipment accessories (mobile handsets) and software (mobile Apps) ecosystem is also well established. The industry is young and the technology is growing at a very high speed, enabling more computing, performance and features at highly reduced price. Hence, this seems to be the best fit technique to extend the communication link in the deep sea. Since it is already available on the land, extending the network to the sea could provide a seamless access. It provides both voice and data network and hence many more services can be developed and offered. This will also have no learning curve and would also benefit the existing investment in the people (to support the network, the customer care, would offer the services at retail/rural level, service center), the handsets, etc.

This would also mean that all the research and development activities happening in the mobile arena (such as Android apps, services like mKRISHI Fisheries, What's App, etc.) can also be easily extended.

The major challenges are the high capital expenditure involved in installing the towers and offer the signal towards the sea. Telecom operators do not see a business in "seeing towards the sea" as Return On Investment (ROI) is challenging,

Though business model was a challenge but the need was critical .we approached various telecom (mobile) operators

with this idea of expansion of the mobile signal in deep sea. The response was not encouraging from most of the operators. Eventually Tata Teleservices Ltd. (TTL) agreed to carry out the pilot along the coast of Raigad district of Maharashtra looking at societal benefits.

4. EXPERIMENT AND RESULTS

The entire mobile signal extension testing approach was divided into four phases:

- Proof of Concept (PoC) and Boat Test Drive
- Selection of cellular technology
- Tower installation and commissioning
- Parametric test based detailed study of major operator's signal

4.1 Proof of Concept (PoC) and Boat Test Drive

TTL had spectrum license for CDMA 850 MHz and GSM 1800 MHz in Raigad. TTL modified the orientation of two sectors of the antenna on one tower to direct GSM and CDMA signals towards the sea and a signal measurement boat test drive was done.

4.1.1 Boat Test drive 1 – Baseline study

The RF engineers, fishermen and the fisheries research scientists carried out 11 hour test drive in the deep sea in a fisher boat. INCOIS Wind direction and speed map were used to plan and guide the journey. The Southward journey was done almost in 15-30 KM range from the nearest shore, whereas northward journey was done almost in 5-12 KM range from the nearest shore (fig. 4).

Following equipment was used for the test drive (fig. 5):

- CDMA and GSM Tough Book for automatic voice and data call testing
- GARMIN GPS
- Mobile Phones (Voice and Data calls)
- Automatic Trackers, supported by Manual observations (Call setup, Quality, data speed, completion, etc.
- Log books

GSM and CDMA Transmitter (Tx) and receiver (Rx) band signal power strength was plotted along the geo-coordinates (Fig. 6). This helped give an idea of the mobile signal strength at various locations during the trip.

Also, the initial tests revealed that even if a dedicated tower is setup, which radiates the signal towards the sea, the line of sight is a major limitation (25-30 KM) because of the earth curvature

4.1.2 Boat Test Drive 2 – GSM and CDMA comparison

Figure 4. Boat Test Drive 1 - Geo path

Figure 5. Mobile signal testing at sea

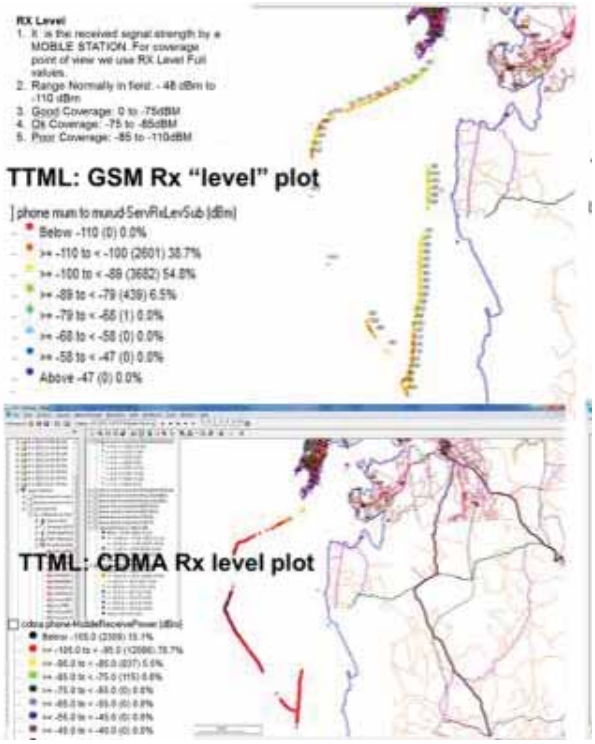


Figure 6. Boat test drive results - call quality

The RF Engineers from the TTL Network planning team adjusted the power level, height and the direction of the antennas at two different locations (towers) (see Results in

Fig. 7). This helped to narrow do wn the testing parameters such as technology, frequency band, tower height, etc.

4.1.3 Boat Test Drive 3 – BTS configuration

Encouraged by the above analysis, TTL decided to carry out a detailed technology comparison pilot. It installed CDMA Base Station Transceiver System (BTS) at two sites in Raigad district i.e. at Sasawane and Varsoli and a CDMA + GSM BTS at Borli-Mandla site.

The boat test drive started from Mandawa port (north point of Raigad) in the boat *Dronagiri Mauli* and journeyed south-west cruising at 10kmph. We reached towards a spot in the sea, 29km from Borli landing center. From there we returned to a distance of approx. 15km from the coast and moved towards Mandawa. During the journey, two tough books kept making automatically latching to nearest BTS in the region and measured the call and data quality. The results have been summarized in the Table 1 and plotted in the fig. 8 and 9.

This drive helped narrow down the choice of the cellular technology and type of the BTS to be used to achieve the maximum distance in the sea.

4.2 Selection of the cellular technology

While comparing both the technologies, CDMA technology has certain advantages over GSM technology, as listed below:

- The link budget comparison indicates that the coverage reach of CDMA will be higher than the coverage provided by GSM technology due to lower band i.e. 850MHz.
- CDMA technology employs processing gain concept which enhances the coverage.
- CDMA technology has superior quality by virtue of soft handoff.
- The data speeds are higher in CDMA technology because of its robust designing, this will help enhance the access of mKRISHI® Fisheries app.
- CDMA is less prone to interference In GSM, managing interference in open area like sea is a big challenge
- In CDMA, more users can make simultaneous calls compared to GSM

Table 1 Boat Test Drive 3 - Test Results

CDMA TRIAL SETUP	GSM TRIAL SETUP
CDMA 850 MHz	GSM 1800 MHz
Coverage reach for CDMA (Voice + Data) – approx 25 km	Coverage reach for GSM (Voice + Data) – approx 16 km
This is with conventional solution –	This is with un-conventional solution –
Standard BTS – No Tower Top solution	Tower Top BTS. .
Normal Gain antenna used	Normal Gain antenna used
No special feature used	Special feature PBT deployed
Incidental coverage beyond 25km may be possible	Incidental coverage beyond 16km may be possible
The coverage reach can be further enhanced, with Tower Top BTS, High gain antennae (20.5 dB gain), to 30+ km	The coverage reach can be further enhanced, with High gain antenna (20.5 dB gain), to 20+ km

Latitude /Longitude	Time	Depth (Meter)	Distance (Km)	CDMA Signal	GPRS availability
18° 30.279 N 72°54.445 E	7.20 am	-	-	Very Good	Very Good
18° 30.216 N 72°52.272 E	7.48 am	4.26	5	Very Good	Satisfactory
18°30.674 N 72°49.981 E	8.19 am	7.6	10	Very Good	Satisfactory
18°31.303 N 72°48.447 E	8.46 am	9.1	15	Satisfactory	Unavailable
18°31.688 N 72°46.262 E	9.23 am	16	20	Satisfactory	Unavailable
18°31.983 N 72°46.262 E	10.02 am	21.6	25	Satisfactory	Unavailable
18°32.105 N 72°43.188 E	10.15 am	23.4	29	Unavailable	Unavailable

Figure 7. Boat Test Drive 2 for GSM vs. CDMA test

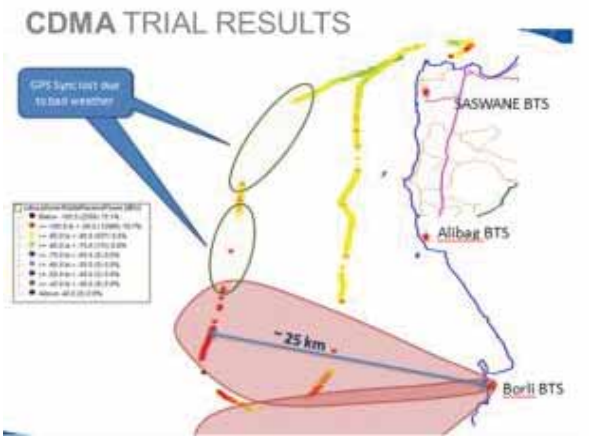


Figure 8. Boat Test Drive 3 - CDMA tower call results

Figure 9. Boat Test Drive 3 - GSM tower call results

Table 2. Boat Test Drive 4 Results

Antenna Height	Distance covered Signal level > -95dBm	Remarks
35 ft.	17 KM	Observed
60 ft.	24 KM	Model prediction
100 ft.	35 KM	

4.2.1 Boat Test Drive 4 – Different Telecom operator with dual technology

Another test drive was done with the government telecom service provider BSNL, which had two GSM Towers (900 MHz) at Saswane and Borli Mandla, both at a height of 35 ft. These towers were approx.150-300 m from the coast. BSNL modified the antenna to direct one sector towards Arabian Sea.

Mobile signals originating from these towers had good reception up to 17KM from the coast. The data was fed to a modeler and based on the pattern learnt by it, it projected that a tower height of 100 feet can give mobile signal range of up to 35km. (see table 2)

An interesting pattern was observed from a tower situated 35KM from the signal test point (in the sea). It was found that the Chirner tower (LAT 18.8675deg, LONG 73.07471389deg) was situated at a hill of Mean Sea Level (MSL) height of 240 m. The antenna height of the BTS was 30m. Hence, due to a height of the 270 m, the signal was observed at the distance of 35KM (fig. 10).

The data collected helped in analysis on what parameters can affect the range of the signal availability. Some of the parameters are:

- Mobile technologies (GSM vs. CDMA)
- Spectrum availability (800KHz vs 1800 KHZ)
- Height of the tower on which the BTS (Base Station Transceiver Sub system)
- Type of the BTS
- Feed antenna length, Type of antenna
- Antenna angle, Antenna radiation power level
- Type of handsets

4.3 Tower installation and commissioning

TTL surveyed the Raigad districts mobile towers and picked up 7-8 sites where work needs to be carried out to expand the signal up to 30+ KM in the sea (fig.11). To create a mobile network offering services towards sea, they needed to create “physical structure (called tower)”, install the BTS , antenna , power supply units, and other network components on it. Building new towers could have escalated the cost and hence, TTL searched and found six existing towers which fairly met the required specification (such as 200 to 500 m from the coast, 40 m height or so,

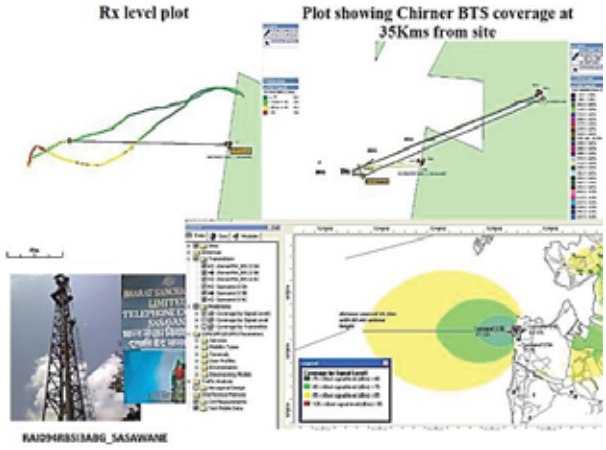


Figure 10. Boat Test Drive 4 Results

etc.). On these towers, following configurations were done to increase coverage.

- Tower Top BTS (Feeder less)
- Std 33deg antenna – High Gain ~20.5 dBi

In addition to this, a new tower was erected and equipped with the following equipment to achieve a continuous 30km mobile signal reach along 120km coastline of Raigad.

- Tower erection, Installation of the tower top BTS Or Remote Radio unit (RRU)s and Feeder cables
- Antennas – MRFU V2 900MHZ CDMA (SUPPORT 4 TRX, 20W TOC,25MHZ IBW)
- Power, Diesel Generator, Microwave Link
- Installation and commissioning
- IP Clearance, Site RFI, WPC certification

4.3.1 Investment for a self-sustainable

An estimated 13.7million Indian Rupees was needed to modify and build these seven sites to start the mobility signal up to a range of 30KM or so. This service benefited approximately 11,620 fishermen families in Raigad and hence, this investment translated to Rs. 98 (i.e. $2 USD) per month per family for a year. This gives a very positive reason on why these investments can break-even in a year and can become self-sustainable within two years.

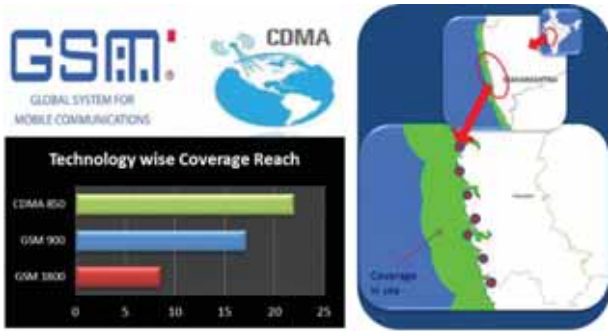


Figure 11. Planned locations of towers and "digital highway" in sea

Table 3. Installed tower locations and details

Site Name	BTS ID	Sector #	Latitude	Longi-tude	Ht. (ft)
Sasavane	380	3	18.7874	72.8654	45
Varsoli_Alibaug	381	3	18.66678	72.86875	35
Borli-Mandla	382	3	18.5105	72.9132	50
Murud_Raigad	387	2	18.3373	72.9591	60
Borli Panchtan	384	3	18.1627	73.0019	55
Shreevardhan	385	2	18.0517	73.0263	50
Aravi	386	2	18.0848	73.0062	50

4.4 Parametric test based comparative study of major operators' signal in same region

It was decided to conduct a comparative study of the installed service's quality, coverage and data speed with the other major service provider in the region. Three test drive kits were prepared to capture TTL CDMA, Vodafone and Airtel GSM network coverage parameters. The handsets and the laptops were connected. The entire setup was connected to the inverter powered by a battery. Boat drive for the field testing started from the Revdanda jetty (which is in the range of Borli Mandla tower and has Alibaug tower in North and Murud tower in South). We cruised towards deep sea in west direction for approx. 32 km. From that point, we cruised towards south approx. 24 km and finally back to jetty travelling approx. 30 km diagonally north-east.

The link budget modeler estimated the coverage upto 25 km, but the actual ground (sea) data indicated the signal availability upto 32+ km in the sea. Beyond the distance of 30-35km, the coverage is restricted by the earth's curvature based on tower/antenna height and line of sight visibility.

Following sections cover the comparative study across three major parameters: coverage, quality and data speed. For certain drive routes, network details were not captured for GSM operators either due to GPS synchronization failure or device issue. So, to bring the parity in the comparison and carry out apple to apple comparative study, the statistics and plots of drive routes which are common to all 3 operators are considered. The individually exclusive details are not captured while comparing with other operators.

4.4.1 Coverage Comparison

Parameters Compared:

- TTL CDMA – MTx Level
- Vodafone GSM – Rx Level
- Airtel GSM – Rx Level

In CDMA, MTx Level produces true coverage indicator while comparing with GSM coverage. It will have better correlation than Rx Level under unloaded condition. It was observed that TTL CDMA coverage is continuous whereas the coverage plots of Vodafone and Airtel seem to be patchy and discrete.

4.4.2 Quality Comparison

Parameters Compared:

- TTL CDMA – Ec/Io Level
- Vodafone GSM – AMR C/I Level
- Airtel GSM – AMR C/I Level

These parameters capture the network and call quality considering the degradation due to interference. Additionally, FER parameter of TTL CDMA compared with Rx Qual (sub) parameters of GSM operators. Rx Qual (sub) is considered to be crude way to indicate quality as it will reflect only idle scenarios. It was observed (fig. 12) that TTL CDMA signal has superior quality compared to Vodafone and Airtel (fig. 13). Vodafone signal quality seems to be deteriorating as the distance increases from the shore. Airtel signal quality seems to be patchy and degraded.

4.4.3 Data Speeds Comparison

Parameters Compared:

- TTL CDMA – Ec/Io Level
- Vodafone GSM – AMR C/I Level
- Airtel GSM – AMR C/I Level

These parameters captures the network and call quality considering the degradation due to interference. Additionally, FER parameter of TTL CDMA compared with Rx Qual (sub) parameters of GSM operators. It was observed (fig. 14) that TTL data speeds are higher due to selective deployment of EVDO technology, which is an evolution of CDMA technology (comparable to 3G) to cater to optimum data speeds. Higher data speeds improve

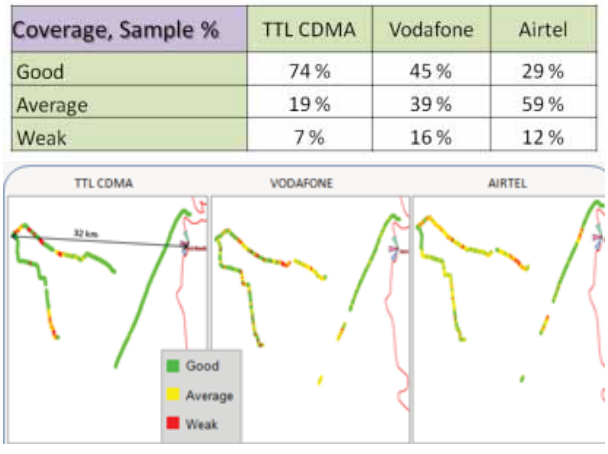
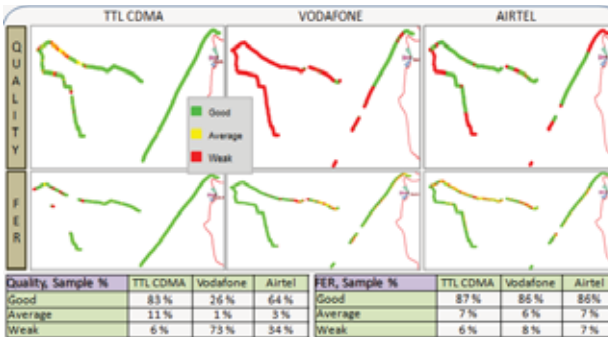


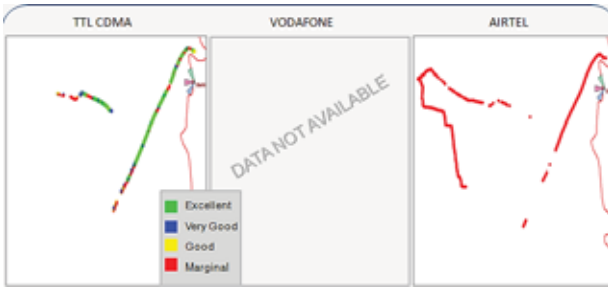
Figure 12. Signal coverage



Quality, Sample %	TTL CDMA	Vodafone	Airtel
Good	83 %	26 %	64 %
Average	11 %	1 %	3 %
Weak	6 %	73 %	34 %

FER, Sample %	TTL CDMA	Vodafone	Airtel
Good	87 %	86 %	86%
Average	7%	6 %	7 %
Weak	6 %	8 %	7 %

Figure 13. Signal quality



Data Speeds, Sample %	TTL CDMA	Vodafone	Airtel
Excellent (> 256 kbps)	70 %	NA	0 %
Very Good (128 to 256 kbps)	10 %	NA	0 %
Good (64 kbps to 128 kbps)	6 %	NA	0 %
Marginal (< 64 kbps)	14 %	NA	100 %

Figure 14. Data Speed

the customer connectivity experience. Airtel data speeds are lower as it works on GSM (2G)

Above comparison is based on the boat drive test covering only 3 sectors (out of total 19 sectors) installed by TTL. A periodic test drive can help optimize the signal strength further.

5. CONCLUSION

From the comparative study it was observed that the pilot provided a good quality mobile signal coverage upto 30-35km in deep sea and a high data rate. This provided an extended accessibility to the ICT network which helped enable a better use of the technology for a better livelihood and safety management. Fishermen used the extended signal to call for help when a boat was overloaded with the catch and could have toppled in the sea. This helped in arranging another empty boat and transfering the excess load thus creating a shared opportunity for all. In another instance, fisher with the large catch quantity, could connect with the local traders at the landing center and negotiate the price. This helped to preserve the quality of the fish and adequate ice, labor and truck could be arranged, even before the catch landed at the sea. A combination of the wireless VHF sets and mobile phones provided a further extended extension in deep sea, upto 60-70 km.

This pilot experience presents a model for extending the early warning system using ICT in deep sea.

6. ACKNOWLEDGEMENT

We are thankful to Dr. Arun Pande to conceptualize this entire initiative. We will like to thank TCS management comprising of Dr. Srinivasu Pappula, TCS CTO K. Ananth Krishnan and VP K padmanabhan and NAIP and World Bank leaders, especially Mr. Ranjan Samantaray for his support. We are also greatful to our partners Dr. S.S.C. Shenioi, Mr. Nagaraja and others at INCOIS and Mr. N. Srinath and Neeraj Dindore from Tata Teleservices to provide required support for the testing.

7. REFERENCES

[1] Gopalakrishnan, A., and Grinson George. "Marine Fishermen: Experiences and Expectations." *Blue Waters* 17.1, pp. 12-14, 2016.

[2] Handbook on Fisheries Statistics 2014, pp. 2, 2014.

[3] "India Marine Fisheries Issues, Opportunities and Transitions for Sustainable Development", Agriculture and Rural Development Sector Unit South Asia Region, World Bank, pp.11-19, Aug 2010.

[4] R Sathiadhas R., K. K. P. Panikkar, and A. Kanakkan. "Traditional fishermen in low income trap—A case study in Thanjavur coast of Tamil Nadu." Marine Fisheries Information Service, Technical and Extension Series 135, pp. 5-10, 1994.

[5] "Safety for Fishermen: The Way Forward", FAO, Rome, pp.17, Sep 2010.

[6] Anabel, Nancy. "Enhancing Lives and Livelihoods of Fisher folk through Innovative Technology applications." 2013.

[7] Chaturvedi, Sanjay, and Vijay Sakhuja. Climate Change and the Bay of Bengal: Evolving Geographies of Fear and Hope. ISEAS-Yusof Ishak Institute, 2015.

[8] Nayak, Shailesh, T. S. Kumar, and M. N. Kumar. "Satellitebased fishery service in India." The Full Picture, pp. 256-257, 2007.

[9] Kimbahune, Sanjay, V. V. Singh, Arun Pande, and D. Singh,. "ICT for fisheries—Environment friendly way: Pilot experience in Raigadh." In 2013 Annual IEEE India Conference (INDICON), IEEE, pp. 1-8, 2013.

[10] Dineshkumar Singh, Veerendra Veer Singh, Arun Pande, and Sanjay Kimbahune, "Approach and challenges for deep sea communication for fishermen –an experiment", 1st World Ocean Science Conference 2015, Mar 2015.

[11] "Can't force operators to give coverage in sea: DoT to Defence ", *Economic Times*, India, 2012.

HUMAN SAFETY CONSIDERATIONS IN THE EMERGING ICT ENVIRONMENT

Shailendra K. Hajela

ABSTRACT

This paper deals with the challenges posed to human safety in the emerging IoT and ubiquitous network scenario by super intelligent machines/ cognitive robotics that would coexist in the upcoming heterogeneous world comprising humans and super intelligent machines. Questions of trust in the super intelligent machines have been raised by scientists, sociologists, industry stalwarts, etc. After deliberating on the issues, aligning AI to human interests has been suggested as a plausible option. Agenda for further action proposed is that the ITU may constitute a multidisciplinary umbrella group, collaborating with existing technical and policy groups working on AI standards at technology and policy levels, for in-depth study of all facets, bridging the technology innovations and the policy considerations, and bring out Guidelines for AI aligning with human interests adhering to ethical code.

Keywords— Human safety, IoT, Trusting Robots, Cognitive informatics, AI Rules for super-intelligent machines

1. INTRODUCTION

Rapid advancements in the fields of Internet of Things, artificial intelligence, super smart machines, cognitive informatics, cognitive robotics, soft computing, and related areas have aroused interest as well as concerns about the emerging interactive man-machine heterogeneous environment, where autonomous super intelligent machines/robot working solo or teaming with, can match and even surpass humans in speed and precision. No longer is this a matter of science fiction when Isaac Asimov in 1942 made the 3 (technically 4) laws to assign paramount importance to human safety [1]:

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm;
2. A robot must obey the orders given it by human beings, except where such orders would conflict with the First Law;
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws.
0. A robot may not injure humanity, or, by inaction, allow humanity to come to harm.

Adaptations and extensions exist based upon this framework. But these have for long remained in the realm of science fiction. The developments in the field of AI, Robotics, Cognitive Informatics, IoT (Internet of Things), and related fields warrant their no longer being considered fictional. A super intelligent machine, say, a cognitive robot is designed to perceive, analyze, reason out and decide its action based on its past experience and predictive knowledge of the results of such action. In a situation, a robot is flooded with a constant stream of perceptual data that can be incomplete, and information and insights reached in a fuzzy way. In order to use this information effectively for determining actions, a robot must have the ability to categorize its experience. Based on these categories, a robot must be able to predict how the environment will change as a result of its actions. Most importantly, this process of development should be driven by an intrinsic motivation to explore the categories of its experience where it can make the most learning progress.

The ITU in its report [2] on IoT (2005) identified the field of robotics as one of the starting points for the creation of a world of smart things, which has since experienced steady growth. During the past decade, the most important factors affecting the rate of growth of the robotics market have been their enhanced capabilities and functionality. The market size of industrial robotics initially has been greater than the market size of personal and service robotics. However, according to the report, the personal robotics segment is likely to lead future market growth. This would include a wide range of different kinds of robots to do house chores, industrial jobs, and personal care giver partner-robots. In Japan, one of the main drivers of the personal robotics market is a rapidly aging population – by 2050, more than a third of all the population is estimated to be 65 or over in age, creating a lucrative market potential for the elderly-care robots, according to the report. Super smart Robots that can also express emotions are already in the market place.

It may be noted that in this paper, Artificial Intelligence (AI) is assumed to be embedded or otherwise leveraged in the super smart machines and robots.

Some years ago, Google reported [3] developing an intelligent machine, an autonomous self-driving vehicle navigating itself through other vehicular traffic and hazards. The resulting boon to society due to its artificial intelligence that would ensure adherence to traffic rules and impart faster navigational and decision making capability to avoid accidents would be enormous. It would eliminate driving fatigue, instant human judgment and navigational errors and would make the commute safer to arrive precisely at the set out destination. While one could feel happy for the truck drivers who drive long hours to ferry goods thousands of kilometers, one would be concerned with its turning into a killer-on-wheels should an inadvertent malfunction occur or due to some rogue person managing to successfully intrude and take over its control over the Internet. Not to forget that the truck drivers role could change or even be eliminated. With the IoT, it also gets its identity established and can be accessed. Due to

this, it raises the risk of cyber security breaches and other cyber forensic issues and the legal concerns about fixing responsibility and accountability in the event of such a vehicle meeting with an accident.
In this converged environment with super intelligent thinking machines endowed with reasoning, determining their own action/reaction in a given situation with fuzzy knowledge of the result of their action, interacting with humans, the abstractions of virtual world changing to reality poses a great challenge to dominance and safety of human society. There appears the need for aligning the super intelligent machines to the ethical codes that incidentally in humans also developed in the process of civilization over centuries as a result of philosophical thought, moral codes of religions in the west and east, renaissance, social reforms and philosophers like Immanuel Kant who influenced greatly the West. Drawing a parallel from human history and civilization, the scientific community engaged in AI/cognitive informatics/cognitive robotics has the opportunity to design and build robots with characteristics that ensure their coexistence in the real world in peace and harmony with humans and collaborative functionality.

2. PRESENT SCENARIO

The present century is driven by Information and Communications Technology (ICT), which involves integration of telecommunications and computer networks being at the heart of all major systems such as energy systems, transportation and communications forming the infrastructure of our living environment. Smart cities of tomorrow envisage urban development vision to integrate multiple ICT and Internet of Things (IoT) solutions in a secure fashion to manage a city's assets – banks, schools, libraries, transportation systems, hospitals, power plants, water supply networks, waste management, law enforcement, and other community services. Smart cities embody smart systems, and AI methodology is at their heart with the main objective to improve the contact between citizens, service providers and government with a view to improve transparency, efficiency, productivity and the quality of life of citizens.

Ethics developed concepts of right and wrong for morals and social conduct of humans living in a civilized society. In earlier days the moral philosophy, religion and culture used to provide guidelines for acceptable social behavior for humans by prescribing do's and don'ts. In the current day, governments make legal, regulatory policy and governance frameworks to maintain and safeguard peace and safety aimed at creating the environment conducive to socio-economic development. As societies become industrialized and as technology becomes part of common human life, many new ethical problems tend to arise that impinge on human safety, which need to be addressed by system designers and regulatory authorities. In the modern world, humans are constrained to coexist with entities created by law such as organizations, companies and regulators on the one hand and on the other hand, technological entities such as robots, driver-less autonomous vehicles, drones and even ubiquitous entities such as wearables, smart phone, tablets, home automation components interlinked to the Internet of Things (IoT). The glaring fact is that the Laws, Regulations and Ethics cannot keep pace with the rapidly emerging new technologies.

In the six decades since AI was born, it has become integrated in some form or the other in many diverse applications, and usage, e.g. in ubiquitous computing, smart phones, mobile apps, mobile internet, Big data, social networks, autonomous vehicles (AV), drones, Internet of Things, Clouds, Cyber Physical Systems (Robotics), advanced Health Care, Surveillance programs, Flexible Manufacturing, Smart cities, and Intelligent Decision Support Systems, to name some.

The ethics of artificial intelligence is the part of the ethics of technology specific to robots and other artificially intelligent agents, and beings. Each domain in the examples listed above raises its own ethical concerns. While the areas listed above have in the background AI programs, Intelligent Agents, Knowledge Bases, smart materials and subsystems, the regulatory policy and ethics have to be considered domain-wise. For example, what happens when a self-driving car has a software failure and hits a pedestrian, or a drone's camera happens to capture images of persons in a private swimming pool or an autonomous robot injures or kills a human? Smartness of systems is achieved by ubiquitous or pervasive computing. Locational Privacy is easily the most often-cited criticism of ubiquitous computing. Can and should there be whistle blower like behavior of bots used for customer support, trading etc. that detect and flag fraud or other unethical behavior it detects? Should a smart robot built for purpose, say security, be extended to or learn to help in the case of a medical emergency and who is accountable if a tragedy ensued?

The extraordinary build-out of the communication networks that link computers together is almost as remarkable as the explosion in computing power. Mobile communication, 4G, 5G, coupled with optical fiber communication infrastructure making available huge bandwidths have resulted in faster emergence and rapid diffusion of the Digital Economy more widely throughout the world than previous technological revolutions [4] giving rise to a new era in human development. Pervasive computing, fuzzy logic [5], cognitive informatics [6] - a new frontier located in an interdisciplinary area, which encompasses informatics, computer science, software engineering, mathematics, knowledge theory, cognition science, neurobiology, psychology, and physiology, tend to make artificial intelligence approach human level natural intelligence. Where would this lead us to? At this point in time no one can exactly predict the ultimate magnitude of the changes set in motion.

The new age where AI and super smart robots are a reality is upon us – as demonstrated by digital assistants incuding Google now, Siri (from Apple), Alexa (from Amazon), Cortana (from Microsoft), Google translate, Autonomous vehicles from Google, Tesla, autonomous Uber and

Singapore Taxis, and Airplane taxis as reported recently in media, to name a few. The positive impact of these innovations such as intelligent vehicles can be enormous. Consider the potential ecological savings of using highways so much more efficiently and from the safety aspect reduce the injuries and deaths in road accidents. Elderly and disabled people would be able to travel around in such cars on their own. People could dispatch their cars to the automated parking warehouse autonomously instead of using surface land for parking and then recall them later. Truly, gains in this area are enormous and most encouraging. The stuff of science fiction is rapidly becoming a fact of life. But this presupposes robust, reliable and secure networks, devices, and systems, in which individuals and society can repose complete trust.

By most indications and practical product developments and integrated systems, AI is now mature, both as a science and as an engineering discipline. Many opportunities exist for AI to have a positive impact on our planet's environment. AI researchers and development engineers have a unique perspective, tools and the skills required to contribute practically to address concerns of global warming, poverty, food production, arms control, healthcare including care of the elderly, education, and demographic issues.

How would real-time monitoring of emotions impact IoT? At the moment, emotional response is being tested by advertising agencies, but the applications don't stop there [7], according to the information in the referred ITU News. Reasoning ability in robots, feeling of pleasure and pain has been reported to be possible in the vast literature published on the subject. Unveiled in June 2014, Pepper, a robot that is capable of expressing and responding to emotions is a humanoid robot jointly developed by Aldebaran Robotics of France and SoftBank of Japan [8]. It has the ability to read emotions and. the ability to analyze expressions and voice tones. It is much more than a robot, designed to be a genuine humanoid companion created to communicate and respond in the most natural and intuitive way, through its body movements and voice. Pepper keeps on learning about the tastes, habits so as to personalize to the human companion, gradually memorizing the personality traits, preferences, and adapts itself. Pepper has been designed to identify human emotions and to select the behaviour best suited to the situation. While the technology behind these advances are astounding and even cause delight, the potential negative ramifications are to be considered. For instance, can fraud be committed by creating a humanoid robot to impersonate someone? Samani in his book on cognitive robotics [9] highlighted the need for aligning humanoid robots mimicking human behavior with human interest.

3. TRUST IN THE EMERGING SCENARIO

The subject of Trust has caught the attention of several scientists, professional institutions, e.g., IEEE, IEEE Robotics & Automation Society, Universities and R&D Labs, to assess impact on society, impact on work culture in the upcoming era with human - robots interaction, whether it would be helpful or hateful, collaborative or competitive, or simply aligned with humans. The recent edition (April 2016) of "ITU News" deliberates about: Do ICTs deserve our trust? Technologies are becoming more autonomous in their interaction and decision-making, but how does one trust in future ICTs? ITU-T has published a new technical report for experts in the field. The impact on society is vast and wide-ranging. Tesla's autopilot [10] highlighted the need to address important issues regarding an autonomous automotive future, including regulatory challenges, human override capabilities; and the pursuit of safer driving.

China is leading the global M2M technologies – an essential component for IoT – thanks to a huge smart phone subscriber base and the country's ambitious and wide-ranging M2M plans, says a new E&Y report. A GSMA report [11] predicts that the country would reach 1 billion M2M connections by 2020. Obviously, therefore, it calls for urgency in addressing the challenges. Another IEEE report [12] on the subject also raises the question of trusting the robots.

According to this report, ever more autonomous machines will present challenges spanning technical, regulatory, legal and even philosophical realms. They will force us to confront deep moral quandaries, and might even tweak our sense of what it means to be human. It has been seen that technology precedes before the laws to deal with its consequences are in place. Self-driving cars are ready before the laws are. Robots are performing surgeries with dexterity and precision – should they be allowed to take over the operation theatre? How much trust can one put in care giver robots? Should robots be only used for hazardous jobs like mining, working in nuclear facilities, chemical plants, etc., or used as warriors or autonomous weapons to decide who lives or dies? Could robots be taught human values and morals and imparted the sense to discern what is physically and morally right and what is absolutely wrong. These, among others are critical issues that need to be urgently addressed.

Wang in his classical paper on Cognitive Informatics [13], an emerging discipline that studies the internal information processing mechanisms of the brain and their engineering applications, via an inter disciplinary approach states how, human beings acquire, process, interpret, and express information by using the brain, and how minds of different individuals are understood that is, in a way, relevant to the design of super intelligence machines. In his recent paper [14], Wang expands the methodology and soft computing to cover philosophical, mathematical, and theoretical foundations for cognitive robotics and eBrain.

This paper does not dwell on the technology aspects of the control structure, cognitive informatics, soft computing, and so on, which are being dealt with in depth by experts in the scientific community and academia, and confines the scope

to discussion on the challenges posed by the emerging environment and need for ensuring human safety and paramountcy by bringing out appropriate ITU recommendations and standards relating to intelligent machines/robots in the emerging environment of IoT where the trends indicate their stupendous growth, together with associated regulatory and legal imperatives to be attended to.

4. POLICY & ETHICAL ISSUES

Alongside the growing interest in exploiting the potential for AI, there are also concerns expressed about its unethical use. Entrepreneur/inventor Elon Musk of Tesla and SpaceX, who spends considerable time on the cutting edge technologies, is reported (Feb. 2015) to have warned that AI could be an existential threat to humanity [15].

Bostrom and Yudowsky [16] have studied and analyzed at length the ethics of AI, and emphasized that thinking machines besides their not harming humans and other morally relevant beings, are also not as versatile as humans. They can have only domain specific intelligence that has been imparted by the designer, suitable for the assigned task but unsuitable for others. However, the possibility of developing super intelligent machines cannot be ruled out. They go further on to discuss the moral status of such machines themselves. AIs with sufficiently advanced mental states may count as persons—though maybe persons very much unlike us and perhaps governed by different rules. The prospect of AIs with superhuman intelligence and superhuman abilities presents us with the extraordinary challenge of stating an algorithm that outputs super ethical behavior. These challenges may seem imaginary, but need to be addressed in future research in order to forestall the hypothesis of I.J.Good (1965) [17] concerning super-intelligence: that an AI sufficiently intelligent to understand its own design could redesign itself or create a successor system, more intelligent, which could then redesign itself yet again to become even more intelligent, and so on in a positive feedback cycle in a kind of "intelligence explosion." To quote Bostrom [18], recursive scenarios are not limited to AI, humans with intelligence augmented through a brain-computer interface might turn their minds to designing the next generation of brain-computer interfaces that increased the IQ manifold.

[19] Notwithstanding whether the next big idea will come using AI, AI is already making a huge impact by creating music [20], making art, writing poetry and even astounding champion Go players by the beauty of the moves [21].
Despite the concerns expressed, none of the darker visions have deterred researchers and entrepreneurs from future work. Reality has lagged those grim imaginings.

Now, five of the world's largest tech companies Alphabet, Amazon, Facebook, IBM and Microsoft are attempting to create a standard of ethics around the creation of artificial intelligence [22]. The importance of the industry effort is underscored in the report recently released by the Stanford University group. The Stanford project, named the "One Hundred Year Study on Artificial Intelligence", would produce a detailed report on the impact of A.I. on society every five years for the next century. The A.I. industry group is perceived by some as having been modelled on a similar human rights effort known as the Global Network Initiative, in which corporations and nongovernmental organizations are focused on freedom of expression and privacy rights. The Stanford report titled "Artificial Intelligence and Life in 2030," states [23] that the basic objective is to ensure that A.I. research is focused on benefiting people and that it will be extremely difficult to regulate A.I., since there is no clear definition of A.I. (it isn't universal but generally, domain/application-specific), and the risks and considerations are very different in different domains. It does not rule out regulation totally but underlines when it would be good to do so.

The Stanford report attempts to define the issues more relevant to people in a North American city in intelligent systems that mimic human capabilities and explores eight aspects of modern life, including health care, education, entertainment and employment, but specifically does not discuss the issue of warfare and military-AI as being outside their current scope and expertise, but they did not rule out focusing on weapons in the future. The report also does not consider the belief of some computer specialists about the possibility of a "singularity" that might lead to machines that are more intelligent and possibly threaten humans.

The Working Group on Robot Ethics of UNESCO's World Commission of the Ethics of Scientific Knowledge and

Technology (COMEST) [24], is deliberating on the major ethical issues surrounding the increasingly widespread development and application of machines that encompass both physical robots and software agents, which are designed to function independently from direct human oversight and can learn by themselves new process or behaviours. The rapidity in the growth of the autonomous robots, both for civil and military purposes, leaves a gap between the effective use of the technology and its ethical application such as human well-being, safety or social benefits. The discussions are centered on the emerging ethical issues from two perspectives. First, the ways in which engineers and researchers design, build and use machines/robots in accordance with human morality and ethics, and second to find answers to such questions as: Do we want machines/robots to make morally important decisions? Are we then abdicating our responsibility to machines? What is the ethical understanding given to machines/robots and whether they should be considered as moral agents with artificial intelligence or even living creatures? Furthermore, considering that in the future, machines/robots will be sharing the world with humans, to explore the extent to which developments in robotics, as an aspect of a broader paradigm of technological convergence, point towards potential new understandings of "human" in respect of neurological implants and enhancing technologies.

5. EVOLUTION OF DESIGN RULES

The Engineering and Physical Sciences Research Council (EPSRC) and the Arts and Humanities Research Council (AHRC) of the UK in 2011 jointly published a set of five ethical "principles for designers, builders and users of robots" in the real world, along with seven "high-level messages" based on a September 2010 research workshop [25]. These are quoted in the following and relate to the design aspects:

1. Robots should not be designed solely or primarily to kill or harm humans.
2. Humans, not robots, are responsible agents. Robots are tools designed to achieve human goals.
3. Robots should be designed in ways that assure their safety and security.
4. Robots are artifacts; they should not be designed to exploit vulnerable users by evoking an emotional response or dependency. It should always be possible to tell a robot from a human.
5. It should always be possible to find out who is legally responsible for a robot.

The messages intended to be conveyed were:

- We believe robots have the potential to provide immense positive impact to society. We want to encourage responsible robot research.
- Bad practice hurts us all.
- Addressing obvious public concerns will help us all make progress.
- It is important to demonstrate that we, as Roboticists, are committed to the best possible standards of practice.
- To understand the context and consequences of our research, we should work with experts from other disciplines, including: social sciences, law, philosophy and the arts.
- We should consider the ethics of transparency: are there limits to what should be openly available?
- When we see erroneous accounts in the press, we commit to take the time to contact the reporting journalists.

In June 2016, Satya Nadella, current CEO of Microsoft Corporation in an interview with the *Slate* magazine is reported to have laid out the six "principles and goals" [26], more or less ethical in nature that he envisages AI research must follow to keep society safe. The artificial intelligence must:

1. "be designed to assist humanity" meaning human autonomy needs to be respected.
2. "be transparent" meaning that humans should know and be able to understand how they work.
3. "maximize efficiencies without destroying the dignity of people".
4. "be designed for intelligent privacy" meaning that it earns trust through guarding their information.
5. "have algorithmic accountability so that humans can undo unintended harm".
6. "guard against bias" so that they must not discriminate people.

> "The beauty of machines and humans working in tandem gets lost in the discussion about whether A.I. is a good thing or a bad thing. Our perception of A.I. seems trapped somewhere between the haunting voice of HAL in *2001: A Space Odyssey* and friendlier voices in today's personal digital assistants—Cortana, Siri, and Alexa. We can daydream about how to use our spare time when machines drive us places, do our chores, and help us make better decisions. Or we can fear a robot-induced massive economic dislocation later this century. Depending on whom you listen to, the so-called "singularity," that moment when computer intelligence will surpass human intelligence, might occur by the year 2100—or it's simply the stuff of science fiction."
> - Satya Nadella

Hopefully, the computer intelligence created by humans in the first place, never surpasses that of its own creator. Also, integration of ethics in artificial intelligence design and control structure is necessary and sufficient controls and checks need to be provided in an intelligent machine/robot to safeguard predominant position of humans that Mother Nature has endowed them.

As regards the judicial developments, comprehensive terminological codification for the legal assessment of the

technological developments in the robotics industry has already begun mainly in Asian countries [27]. This progress represents a contemporary reinterpretation of the law (and ethics) in the field of robotics, an interpretation that assumes a rethinking of traditional legal stipulations. These include primarily legal liability issues in civil and criminal law.

Mark W. Tilden [28] proposed three guiding principles/rules for robots, which do not pertain to humans or humanity, but to robots themselves. They stipulate that: (i) a robot must protect its existence at all costs; (b) a robot must obtain and maintain access to its own power source: and (iii) a robot must continually search for better power sources. Without regard to Asimov's laws, one finds these untenable and counter for harmonious and collaborative coexistence in the upcoming heterogeneous world scenario of super intelligent machines with humans.

6. AGENDA FOR ACTION

In a research report [29], Soars and Fallenstein have discussed the subject of aligning superintelligence with human interest, which appears to be one of the plausible solutions. Human society has evolved over the ages influenced by philosophers, sociologists, politicians, economists, religious leaders, technology, and adherence to certain values and principles based on recognition of individual human rights, independence, free thought, empathy, compassion, mutual respect, love, peace, and so on. According to an eastern philosophy [30], a person is considered divine or flawless if he embodies traits like fearlessness, cultivation of knowledge, charity, self-control, austerity and simplicity; nonviolence, truthfulness, honesty, courage of conviction, freedom from anger, tranquility, aversion to faultfinding, compassion and freedom from covetousness; gentleness, modesty and steady determination; vigor, forgiveness, fortitude, cleanliness, freedom from envy; and demonic: if a person embodies arrogance, pride, anger, conceit, harshness and ignorance. Could a super-intelligent machine be governed by a control structure or inbuilt system that will make it distinguish right from wrong and forbid what humans consider demonic behavior? In IoT, such a robot would have its own individual identity and could directly communicate with man or machine/ another robot and decide its own behavior based on what it considers right. As indicated [31], could a super intelligent robot be imparted with a conscience, to distinguish right from wrong according to its given moral code to safeguard human interests? Could a check list of Do's and Don'ts be integrated into it that it must abide by? Could super intelligent machines/robots coexist with humans and serve humanity in a positive way? It appears possible from the extensive research done and currently underway [32].

Aligning super intelligent machines with human interests is perhaps one way to go forward. Recent studies [33] show that a number of unanswered foundational questions relating to the development of general intelligence have been identified, and it seems possible to make some promising inroads. In scientific pursuit, a problem can have multiple solutions.

However, considering that the ITU is already actively working in this field, and recognizing the excellent work on IoT, cognitive informatics, artificial intelligence being carried out in relevant Study Groups, it is suggested that a multidisciplinary experts umbrella group, comprising scientists, academia, R&D Labs., sociologists, international law professionals, civil society, industry, and Government representatives, may be constituted under ITU auspices to deliberate on this very important subject to bring out Guidelines to be followed for designing super intelligent machines/robots aligned to human interests, by gaining consensus from the various stakeholders. Rather than having the various groups pursuing separate goals, laudable as they are on their own, having such an umbrella group will bridge the various technology innovations and policy and ethical needs, to ensure a harmonious outcome of their efforts and put human safety at the forefront. An item of particular importance is to come up with the consistent taxonomy of the AI technology and classification of AI systems by sophistication, in order to have a more nuanced regulatory and policy regime, instead of going with a single universal policy. There are existing attempts to classify AI by the level of sophistication using terms such as Artificial General Intelligence or Strong AI, Narrow/Weak AI or alternately by level of difficulty of the problem being solved using computer science terminology like AI-complete, AI-hard [34]. It might be beneficial to additionally categorize AI by considerations whether there is a physical component to the intelligence (in the form of mechanics and electronics, also referred to as mechatronics) or whether it is purely software – Visible AI versus hidden AI.

As stated before, these intelligent/ super intelligent machines may be designed not only to work solo or with other machines, but also with humans in a team. Their AI would need to be aligned to fall in line with the ethics of humans for harmonious team work.

With super-intelligent systems still being considered futuristic, it may seem premature to undertake the alignment of AI now. However, it would be prudent to develop a clear understanding of AI alignment well in advance of making design decisions about smarter-than-human systems. As the Research report of Machine Intelligence Research Institute says:" by beginning our work early, we inevitably face the risk that it may turn out to be irrelevant; yet failing to make preparations at all poses substantially larger risks."

In summary, AI is a very exciting technology area with astounding advancements in how machines (with or without mechanical components) can mimic human cognition, learning and problem solving. AI - as with the term cloud - is not a single entity but a broad collection of different technologies and techniques that target different problems and use cases in a variety of market segments. As with any

technology, the art of the possible far outstrips what is acceptable by society and the laws of each land. Additionally, as with any technology, while there are significant benefits, there are also downsides both known and unanticipated/unintended. In a sense, AI and super smart robotics could do to knowledge workers what industrialization and factory automation did to manufacturing workers. On the one hand, production output and efficiency could be exponentially improved while on the other hand peoples' jobs and livelihoods are at stake. Leaving the decision of where AI evolves thus cannot be left up to the industry and developers alone to tackle. The societal and governance inputs ought to be in place from the get go.

7. Acknowledgement

The author would like to acknowledge quotes from the personal lecture notes of Prof. VVS Sarma of Indian Institute of Science, Bangalore, India; and valuable assistance provided by Dr. CR Kasarabada, fellow Research scientist and Mr. Vasu S. Sankhavaram, an eminent IT professional.

REFERENCES

[1] https://en.wikipedia.org/wiki/Three_Laws_of_Robotics
[2]https://www.itu.int/net/wsis/tunis/newsroom/stats/The-Internet-of-Things-2005.pdf
[3] Google self-driving cars
[4] Carl Shapiro and Hal Varian (1998), *Information Rules: A Strategic Guide to the Network Economy* (Boston: Harvard Business School Press), p. 9.
[5] Lotfi Zadeh; WCECS 2014 - KN "A way for progression toward the achievement of long-sought human level machine intelligence."
[6] Wang, Y; IEEE ICCI'02 – KN – Cognitive Informatics
[7] www.rcwireless.com/20160630/internet-of-things/emotional-internet of things-tag28
[8] www.ald.softbankrobotics.com/en/cool-robots/pepper
[9] Hoonan Samani: Cognitive Robotics (Book), CRC Press
[10] Guardian newspaper of 30 June 2016
[11] GSMA M2M Report 2020
[12] http://spectrum.ieee.org/robotics/artificial-intelligence/can-we-trust-robots
[13] Proceedings of the First IEEE International Conference on Cognitive Informatics (ICCI'02) 0-7695-1724-2/02 $17.00 © 2002 IEEE
[14] Yingxu Wang, WConSC'16, UC Berkeley, USA, May 22-25, 2016 (The 6th World Conference on Soft Computing, UC Berkeley), Soft Computing: Philosophical, Mathematical, and Theoretical Foundations for Cognitive Robotics and eBrain
[15]http://www.japantimes.co.jp/opinion/2015/02/04/editorials/ethics-artificial-intelligence; Japan Times dated February 4, 2015, Editorial.
[16] Bostrom, Nick, and Eliezer Yudkowsky. "The Ethics of Artificial Intelligence." *Cambridge Handbook of Artificial Intelligence*, edited by Keith Frankish and William Ramsey. New York: Cambridge University Press.
[17] https://en.wikipedia.org/wiki/Technological_singularity
[18] http://www.nickbostrom.com/ethics/ai.html
[19] https://cns.utexas.edu/news/the-future-of-ai
[20]http://www.eetimes.com/document.asp?doc_id=1145348
[21] https://www.wired.com/2016/03/sadness-beauty-watching-googles-ai-play-go/
[22] ITU News, September 2016
[23] Artificial Intelligence and Life in 2030, One hundred year study of Artificial Intelligence, Report of the 2015 Study Panel Sept 2016
[24] http://www.unesco.org/new/en/social-and-human-sciences/themes/comest/
[25] Principles for designers, builders and users of robots
[26] Slate interview with Satya Nadella
[27] Wikipedia, "Laws of Robotics", http://en.wikipedia.org/wiki/Laws_of_robotics
[28] https://en.wikipedia.org/wiki/Mark_Tilden
[29] Research supported by the Machine Intelligence Research Institute (intelligence.org). Published as technical report, 2014{8}.
[30] Bhagavad Gita, Chapter 13, Bhaktivedanta, Published by ISKCON 1972
[31] Quote from colleague Dr W Selvamurthy, Former Scientific Advisor, Govt of India
[32] http://blogs.wsj.com/japanrealtime/2014/06/06/why-softbank-turned-to-france-for-an-emotional-robot/
[33] https://www.cs.bath.ac.uk/~jjb/ftp/BrysonAISBQ12.pdf
[34]https://en.wikipedia.org/wiki/Artificial_general_intelligence

SESSION 2

SERVICE AND QUALITY STANDARDS

S2.1 Invited paper: Ageing, well-being and technology: from quality of life improvement to digital rights management. A French perspective.

S2.2 Universal Service, quality caps and net neutrality.*

S2.3 Quality and standardization in technology-enhanced learning.

AGEING, WELL-BEING AND TECHNOLOGY: FROM QUALITY OF LIFE IMPROVEMENT TO DIGITAL RIGHTS MANAGEMENT A FRENCH PERSPECTIVE

Nathalie Devillier LLD
Research Professor
Grenoble Ecole de Management

ABSTRACT

ICTs are a critical enabler for a sustainable development that helps with unlocking human capabilities. Technology related to ageing belongs to several sustainable development targets and action lines: social inclusion, access to scientific knowledge, capacity building, cyber-security and ehealth. While the objective is the improvement of quality of life for citizens, a special monitoring is needed so as to regulate ICT use in compliance with the rights and freedoms of data subjects. How can we ensure that these tools are implemented in a way that is relevant and appropriate? This article aims at finding the right balance between multiple stakeholders' interests: in particular the relationship between risk management and respect for private life and lighting providers about integrating data protection into the design of their products so as to improve the quality of life for citizens and preserve their privacy.

***Keywords*—** Data protection, ethics, health, privacy, robot.

1. INTRODUCTION

While the population in Europe and Northeast Asia is aging and increases, the World Health Organisation (WHO) Director-General, Dr Margaret Chan, said: "we need to ensure these extra years are healthy, meaningful and dignified. Achieving this will not just be good for older people, it will be good for society as a whole" [13, 14]. ICTs are a powerful lever for sustainable development and a way to attain both the fundamental human right to healthcare (Art.25 of the Universal Declaration of Human Rights or UDHR, Art.12 of the International Covenant on Economic, Social and Cultural Rights or ICESCR) and the right to enjoy the benefits of such scientific progress (Art.27 UDHR, Art.15 ICESCR). The relationship with the sustainable development goals (SDG) n°3 to "Ensure healthy lives and promote well-being for all at all ages" is the most relevant to analyze how ICT applications and services achieve sustainable development.The elderly is connected and easily adapts to its environment and frequently uses smartphones and tablets. Still, some active retirees may start to feel a weakness that can be physical (fatigue) or social (isolation), and others lose their autonomy. Gerontechnology is an academic and professional field, interdisciplinary, which synthesizes gerontology and technology. It aims at creating technological environments for health, housing, mobility, communication, and recreation for the elderly. Research findings in this area provide a basis for ergonomists, designers, engineers, manufacturers and health professionals (geriatricians, nurses, psychiatrists, psychologists, gerontologists, etc.) to establish remote monitoring tools: bracelet, watch, actimetrics, sensors, accelerometers, or vital parameters of health, video, robots... Thus, ICTs achieve the sustainable development goal 3 « Good health and well-being » and are an efficient tool to implement public policies for social inclusion. A example at the EU level is the European Commission « Active and Assisted Living Joint Programme » and the « European Innovation Partnership on Active and Healthy Ageing » which brings together government, care professionals, industry and users across borders to scale up and bridge the gap between seller and buyer, producer and user. In this field ICTs play a key function to strike the right balance between various interests: governments, the private sector and citizens. As a result, several EU Member States have launched specific strategies on the Silver economy, emphasizing the potential of SMEs to benefit from the increase in global demand for goods and services for seniors. France is one of them with the industrial contract for the Silver Economy initiated by the Ministry of economic renewal. It finds its rationale in both the ability to identify innovative ways to meet the care needs of this segment of the French population, but also to emphasize that this is an emerging growth area of the economy where the creation of innovative services and products is conducive to the creation of new jobs. With this framework, public authorities and private stakeholders commit to the implementation of concrete actions aimed at making the Silver economy a genuine industrial sector to position France as a world leader in this sector which includes: home automation, housing (homes, almshouses, inter-generational buildings), smart cities, mobility (transportation) and health, including e-Therapy [6].

However, assessing such ICT applications and services goes beyond technical issues, their legal and ethical context are as important as international standards and specifications [11]. ICTs involve articulating fundamental rights with others like privacy, freedom of movement, and ethics [8]. International standards, French and European legal requirements and ethics are strong levers to promote a sustainable development of ICT applications and services. This article

will explain how technology related to ageing and ambient assistive living contribute to sustainable development goals, targets and actions (2), and then examine how freedom of consent to the use of ICT applications and services, privacy, and data protection are implemented by the industry with the incorporation of digital rights management considerations (3).

2. IMPROVING THE QUALITY OF LIFE FOR CITIZENS.

2.1 Sharing knowledge and improving access to information for effective ICT use.

ICT applications and services for sustainable development offer a better access to scientific knowledge and therefore contribute to the implementation of the World Summit on the Information Society (WSIS) Action Line C3 "Access to information". Information is gathered and shared with innovative ICTs and contribute to the emergence of a "Science 2.0" (wikis, blogs, video journals) and now Watson (deep question answering natural-language computer system), from the IBM laboratory. Watson's natural language, hypothesis generation, and evidence-based learning capabilities allow it to function as a clinical decision support system for use by medical professionals! Robots are today able to detect health problems as the Roberta Ironside project showed, robotic evolution embodies the expertise learned during the development of pure vocal personal assistants for dependent persons [12]. Other robots are equipped with electronic stethoscopes, otoscopes (inspection of the ear canal) and ultrasound equipment that allows a physician to diagnose remotely and help maintaining the independence of older people. Restorative oriented robots prostheses allow return of missing functions: traction for artificial legs, arms and grasping for artificial hands. Robots are considered by national and international agencies of health products as "medical devices" and are evaluated as such.

New terms are used or have an ICT connotation such as Massive Open Online Course (MOOCs), Access to Scientific Knowledge (A2SK), big data, digital identity… New products and services emerge, new tools and… new research questions! In the legal field for instance, lawyers don't have the same views on data legal status: is it possible to have an "ownership" of data when it is available on the Internet? Is it possible to sell your data to a private entity? What about liability when a robot is involved? For instance, when Watson helps a general physician to make a diagnosis, would it be also liable in case the patient suffers damage just like a physician is?
Profound transformations for the scientific community as well as citizens are open access to science and open data. Open data in the field of health is implemented in France with the recent Law on the Modernization of our Health System (Law n°2016-41 of 26 January 2016). Participants to the WSIS 2016 Forum discussed those issues and considered it was urgent to identify tools available to answer needs [15, p.91] which shows it is now time to have a mature and effective use of ICTs.

2.2 Capacity building, e-health and telemedicine.

Capacity building for e-health applications is not just for technologists but also for policy makers and health workers. WSIS Action line C4 is relevant to understand the impact of ICTs on a better access to health care and how different interests are interacting with each other's: public health policy, emerging processes, products and services from the private sector and patients. Regarding eHealth applications, training and capacity building are critical in the course of health delivery. This is true in particular in rural areas where a general practitioner will for instance make an ECG and see the result on his smartphone and then send it for second opinion to a cardiologist. Capacity building is addressed by the World Health Organization with several partnerships and initiatives such as "Be Healthy, Be mobile" which focus on large-scale implementations of mHealth for Non-Communicable Diseases. This is leveraged at national level for tobacco cessation (mCessation, an application used by India and Tunisia) and the fight against Ebola (SMS campaign in a massive public awareness effort by Senegal). Multi stakeholders' partnerships are essential to strengthen capacities of health workers just like the United Nations Development Programme (UNDP) Telemedicine initiative showed. The proliferation of e-health apps is creating a need for specific skills as the OEDC explains [10]. This also means that teachers and professors should have adequate skills to educate and train people for their initial and long-life learning. Capacity building is required at all levels: health professionals, processors, patients, end-users of well-being apps are of particular importance. Indeed, it is difficult for users to select an app which relies on relevant scientific evidence (cf Lumosity was fined $2 million for deceptive brain health ads by the Federal Trade Commission in the US). In the specific case of gerontechnology for instance, the tool is granted by the elderly on a proposal by the parent or relative, the person of trust, or trustee, the guardian, a GP (medical monitoring or remote assistance, therefore telemedicine), or the Director the health facility. Gerontechnology remains subsidiary, respect of the person wishes is essential; the choice should be driven by a logic of individual risk prevention, not on the logic of general precaution. Here again, capacity building is essential to help users in selecting the relevant application or service.

In the field of telemedicine, an interesting global initiative is "WeTelemed" (Women for Telemedicine) which promotes capacity building in telemedicine and gender equality meanwhile in line with SDG 5. WeTelemed actions include programmes in telemedicine and e-health targeting young girls with the aim of increasing the number of women in ICT careers as well as promotion the exchange of best practices on the integration of gender perspectives in ICT education. The French language Virtual Medicine University is placed

under the auspices of both the National Conference of Deans of Medical Schools and the International Conference of Deans of French-Speaking Medical Schools. It has been developed to create a major Internet portal for French-speaking distance medical learning and teaching. The national initiative "Patients' University" (Grenoble, Marseille, Paris) considers patients as knowledge producers but also as subjects performing a number of activities in the service of their "self keeping alive". The university offers patients both to take account of the acquired experience of the disease and meet the needs of patients in terms of degree programs in the field of education and training. The project also involves using the university as an intangible heritage protection area comprising the knowledge of experience and action acquired by several generations of patients.

E-Health applications are also a WSIS Action Line (C7) and the interaction with e-agriculture is quite strong regarding nutrition, which means that it is possible to address multiple interlinked goals. For instance, platforms mix different health information such as malnutrition and malaria: the data are not delivered with statistics but are understandable for its users (SNISI in Mali). This shows how the holistic nature of SDGs can be addressed thanks to ICT in a quite innovative way. The private sector is also considering its contribution to the society as a whole as Tomas Lamanauskas (Vimpelcom) declared during the WSIS Forum 2016: "We are a company. What we do is not charity but we do it out of our business interest. But for our business to play a key role in socio-economic issues of the communities we need to provide high quality services to our customers and leverage our technology to support development" [15, p.112]. It is now clear that ICT applications and services for sustainable development will be addressed beyond 2016 in a cross-sectorial perspective with all stakeholders involved: "Transcending from infrastructure to applications" has been introduced to Action line C4 to focus on e-applications and imperatives that will enable people to adopt, use and benefit from different applications.

Capacity building activities also include conformity, interoperability and testing so as to reduce the digital divide and standardization gap in developing countries especially introducing new technologies. The objective is to clarify fundamental aspects as accreditation, certification and mutual recognition agreements. It is essential to ensure a consistent and interoperable ecosystem of personal connected health devices via international standards. All required information for medical decisions must be accurate and available to healthcare professionals, standards also cover machine-to-machine communications for medical devices such as the Institute of Electrical and Electronics Engineers (IEEE) standards. Other standards regard higher business-level languages for enterprise-level electronic health data exchange: the Health Level Seven (HL7) and Integrating the Healthcare Enterprise (IHE) standards, another regards patient consent preferences (HL7 Clinical Document Architecture Release 2 Consent Directive – HL7 CDA R2). International standards cover the personal area network device (such as a blood pressure meter, pedometer, or personal alarm), the application hosting device (cell phone, PC, or specialist device), the wide area network device (a telehealth remote monitoring service platform) and the health record (a physician's electronic health record). However, it remains difficult to align ICT applications and services with customers' rights. Service contracts' are not negotiable by the customer and rely on the acceptance of the Terms and conditions of processing, storage and use of personal data collected by the application or service. Such clauses are of paramount importance as they determine the extent of monitoring (video, geolocation): where and when monitoring takes place (hall, lounge, kitchen, bedroom, bathroom, 24/24). Does the customer really master the parameters? Is it possible to disable or limit its functionalities? Does the customer have an effective control? Is the Hotline informed when the client disables a function? And caregivers? Are the original features restored automatically? How long is the data stored? What are the purposes of the processing? What is the cookie policy? All these elements need to be clarified in the Terms of Use both for the customer, the platform services and caregivers. Therefore, capacity building should include this legal perspective so as to deliver sustainable ICT applications and services contracts.

2.3 Consent and special needs (disabled & elderly).

ICT applications and services for health and well-being of the elderly are directly relevant to WSIS Action Lines C2, C3, C4, C7, and C8, which refer to disability, to accessibility and to 'disadvantaged and vulnerable' groups (which includes persons with disabilities) which reflects the SDG commitment on leaving no-one behind. WSIS Forum 2016 outlined the need for trade, intellectual property and competition-related regimes that do not restrict the right of persons with disabilities to benefit from technologies that would enhance accessibility [15, p.255]. The French Project of Law for a Digital Society provides better protection for users, increased openness of public data and improved access to the Internet. With this new law, France is preparing to improve access to its public sites for disabled (Art.44). The French Law on Society adaptation to ageing is also a lever in the development of gerontechnology for people with low incomes (Law n°2015-1776 of Dec. 29, 2015). Here again, it is the government duty to preserve the elderly. The first instrument a stronger consent requirement: the health professional must obtain the informed consent of the person he treats or his legal representative when the person is "vulnerable", "fragile" or "out of state to express his will" [4]. Favier suggests to "better define the role of third parties that might accompany the vulnerable person in a decision-making" in order to avoid a formal consent resulting in the denial of the person. Reflection on a theory of vulnerabilities would renovate the narrow approach of consent prevailing in law. Another protection is surprisingly provided by the data protection authority (French National Commission for information technology and civil liberties, CNIL) that have highlighted the need for a relevant control

of ICTs for the elderly: it must be based on a medical opinion issued after a peer review conducted by a multidisciplinary team that supports the person and give rise to regular reassessments (director, manager, coordinating physician, nurse, caregiver, psychologist, instructor, educator, occupational therapist, psycho-motor therapist, hospital service agents, home care assistant, gerontology nursing assistant, physiotherapist). The processing of personal data is monitored by the CNIL and implies compliance with security and confidentiality requirements. CNIL has set up a repository to enable a self-assessment by the candidate to host personal data and an efficient processing of applications for approval, together with the Agency for Shared Health Information Systems in consultation with the French industry. Here, the physician certifies that the person presents difficulties or disorders of intellectual functions justifying the use of the application or service. The consent of the person and the aforementioned third is collected. It is also possible to collect the consent following an early diagnosis of a disease involving forward the possibility of disorders of intellectual functions (Art.L311-4 of the French Social Action and Families Code). If the person is the subject of a measure of protection, then the legal guardian decides without having to justify any refusal; it is still subject to the concurring judgment of a physician chosen by the public prosecutor (Art.431 French Civil Code). If the person is accommodated in a health facility, social or medical, the decision rests with the director of the institution upon medical advice and following a peer review of the multidisciplinary team care, accompanied by a support person and family members. If the person is hosted at his home, the decision belongs to the parent, upon medical advice and following a peer review of the multidisciplinary team of care [7]. This is concretely how the elderly special needs are addressed by the law-making authorities.

Alzheimer's disease and patient “consent”. A person with Alzheimer's has not all his powers of judgment, therefore consent to the implementation of the service or application is questionable. People with a degenerative neurological disease tend to refuse a service like GPS tracking even if it aims at protecting themselves [5]. This refusal may also result from family conflicts or with relatives, or even a desire to endanger. In such a situation, ethical views help in imposing the ICT application or service to the patient [1]. The French Law of 5 March 2007 on the legal protection of adults is intended to apply to all people with alterations of their personal faculties (mental and / or physical) as soon as the expression of the will is prevented or lessened to the point that the individual can not only take informed personal decisions. The legal representative must ensure access to personal care to promote living conditions and should involve the person in any decision regarding his health. In case the consent cannot be obtained, information must be delivered to the person before the implementation of the application or service and a detailed medical examination by a physician is required as well as a list specially prepared by the prosecutors and by the guardianship judge. An early consent, as well as advance directives, is also an option to consider.

The concept of assent was proposed by the EREMA (National Ethics Thinking Space on Alzheimer's Disease, a French Association) because of these obstacles, and in the name of respect for the person, it seems to be necessary to go beyond the exclusive reference to the notion of consent and to recognize the importance ethics and the legal relevance of "consent" to aid and care: it is to give full meaning and importance required to the opinion of the person unable to issue a fully free and informed consent, but still able to participate in decision making, by strengthening the peer review intended to put the patient at the heart of the accompaniment. This concept is today in the Helsinki Declaration of the World Medical Association since 2013 (Art.28-30) which is neither a standard, nor a binding instrument.

3. DIGITAL RIGHTS AND IDENTITY MANAGEMENT.

3.1. Enabling a trusted connected world.

Data collected by a personal health device is shared with application hosting devices (personal computers, gateways, tablets, phones) using various transport technologies like USB, Bluetooth, Near Field Communication (NFC), Bluetooth low Energy and ZigBee. As a result, ICT applications and services impact the fundamental right to privacy, the fundamental freedom of movement, dignity, integrity, and human security. All those rights could be undermined in the absence of specifications regarding functionalities and tasks of the application or service i.e. ethical limits. They must be deactivated and reactivated easily and at any time by the concerned persons: no permanent monitoring by a third party, the person retains the initiative of the request for assistance. When cameras are used, they must not be placed in areas where respect for privacy is needed (toilet, shower...). The respect for private and family life is recognized by the European Charter of Fundamental Rights (ECFR) which declares that “Everyone has the right to respect for his or her private and family life, home and communications” (Art.7). The disclosure or improper discovery by third persons of facts relating to physical condition, health or personality may undoubtedly interfere with one’s privacy and private life (Eur. Comm. H.R., DVO v. Belgium (Appl. No. 7654/76), Rep. of 1st March 1979). The wide application of the principle of privacy necessitates protection of personal data used, for example, for social security purposes. The European Court of Justice (ECJ) has recognized that the right to respect for private life secured under Article 8 of the ECFR includes a person’s right to keep his state of health secret (ECJ, Case C-404/92 P, X. v. Commission of the European Communities, [1994] ECR I-4737, judgment of 5 October 1994). Important regulations concerning data security are in the Council Directive 2002/58/EC on privacy and electronic communications, i.e., the protection of individuals in connection with the treatment of personalized data its

Preamble - §§ 11 & 24 - refers explicitly to Article 8 of the ECFR and the case-law of the European Court. The Directive emphasizes the right to privacy in Article 4, which obliges the provider of a public available electronic communication service to take appropriate measures to protect the security of its services. If there is a particular risk of a breach of the security of the network, the provider must inform the subscribers of such a risk. As a consequence, personal data belong to this fundamental right to privacy, it is therefore not a subject of property. The French Data Protection Act highlights this humanist philosophy as well as European law.

Whereas the expansion of connectedness brings more people online, the question of trust for the use of ICT applications and services becomes crucial in terms of networks infrastructure, and human resources. This topic is of importance for every one of us since we are all internet users and potentially clients (consumers) of applications or e-services. From privacy to data portability, right to be forgotten and the protection of the rights and freedoms of data subject, consumer protection is impacting the regulatory framework of ICT apps and services. Terms of service and confidentiality policies impact customers' privacy. For instance, temperature sensors placed at home, biometric devices, remote monitoring systems by Internet... Assistant robots are also mobile and interactive, they are integrated in an intelligent living [12] to ensure the safety of single senior and ensure their comfort while stimulating their cognitive abilities (relief task of caregivers and maintaining the social link) while recording data on the past life of the person [16] which raises the question of an anticipatory technology governance to ensure it.

Here again international standards have security purposes: standardization has a significant role in reinstating confidence. For the services interface and the hospital information systems (HIS) interface, security is achieved through several standards: consent management (HL7 CDA R2 Consent Directive) and consent enforcement (IHE Document Encryption Profile); auditing (IHE Audit Trail and Node Authentication - ATNA), confidentiality, integrity and service authentication (WS-I BSP, TLS v1.2, TLS v1.1 and IHE XDM S/MIME), and entity authentication (WS-I BSP, WS-Security + SAML 2.0, or OAuth, IHE XUA, IHE XUA++); identity management (IHE Patient Identity Feed HL7 V3, IHE PIXV3 Query transaction, and IHE Patient Demographics Query HL7 V3 transaction). Those standards support personal health devices health and fitness, aging independently, and disease management. Other standards apply to health record networks interface, messaging infrastructure and transport.

When we look at national policy instruments and laws, it is unclear whether elderly people, just like children, are targeted in terms of online security (USA, France and ITU's Child Online Protection Initiative). Still this population should be more protected than others. While for children parents play a key role, for the elderly, relatives, health professionals may take the decision to use ICT applications or services. In such extreme situation, ICTs for ageing and assistive living become a response to a loss of independence of a person whose mental faculties are affected, they have a function of prevention and assistance. Here, gerontechnology is implemented by family care givers and the recipient (or imposed by the caregiver), or worse, it can be canvassed by an unscrupulous business (abuse of weakness). In this case, common law applies: consent is naturally required, still the law doesn't positively define its conditions; a consent is valid as long as it is not vitiated by error, fraud, violence, or injury. For instance in France, it is based on a real, healthy and conscious will corresponding to the absence of mental disorder (mind insanity, Art.1109-2 Civil Code) regardless of the origin of the mental impairment, such as old age (senile dementia).
Trust in a connected world is redefining state agencies' powers to protect consumers online: data protection authorities, consumers' protection bodies as well as competition agencies have a growing impact on preserving fundamental rights of users. Investigation powers, administrative fines for instance are strong levers for getting companies' compliance with regulations. Meanwhile, the private sector is integrating those requirements with the adoption of best practices, Codes of conduct or other self-regulation instruments. Those elements are specific to the digital environment whose governance is questioned at the highest level by institutions and NGOs.

3.2 Cyber-security.

Digital agendas and cybersecurity are not managed by the same agencies at national level (finance, trade, health etc). However, privacy, encryption, prevention of backdoors must be discussed, promoted and aligned with digital strategies to better achieve WSIS Action Line C5 "Building confidence and Security in the use of ICTs". The Global Cybersecurity Index measures the commitment of countries to cybersecurity regarding five dimensions: legal, technical, organizational, capacity building and cooperation. The empowerment of developing countries can be achieved with case studies and highlight the different options available for instance regarding data breach notification (voluntary vs. mandatory reporting) and ITU's national CIRT Programme (Computer Incident Response Teams). The question of whether or not a binding legal instrument should be adopted to ensure cyber-security is also raised. For instance, the European Union has adopted a comprehensive data protection regulation which includes data breach notification, administrative fines and a right to sue controllers and processors. On the basis of the ECFR which states that "Everyone has the right to the protection of personal data concerning him or her. Such data must be processed fairly for specified purposes and on the basis of the consent of the person concerned or some other legitimate basis laid down by law" (Art.8). Article 16 of the Treaty on the Functioning of the EU also states that "Everyone has the right to the protection of personal data concerning them". The European Regulation passed in April

imposes that the data should be “collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes” (Art.5-1-b Regulation 2016/679 on the Protection of natural persons with regard to the processing of personal data and on the free movement of such data, OJEU, 04/05/2016, L119/1). As a consequence, ICT applications and services cannot be used to gather information on the person, or profiling her for future sales prospection (Art.16 Regulation). The controller must ensure that adequate measures are taken and is accountable to ensure the security, privacy and confidentiality of the data processed to prevent destruction, loss, or theft of personal data (information system security policy, Code of Conduct...). The Regulation says that data must be “processed in a manner that ensures appropriate security of the personal data, including protection against unauthorized or unlawful processing and against accidental loss, destruction or damage, using appropriate technical or organizational measures” (‘integrity and confidentiality’, Art.5-1-f). Privacy by design promotes respect for privacy from the design software and services to "make *ab initio* the guarantee of privacy a monitoring unit placed within the technology design stage" [9]. The ICT applications and services should thus easily be disabled by the user and offer all the facilities previously stated. The European Regulation on the protection of natural persons with regard to the processing of personal data offers a legal basis for data protection by design and by default (Art.25 Regulation 2016-679 of 27 April 2016). Data protection by default requires that only personal data which are necessary for each specific purpose of the processing are processed. The European Regulation imposes heavy liability on the controller who shall be responsible for, and be able to demonstrate compliance with the Regulation basic principles (‘accountability’, Art.5-2). When the technology is likely to result in a high risk to the rights and freedoms of natural persons, the controller shall, prior to the processing, carry out an impact assessment including the risks to the rights and freedoms of data subjects (Art.35-7-c Regulation). Eventually, data protection authorities have the power to impose administrative fine up to €20,000,000 or 4% of global turnover for breach of data protection principles such as: absence of free and unconstrained consent; security breach, irrelevant processing, and failure to comply with objection to profiling (Art.83-5 Regulation). Data protection by design and default, security and privacy impact assessment breaches are also subject to fines up to €10,000,000 or 2% of global turnover (Art.83-4 Regulation). Civil and criminal liability of the health data host can be sought in case of disclosure to an unauthorized person.

4. CONCLUSION.

Technology, standards, legal norms and ethics are important drivers to reach SDGs regarding ageing and well-being in order to improve quality of life and manage digital rights. Benefits of ICT applications and services must outweigh its constraints, then it will result in a better individual and social life for every people. Everyone should organize the various implications of his advanced age and anticipate his choices by formalizing its decision to use gerontechnology individually. Conversely, a systematic removal of the alarm button for instance should be considered as a refusal, although it is not formalized: a specific approach will then be considered collegially. Reflection on a theory of vulnerabilities would renovate the narrow approach of consent prevailing in French law. Ethical reflection on gerontechnology seems to lag behind the technical development.

Older people have ambivalent experiences with technology, as it gives rise to possibilities as well as constraints, and safety as well as worries [3]. Therefore, evaluation processes are a key topic to make sure legal and ethical requirements are met. Evaluation of gerontechnology is now needed to help users, patients, health professional to select relevant tools that comply with ethics. The French data protection authority, CNIL, published an opinion regarding gerontechnology for elderly people [2] saying it shouldn’t unduly affect people’s rights and freedoms, in particular human dignity. Its use should rely on a case by case analysis by the health professionals and subject to re-evaluation. The Welfare and elderly people rights national Committee adopted a Good practices Charter on Geolocation in 2013 which focus on dignity, privacy, security and risk prevention. The information should be adapted to the person discernment ability and also delivered to the family; the gerontechnology should be approved by a GP, and data processing remains under the monitoring of the CNIL. Eventually, safety and security are integral part of human rights just like ethical dimensions of ICTs for common good and values. They need to be formalized so as to ensure all citizens access relevant and trustful applications and services worldwide. The objective is to prevent abuses of such tools and find an equilibrium with fundamental freedoms. Some initiatives aim at monitoring the different technology options via certification and labelling. Associations check medical, technical, ethical and legal requirements and deliver the label. Still to date, there is no official recognition of those evaluation systems.

REFERENCES

[1] Alzheimer Europe, *The Ethics of Dementia Research*, Alzheimer Europe Research, 2011, 184 p.

[3] CNIL, « Systèmes de suivi et d'assistance électroniques des personnes âgées ou désorientées : les recommandations de la CNIL », online Press Release, 23/07/2013.

[3] Dahler A.M., Rasmussen D.M. and Andersen P.T., « Meanings and experiences of assistive technologies in everyday lives for older citizens: a meta-interpretive review », *Disability & Rehabilitation: Assistive Technology*, 2016, vol.7, pp.1-11.

[4] Favier Y., « Vulnérabilité et fragilité: réflexions autour du consentement des personnes âgées », *Revue de Droit Sanitaire et Social*, 2015, pp.702-713.

[5] Jefferson, A.L., Lambe, S., Moser, D.J., Byerly, L.K., Ozonoff, A., Karlawish, J.H., "Decisional Capacity for Research Participation in Individuals with Mild Cognitive Impairment", *J. Am. Geriatr. Soc.*, 2008, vol.56, pp.1236-1243.

[6] Kotsopoulou A., Melis A., Koutsompu V.-I., and Karasarlidou C., "E-Therapy: the ethics behind the process", *Procedia – Social and Behavioral Sciences, 2015,* vol.65, pp.492-499.

[7] Landau R., Auslander G. K., Werner S., Shoval N. and Heinik J., "Who should make the decision on the use of GPS for people with dementia", *Ageing & Mental Health*, 2011, vol.15 no.1, pp.78-84.

[8] Lobato Faria P. and Cordeiro J. V., "Managing difficult ethical and legal equilibrium in healthcare: Assuring access to the basics while keeping up with innovation", *Revista portuguesa de saude publica*, 2014, vol.32 no.2, pp.121-122.

[9] Loiseau, G. « De la protection intégrée de la vie privée (Privacy by design) à l'intégration d'une culture de vie privée », *Légipresse*, 2012. p. 71.

[10] Organisation for Economic Cooperation and Development, Skills for a Digital World, DSTI/ICCP/IIS(2015)10/FINAL, 25 May 2016, 58 p.

[11] Rialle V., « Quelques enjeux de l'éthique évaluative en gérontechnologie », *Revue Francophone de Gériatrie & Gérontologie*, 2009, vol. XVI, pp. 310-314.

[12] Sansen H, Chollet G., Glackin C., Jokinen K, Badii A., Torres M-I., Petrovska-Delaz D., Boudy, J, Schlögl S., (2016). The Roberta IRONSIDE project, A cognitive and physical robot coach for dependent persons, Handicap 2016, Paris, June 2016.

[13] WHO, World report on ageing and health, 2015, WHO, Geneva, 260 p.

[14] WHO, Dr M. Chan, "Number of people over 60 years set to double by 2050; major societal changes required", 30/09/2015, WHO Press Release, online: http://www.who.int/mediacentre/news/releases/2015/older-persons-day/en/.

[15] WSIS Forum, Outcome document, WSIS Action Lines: Supporting the implementation of the SDGs, International Telecommunication Union, 2016.

UNIVERSAL SERVICE, QUALITY CAPS AND NET NEUTRALITY

Emilio Carrera Félix

UNIVERSITÉ PARIS II - PANTHÉON ASSAS
Paris Center for Law and Economics, Paris, France.
Emilio.Carrera-Felix@u-paris2.fr

ABSTRACT

In the actual network configuration, content providers are not involved in the universal service programs with a corresponding participation being taxes, funds or compensations. We propose a regulated Internet contract with a free or very small economic participation to access into a limited version of the service with content providers financing participation to broadcast their high bandwidth content. We study a minimum quality of a service as a strategy of public policy over a broadband telecommunication services to create better absorption of technological benefits as a welfare measure for the users. As results we show a positive effect to propose universal service by quality as complement of universal service obligation, the conditions to determine the prices have to be ruled ex-ante by authorities and finally the best market scenario for welfare superior is determined by competition.

Keywords— Quality standards, Universal Service, Regulation, Net neutrality, Zero price rating.

1. INTRODUCTION

If we ask a simple question: is it possible that a broadband provider proposes a free Internet service? Suppose that it is possible, so better ask otherwise; in which context could a broadband operator propose a free Internet service as part of a Universal Service policy on broadband market?

To revisit this idea in the actual configuration of broadband services, we need to understand the importance of quality in a broadband context. Nowadays, the perception of Internet speed is considered as the single most important metric of interest in characterizing the "quality" of broadband service [1]. More speed perception in a "*ceteris paribus*" situation, especially over the price, will announce a better customer satisfaction. However, the different perceptions of broadband speed (capacity, advertised and achieved) difficult the analysis of data speed.

Nevertheless, speed or bandwidth does not depend only on the technical capacity of the broadband service at the present time. The high data demand from content providers (CP) has become one of the reasons why the average speed of broadband does not depend only on the Internet service provider capacity (ISP). ISPs have been turning to pricing as the ultimate congestion[1] management tool to mitigate the fear of a data saturation [3].

Notwithstanding this situation, the financial transfers between residential ISPs and CPs would be prohibited by the zero-price rule on network neutrality [4]. ISPs do not have the right to make CPs pay a termination fee for access to Internet consumers; the residential consumers pay uniquely to be connected to an ISP and the CPs pay to be connected to the network. This rule implies that there is a "missing price" prohibiting financial transfers between CPs and ISPs according to their explanation.

In this sense, the possibility that an ISP's differentiating charges and treatment of data will allow a better and more efficient management of the network, providing an ongoing incentive to create a more accurate service. The interaction between price and quality of service is produced when networks are under congestion. However a major source of confusion for the economic analysis comes from the different bandwidth uses of the Internet applications [5].

The presence of net neutrality addresses new issues to maintain the competitive balance between all actors, especially when one side of the network is not participating in the universal service funding. A new interpretation of universal service become a commitment to increase the benefits of the development of networks, as suggested in [6]. Net neutrality and quality of service[2] introduce a new opportunity where Universal Service could be suggested as policy over the data bandwidth.

Allowing people that where previously in unconnected areas to join the global information society has even become one of UN[3] and ITU[4] major concerns. United Nations Sustainable

[1] One interesting work is presented in [2] , in a model of economic analysis of Internet flat rates, the author explains that Internet flat rates may create network congestion problems and one way to solve network congestion is to over-provide bandwidth. If it is too costly to do, appropriate pricing schemes might be introduced to allocate bandwidth between users in a more efficient way.

[2] A good description of this context is developed by [7] .

[3] SDG-9 : Build resilient infrastructure, promote inclusive and sustainable industrialization and foster innovation [8].

[4] Universal Service is not new issue on public policy debate. However, Broadband Universal Service correspond an new issue for governments and International organizations. Indeed, as part of ITU concerns, an interesting

Development Goals could be the perfect scenario to propose ISPs and CPs to participate as global partners for sustainable development[5] through a broadband universal service policy.

As we note, content providers are not involved directly in the universal service obligations with a corresponding participation being taxes, funds or compensations. This legal vacuum over the content providers make a challenge in the research of a possible strategy compensation (price) as part of universal service policy to improve costumers welfare.

The lack of participation deepens much more when the demand growth for bandwidth is the current argument of technological innovations on broadband and the main economic winners are the content providers.

1.1. Universal service as zero price rate policy

Zero rating policy[6] could produce an externality effect over the universal service in a context of competition or vertical integration. In a competitive market some external effects could arise over the market equilibrium [7], explained in [5].

However an "unlevel playing fields" as zero price rating policy could become an opportunity for universal service access. Through this policy, some services could be offered for free (e.g. free or very small priced access to a limited version of the Internet) that could support the universal service efforts to increase the access to Internet service.

We develop a normative approach of zero price rating Internet service policy as universal service strategy. Our objective is to describe the market as it should be. Our contribution is to give some public policies insights to be considered by regulation authorities.

We focus our analyze on a scenario where a regulator decides to impose a universal service as a limited version of Internet service (bandwidth limitation)[8]. This service will be financed by subsidies[9] or by a really small price from subscriber. The ISP impose data caps on use. This data caps will shape competition between content providers because the data limitation create an artificial scarcity, making users perceive different qualities of service from each content providers as explained in [5]. [10]

Our contribution tries to answer the question about zero pricing Internet services in the context of infrastrucutre access (free or a very small economic participation to access) to a limited version of the Internet, proposed in [5]. *Does expansion of use provide a benefit that merits less concern about the competitive effects or not?*

This paper is organized as follows. Section 2 presents an application of the concerns proposed in [5], in Section 3 we start our analysis under a monopolistic ISP that provides a limited Internet service allowing a differentiated speed access to CPs that are interested to diffuse their high bandwidth content. In Section 4 we introduce the zero price rating Internet under a duopoly competition. As result we obtain a welfare superior under imperfect competition scenario. Finally, we conclude this paper.

2. THE MODEL

Following the schema presented in [5], CPs could participate with a compensation as part of universal service policy for a target population that has access to a limited Internet service and could be able to access their content with an increase of bandwidth financed by CPs [11].

Nowadays, universal service as zero price rule is not present in the Internet market structure. End users pay a price for a service provided by ISPs. CPs have no technical limitations to their content access and users have a unlimited content access. In this situation Internet and content access are perfect complements. Internet users' utility is defined by

$$u = f[b] - p_{isp}$$

The Internet service price p_{isp} depends on the technology deployed on the network (broadband, ultraboadband, etc). u represents the utility level of Internet users. We assume that $f[b]$ is a concave non decreasing function with decreasing marginal value as a preference function. b represents the acceded data volume[12] by Internet users. Therefore the price of Internet service depends on data speed capacities.

In the case of ISP, their benefits depend on the price charged, the number of final users α and the cost c_{isp} of providing the service. CP's profits depend on $r \cdot g[b]$ as the revenue, where r is a parameter that differs among the different CPs[13]. We

report published by ITU-D study group 1, entitled : question 7-3/1, "Implementation of universal access to broadband services.", presents some national initiatives and possible policies in relation with universal service and access.

[5] SDG-17: Revitalize the global partnership for sustainable development [8].

[6] The example of "zero-rating" plans offered by local mobile operators in India to roll out Internet.org Facebook's project, and the preoccupation of government regulators about content neutrality, show the inherent tension between infrastructure access, content and net neutrality. An interesting world panorama with an analysis of zero-rating debate is exposed in [9].

[7] Selected websites or Internet services rated at zero cost to the customers, would not respect the essence of net neutrality, which requires non-discrimination between different content and applications.

[8] A rustic Internet service with a minimum of quality of service defined by the regulator. A difference as well as a complement to the existing free wifi access policies, a broadband Universal service policy, focused on bandwidth (universal minimum quality in Mb/s), looks to create better absorption of technological improvements as a welfare measure on a greater part of population.

[9] Traditional universal service financing mechanisms.

[10] A possible race for innovation in data demand and content transmission from the content providers could happen in the future. However our interest is to define if this data caps delimitation become positive for consumer access under the scenario presented before.

[11] there are different strategies nowadays that CP employs to finance their business models as publicity or database collection.

[12] referring us to the symbol for bytes as a unit of data traffic

[13] [10] explain that revenues could result of advertisement or value-added services. The different types of CPs may produce much different revenue levels.

assume that $g[b]$ is a non decreasing concave function. Both profits functions are given by

$$\pi_{ISP} = (p_{isp}[b] - c_{isp}[b]) \cdot \alpha \quad (1)$$

$$\pi_{cp} = r \cdot g[b] - c_{cp}[b] \quad (2)$$

3. MONOPOLY ISP

We assume that ISP provides a limited (bandwidth)[14] Internet service to a defined population at a price p_z which is lower than the market price p_{isp}, where $p_{isp} > p_z \geq 0$. Considering universal service from the legal definition, p_z corresponds to the net cost of providing a limited bandwidth Internet service over the network. Due to the low stream bandwidth of this network, some contents from CPs will not be able by the technical limitations, however CPs have the option to be viewed via the payment of a termination fee t (financed by themselves e.g. publicity or database collection) for access to the fast lane network bitstream[15].

As explained in [5], without network neutrality ISP could give preferential treatment to their own services in fast lanes, driving out the content competitors when the ISP has a vertical integration strategy with a content provider. However the cost of a vertical integration for ISP is exogenous to Internet users at least in the first step of the strategy because zero rating price p_z under an universal service strategy is regulated and ex-ante considered. Costumer welfare is not affected and the principal objective is to increase the universal access to the Internet but the game rests between ISP and CPs.

3.1. CP optimal Traffic data

The monopolistic ISP plays the role of a gateway between CPs and Internet users to transfer data content. We define b as the acceded data from a CP by Internet users. The ISP charges a termination fee t per unit of data speed traffic to the CP that decided to propose the access to its content by a fast lane network. The CP is free to decide the kind of content or service to be acceded or displayed. The only condition to respect is the termination fee payment. Content providers chose b such that

$$\begin{aligned} b \in \arg\max_{b} \quad & v_{cp}[b] = r \cdot g[b] - b \cdot t \\ \text{s.t.} \quad & b \geq 0 \end{aligned} \quad (3)$$

Lemma 1: *Content provider choses b such that*

$$g'[b] = \frac{t}{r}$$

Proof. See the Appendix.

[14]For our research we rely on concepts and assumptions presented in [11, 12, 13, 14, 5]

[15]An interesting application of this strategy is presented by [15].

Traffic data volume is limited by CP in reason of r and t. The traffic data choice of CP will depend on the relation between termination fee t and CP revenues $r \cdot g[b]$. Indeed, the possibility of a renting strategy for CPs in this scenario is crucial for their participation in the policy.

We note that under a monopolistic ISP without a specific regulation, ISP will define $t \geq r$ as a result of its market power influence. A non incentive situation for CP to participate in the market could happen.

Lemma 2: *the variation of traffic data volume for CP b affected by a variation of termination fee (t) at the optimum is given by*

$$b'_{cp}(t) = \frac{1}{r \cdot g''[b]} \quad (4)$$

Proof. See the Appendix.

3.2. ISP optimal termination fee

Fixed by regulations, zero rating price p_z corresponds to net cost of providing a limited bandwidth Internet service over the network. Without regulation on termination fee t, monopolistic ISP maximizes its profits in function of CP revenues, impacting CP termination fee. Thus, the total profits of ISP is given by

$$\begin{aligned} \max_{t} \quad & \pi_{ISP} = (p_z - c_z)\alpha + (t)\, b^*[t] \\ \text{s.t.} \quad & t \geq 0 \end{aligned} \quad (5)$$

Where c_z corresponds to the cost per-subscriber. ISP determines the optimal price of t by $\max_t \quad \pi_{ISP} \quad s.t. \quad t \geq 0$.

Theorem 1: *with the optimal traffic data volume for CP and under universal service condition $p_z = c_z$, considered as the net cost of providing a limited bandwidth Internet service over the network, the optimal termination fee (t) charged by the monopolistic ISP to the CP :*

$$t^* = \frac{-b[t]}{b'[t]}$$

$$t^* = -b \cdot r \cdot g''[b] \quad (6)$$

The optimal termination fee for the monopolistic ISP is positive $t^* > 0$ because $g[b]$ is concave: $g''[b] < 0$. If monopolistic ISP tries to discriminate CPs, ISP will try to reduce the gap between the optimal traffic data volume b^* and termination fee t^* , by pricing the latter at the same level as the revenue of CPs. [16]
Under these circumstances, CPs would not participate into the market. Hence, ISP determines t^* considering only the value of $b[t]$ and not the revenue level $r \cdot g(b)$ of CP .

Proof. See the Appendix.

[16]An interesting methodology to calculate the benefits of sponsored data for an individual content provider is proposed by [16].

Theorem 2 *the optimal data traffic of CPs in terms of $g[b]$ is given by substitution of* (6) *in lemma 1*

$$b^* = -\frac{g'[b]}{g''[b]}$$

As an intuitive result, we can state that the higher the marginal gains, the better the quality. The higher the gap between r and t , the bigger the incentive for CPs to participate. Given $t^*$and $b^*[t]$ conditions, termination fee is inelastic.

$$\epsilon_t = -\frac{t}{t^{*2}}$$

Proof. See the Appendix.

3.3. Universal service Internet as a substitute service

Now, we study the effect of universal service adoption as a zero rating price policy concomitantly to a normal Internet service. ISP offers both a normal Internet Service and a limited Internet service based on a zero price rate. We denote two types of contracts that users could choose, α and β corresponding to the number of users of each service. c_z and c_{isp} correspond to the cost per-subscriber of a limited Internet Service and a normal Internet Service respectively. p_z and p_{isp} corresponding to each service price of service

$$\begin{aligned} \max_t \quad & \pi_{isp} = (p_{isp} - c_{isp})\,\alpha + (p_z - c_z)\,\beta + t \cdot b[t] \\ \text{s.t.} \quad & t \geq 0 \end{aligned}$$

Following the legal definition of universal service price, the net cost of providing the service is considered as the price of the limited Internetg, $p_z = c_z$.

Theorem 3 *under contract substitution, the optimal termination fee t is*

$$t^* = \frac{-b[t]}{b'[t]}$$

Proof. See the Appendix.

The cost of the limited Internet service (c_z) could be altered in the next steps by the monopolistic ISP as a result of a possible Internet users contract shifting. ISP could argue a loss of benefits as a result of the shifting of users. However as an *ex-ante* policy and regulated tool, the monopolistic firm cannot change the price of the limited Internet service p_z. As $p_z = c_z$, the limited Internet service net cost rests unchangeable. Only if the contract between ISP and CP is defined as a dynamic termination fee t, monopolistic ISP will re calibrate t to include the lost benefits of Internet users shifting. The CP will reevaluate their optimal data traffic just to the equilibrium.

$$\begin{aligned} \max_t \quad & \pi_{isp} = (p_{isp} - c_{isp})\,\alpha + (p_z - c_z)\,(\beta + (1-\alpha)) \\ & + \left(1 + \frac{(p_{isp} - p_z)\,(1-\alpha)}{\beta + (1-\alpha)}\right) t \cdot b[t] \\ s.t. \quad & t \geq 0 \end{aligned}$$

In this case, optimal termination fee t is the same as the initial situation: $t^* = \frac{-b[t]}{b'[t]}$.

Proof. See the Appendix.

Proposition 1 The termination fee t defined by ISP has to be the same no matter if it is a unique or multiple contract strategy even when there are shifting Internet users contracts, i.e. as a monopoly. ISP do not make difference in the allocation of termination fee t under different contracts. The contracts are treated separately under a price regulation p_z.

4. DUOPOLY

4.1. Symmetric ISP behavior

Universal service adoption as a limited Internet service based on a zero price rate in a broadband duopoly could produce a different behavior from ISPs. A first case is a symmetric situation, where both ISP operate with a Normal Internet Service and propose the limited Internet service within their contracts. This case reflects a situation where the limited service is an open regulated policy with defined specifications provided by the regulaton authorities to any ISPs. However if one of the ISPs decides not to provide this kind of contract, the second will take advantage of this situation, becoming a monopoly of the limited Internet Service.

We consider two ISPs defined as $ISP_i = 1, 2$, competing ISP set termination fees $t_i, i = 1, 2$ to maximize profits. ISPs' bandwidth are homogeneous, perfect substitutes. CPs prefer the ISP with the lowest price as long as $v_{cp}[b] \geq \min\{t_1, t_2\}$. If $v_{cp}[b] < \min\{t_1, t_2\}$, CPs do not have incentives to participate in a limited Internet service policy as the result of higher termination fees. For equal termination fees $v_{cp}[b] \geq t_1 = t_2$ CPs are indifferent to the choice between ISPs.

$$b_{cp}(t_1, t_2) \begin{cases} b_{cp}(t_1) & \text{if } t_1 < t_2 \\ \frac{1}{2} b_{cp}(t_1) & \text{if } t_1 = t_2 \\ 0 & \text{if } t_1 > t_2 \\ \\ b_{cp}(t_2) & \text{if } t_2 < t_1 \\ \frac{1}{2} b_{cp}(t_2) & \text{if } t_2 = t_1 \\ 0 & \text{if } t_2 > t_1 \end{cases}$$

Under a Bertrand's paradox, $t_1 = t_2 = c_b$. Nash equilibrium in prices is determined in termination fee (t_1^*, t_2^*)

Proposition 2 If ISPs adopt the same strategy over a limited Internet service policy, under a Bertrand duopoly, ISPs define termination fee t at marginal cost of data traffic. The long-run marginal cost of increasing bitstream data traffic define the price of t. Additionally, termination fee t becomes minimal due to the fact that a big cost componet of providing service is including in the net cost c_z. In a perfect competition scenario $t \approx 0$.

As in the monopoly situation, ISPs define termination fee t under substitution contracts at the same value as a unique service contract. ISPs do not make difference in the allocation of termination fee t under the presence of different contracts. The contracts are treated separately under a price regulation p_z.

4.2. Universal service Internet over an exclusivity ISP

An asymmetric situation could by presented when only one ISP provides the limited Internet service. This scenario reflects the case where the policy is targeted to only one ISP. In several cases, we see in the literature that universal service obligation is provided by the incumbent or as part of a contract specifications defined to a new player in a market. In the same case, ISP that proposes the limited Internet service take advantage of this situation in a monopoly position. In this scenario the market power of ISP is no lower than that of a monopolistic ISP, because it is the only provider of this contract. In fact, ISP may compete to obtain the exclusivity to offer the service by public bidding or paying the competitors to do not provide this service[17].

Exclusive ISP contracts to provide a limited Internet service induce a monopolistic behavior of ISP. The effect of Exclusivity contracts have to be evaluated over a welfare comparison.

4.3. Welfare Analysis

According to the legal definition of the universal service policy $p_z = c_z$ the price of the subscription to this Internet service will be defined as the net cost of delivering the limited Internet technology. The variation of costumers' welfare in this situation is highly positive because $p_z < p$. The total social welfare is determined by

$$SW^z_{monopoly} = r \cdot g\left[b\right] - c_{cp} + \Delta CW$$

Under a duopoly with a symmetric ISP behavior that provide a limited Internet service, termination fee is defined as $t_1 = t_2 = c_b$ in a Nash equilibrium in termination fee prices (t_1^*, t_2^*) and according to the universal service condition $p_z = c_z$. The total social welfare is determined by

$$SW^z_{duopoly} = r \cdot g\left[b\right] - c_{cp} + \Delta CW$$

Proposition 3 Duopoly scenario is welfare superior than monopoly scenario. Indeed $r \cdot g\left[b\right]$ is a function of t because $b = b\left[t\right]$. With the lowest pricing of termination fee, the revenue level $r \cdot g\left[b\left[t\right]\right]$ will be higher on duopoly.

$$SW^z_{monopoly} < SW^z_{duopoly}$$

Proof. See the Appendix.

[17] This kind of practices are allowed under a regulated market, at least not under the presence of a regulator.

Under duopoly competition, the ISPs define termination fee t above their marginal cost. Limited Internet service in duopoly has a positive impact on end users welfare. indeed following theorem 2: the higher the marginal gains, the better the quality. With a reduction of t in relation to monopolistic ISPs, CPs will allow more content financed by themselves in the limited Internet service. Users will access more content without any extra cost.

5. CONCLUSION

We presented what we believe is a possible alternative for universal service on broadband: a kind of alternative Internet contract with a free or very small economic participation to access a limited version of the service. CP finance their participation to broadcast their high bandwidth content. We focused our analysis on a scenario where a regulator decides to impose a universal service as a limited version of the Internet service through a bandwidth limitation, financing the Internet access by subsidies or a really small price from subscriber.

The objective of this paper was to contribute to the literature with some public policies insights to be considered by regulation authorities. We have developed a very simplified model that may overlook some issues. First, we developed a normative static approach, we described the market as it should be and not as it is. Second, the vertical integration between ISPs and CPs was not included as an issue because price under a universal service strategy is regulated and *ex-ante* considered. Vertical integration becomes a dynamic evolution of data caps universal service strategy.

As mainly results we show:

- There is a positive effect to propose a Broadband universal service by quality as complement of Universal Service obligation at least in terms of access due to the reduction of subscription price to a limited bandwidth Internet service over the network.
- Limited Internet service price has to be calculated and fixed by regulations, taking into account the technical components of the service and not the possible loss of benefits or opportunity costs for Internet service providers. ISP will try to defend or justify their position in order to extract all possible benefit of this policy.
- Network neutrality policy has to be imposed to ISP in terms of termination fee t over the different CPs. Indeed, if monopolistic ISP tries to discriminate CPs, ISP will try to reduce the gap between the optimal traffic data volume pricing termination fee at the value of revenue level of CP.
- Duopoly competition scenario is welfare superior than monopoly scenario. Zero price rating Internet service in duopoly competition has a positive impact on end users' welfare.

- If this policy is to be proposed by the regulation authority, it has to be considered in a competition market structure. A reduced termination fee will be granted to CPs and more content financed by themselves will be available to end users.

In conclusion, the expansion of this policy provides benefits in terms of participation of indirect financing of universal service. Indeed, CPs are not involved in the universal service programs with a corresponding participation being taxes, funds or compensations.

Defining a limited Internet service as a policy to increase Internet penetration, the benefits merit less concern about the competitive effects, specially in regions where a large income difference between their population is present.

Appendix

Proof of Proposition 1: ***optimal traffic data volume for CP, on page 3***

We consider the following payoff from content provider:

$$\begin{aligned} \arg\max_{b} \quad & v_{cp}[b] = r \cdot g[b] - b \cdot t \\ \text{s.t.} \quad & b \geq 0 \end{aligned}$$

where b is the optimization variable, v_{cp} is the utility function for content provider, in order to maximize $v_{cp}[b]$, we define a Lgagrange's function on Kuhn-Tucker conditions:

$$\mathscr{L} = r \cdot g[b] - b \cdot t - \lambda(-b)$$

The first order condition is defined by:

$$\frac{\partial \mathscr{L}}{\partial b} = r \cdot g'[b] - t + \lambda = 0 \tag{7}$$

complementary slackness Kuhn-Tucker conditions for a point to be a maximum are:

$$\lambda(-b) = 0 \tag{8}$$

$$\lambda \geq 0 \tag{9}$$

Solutions of (7) for λ are given by

$$\lambda = t - r \cdot g'[b] \tag{10a}$$

$$g'[b] = \frac{t}{r}$$

Proof of Lemma 2: ***optimal traffic data volume for CP, on page 3***

If $t - r \cdot g'[b] > 0 \Longrightarrow \lambda > 0$. the condition $-b = 0$ is strictly necessary to obtain $\lambda > 0$, according equation (8). In other words, the constraint is binding.

In contrast, if $t - r \cdot g'[b] \leq 0 \Longrightarrow \lambda = 0$. According to equation (10a) where $\lambda = 0$, and replacing λ in equation (7), we get:

$$\underbrace{r \cdot g'[b] - t}_{h} = 0 \tag{10b}$$

Following the implicit function theorem where :

$$\frac{db}{dt} = -\frac{\frac{\partial h}{\partial t}}{\frac{\partial h}{\partial b}} \Longrightarrow \frac{db}{dt} = \frac{1}{r \cdot g''[b]} \tag{11a}$$

The exact value of b in equation (11a) is $b[t]$, according to equation (10b), b depends on t at the equilibrium. However we note it only b to simplify concepts, in a strict sens the exact value is defined as:

$$\frac{db}{dt} = \frac{1}{r \cdot g''[b[t]]} = b'[t] \tag{11b}$$

From equation (11b) we can get:

$$\frac{d^2b}{dt^2} = -\frac{b'[t] \cdot g'''[b[t]]}{r \cdot g''[b[t]]^2} \tag{12}$$

Replacing (11a) in equation (12):

$$\frac{d^2b}{dt^2} = -\frac{\frac{1}{r \cdot g''[b]} \cdot g'''[b]}{r \cdot g''[b]^2} = -\frac{g'''[b]}{r^2 \cdot g''[b]^3} = b''[t] \tag{13}$$

Proof of Theorem 1: ***optimal termination fee (t) monopolistic ISP, on page 3***

We now optimize the ISP program under the assumption of a concave function:

$$\pi_{ISP} = t \cdot b[t] + (p_z - c_z)\alpha \tag{14}$$

Under universal service condition p_z is defined as the net cost to provide a limited bandwidth Internet service over the network. $p_z = c_z$. Replacing in (14), we get:

$$\pi_{ISP} = t \cdot b[t]$$

$$\frac{\partial \pi_{ISP}}{\partial t} = b[t] + t \cdot b'[t] \tag{15}$$

We obtain $t^*$from equation (15):

$$t^* = -\frac{b[t]}{b'[t]}$$

Replacing (11a) in (15) we obtain:

$$t^* = -b \cdot r \cdot g''[b] \tag{16}$$

$t > 0$ because $g[b]$ concave $g''[b] < 0$.

In the same way, the second derivative of π_{isp} is simplified by (11b) and (13) .

$$\frac{d^2\pi_{isp}}{dt^2} = \frac{1 - t'}{r \cdot g''[b]} - \frac{g'''[b]}{r^2 \cdot g''[b]^3} t$$

Where $g'''[b]$ depends on the capacity of CP to find more clients in their business model. We make the assumption $g'''[b] < 0$ because in the first step, CP works with a very defined clients base. CPs improve their strategies in the further steps. As result $\frac{d^2\pi_{isp}}{dt^2} < 0$.

Proof theorem 2: ***Optimal data traffic CP in terms of*** $g[b]$ ***, on page 4***

$$g'[b] = \frac{-b{\cdot}r \cdot g''[b]}{r}$$
$$b = -\frac{g'[b]}{g''[b]}$$

The price elasticity of CP is defined by:

$$\epsilon_t = \frac{\frac{\partial b_{cp}[t]}{b_{cp}[t]}}{\frac{\partial t}{t}} \tag{17}$$

Replacing (15) and (11b) on (17), we obtain:

$$\epsilon_t = \frac{\frac{\frac{1}{r\cdot g''[b]}}{b}}{\frac{-b\cdot r\cdot g''[b]}{t}} = \frac{\frac{1}{b\cdot r\cdot g''[b]}}{\frac{-b\cdot r\cdot g''[b]}{t}} = -\frac{t}{(b\cdot r\cdot g''[b])^2} = -\frac{t}{t^{*2}}$$

Proof of Theorem 3 : ***optimal termination fee (t) from ISP under Zero rating price as an option to subscriber,*** **on page 4**

$$\max_t \quad \pi_{isp} = (p_{isp} - c_{isp})\,\alpha + (p_z - c_z)\,\beta + t \cdot b[t] \tag{18}$$
$$\text{s.t.} \quad t \geq 0$$

termination fee t is measured in function of c_z and c_{isp} corresponding respectively to the cost per-subscriber of a limited Internet service and to the normal Internet service. p_z and p_{isp} corresponding to each service price. Under the universal service legal definition, price is measured as the net cost of provide the service, $p_z = c_z$. Simplifying 18 we obtain:

$$\pi_{isp} = (p_{isp} - c_{isp})\,\alpha + t \cdot b[t]$$

where:

$$\frac{\partial \pi_{isp}}{\partial t} = b[t] + t \cdot b'[t] \tag{19}$$

From 19 we obtain t^*:

$$t^* = -\frac{b[t]}{b'[t]}$$

The optimal termination fee t^* does not change under the distribution of a normal Internet service.

As a second case, the optimal termination fee (t) ISP under the consideraton that monopoly benefits lost by the Internet users shifting is defined as :

$$\max_t \quad \pi_{isp} = (p_{isp} - c_{isp})\,\alpha + (p_z - c_z)\,(\beta + (1-\alpha)) \tag{20}$$
$$+ \left(1 + \frac{(p_{isp} - p_z)\,(1-\alpha)}{\beta + (1-\alpha)}\right) t \cdot b[t] \tag{21}$$
$$s.t. \quad t \geq 0$$

Under universal service condition p_z is defined as the net cost to provide a limited bandwidth Internet service over the network. $p_z = c_z$. Replacing in (20), we get:

$$\pi_{ISP} = \left(1 + \frac{(p_{isp} - p_z)\,(1-\alpha)}{\beta + (1-\alpha)}\right) t \cdot b[t]$$

$$\frac{\partial \pi_{ISP}}{\partial t} = \left(1 + \frac{(p_{isp} - p_z)\,(1-\alpha)}{\beta + (1-\alpha)}\right) b[t] \tag{22}$$
$$+ \left(1 + \frac{(p_{isp} - p_z)\,(1-\alpha)}{\beta + (1-\alpha)}\right) t \cdot b'[t] \tag{23}$$
$$\tag{24}$$

We obtain t^* from equation 22:

$$t^* = -\frac{b[t]}{b'[t]}$$

Proof of Welfare Analysis, on page 5

Under this scenario, Internet users utility is defined by:

$$u_z[b] = f[b] - p_z[b]$$

Corresponding the universal service policy with the legal definition $p_z = c_{isp}$ the price's subscription to this Internet service will defined as the net cost of delivering the limited Internet technology. Welfare superior is present in all the cases and highly positive as result of $p_z < p$. Variation of social welfare is $\Delta CW > 0$

The benefits of CPs and ISP are defined as:

$$\pi_{cp} = r \cdot g[b] - c_{cp} - (t)\,b^*[t]$$

$$\pi^z_{ISP} = (p_z - c_z)\,\beta + (t)\,b^*[t]$$

Under a monopolistic ISP that provides a limited Internet service, the total welfare is determined by:

$$\begin{aligned} Wz &= \pi_{cp} + \pi^z_{isp} \\ &= (r \cdot g[b] - c_{cp} - (t)\,b^*[t]) + ((p_z - c_z)\,\beta + (t)\,b^*[t]) \\ &= r \cdot g[b] - c_{cp} + (p_z - c_z)\,\beta \end{aligned} \tag{25}$$

Under universal service condition p_z is defined as the net cost to provide a limited bandwidth Internet service over the network. $p_z = c_z$.

$$Wz = r \cdot g\left[b\right] - c_{cp}$$

The total social welfare is determined by:

$$SW = W_z + \Delta CW$$

$$SW = r \cdot g\left[b\right] - c_{cp} + \Delta CW \tag{26}$$

Under a duopoly with an asymmetric ISP behavior that provides ZPR, the total welfare is determined by:

$$\begin{aligned} Wz &= \pi_{cp} + \pi^z_{isp_1} + \pi^z_{isp_2} \\ &= \left(r \cdot g\left[b\right] - c_{cp} - \left(t\right) b^*_{1,2}\left[t\right]\right) \\ &+ \left(\left(p^1_z - c_z\right)\beta + \frac{1}{2}\left(t_1\right) b^*\left[t\right]\right) \\ &+ \left(\left(p^1_z - c_z\right)\beta + \frac{1}{2}\left(t_2\right) b^*\left[t\right]\right) \end{aligned}$$

Under a Bertrand's paradox, $t_1 = t_2 = c_b$ a Nash equilibrium in prices is determined in termination fee (t^*_1, t^*_2) and under universal service condition p_z is defined as the net cost to provide a limited bandwidth Internet service over the network. $p_z = c_z$.

$$Wz = r \cdot g\left[b\right] - c_{cp}$$

The total social welfare is determined by:

$$SW = W_z + \Delta CW$$

$$SW = r \cdot g\left[b\right] - c_{cp} + \Delta CW \tag{27}$$

However revenue level for CP in duopoly scenario is superior than that of monopoly scenario. Indeed $r \cdot g\left[b\right]$ is a function of t because $b = b\left[t\right]$. With a lowest price of termination fee, the revenue level $r \cdot g\left[b\left[t\right]\right]$ will be higher on duopoly.

$$SW^z_{monopoly} < SW^z_{duopoly}$$

REFERENCES

[1] Steven Bauer, David D Clark, and William Lehr, "Understanding broadband speed measurements," Tprc, 2010.

[2] Marc Bourreau, "The economics of internet flat rates," *Communications & Strategies*, vol. 42, no. 2, pp. 131–152, 2001.

[3] Soumya Sen, Carlee Joe-Wong, Sangtae Ha, and Mung Chiang, "A survey of smart data pricing: Past proposals, current plans, and future trends," *ACM Computing Surveys (CSUR)*, vol. 46, no. 2, pp. 15, 2013.

[4] Sebastien Broos and Axel Gautier, "Competing one-way essential complements: the forgotten side of net neutrality," *Available at SSRN*, 2015.

[5] Shane Greenstein, Martin Peitz, and Tommaso Valletti, "Net neutrality: A fast lane to understanding the trade-offs," Tech. Rep., National Bureau of Economic Research, 2016.

[6] James Alleman, Paul Rappoport, and Aniruddha Banerjee, "Universal service: A new definition?," *Telecommunications Policy*, vol. 34, no. 1-2, pp. 86–91, 2009.

[7] Eitan Altman, Julio Rojas, Sulan Wong, Manjesh Kumar Hanawal, and Yuedong Xu, "Net neutrality and quality of service," in *Game theory for networks*, pp. 137–152. Springer, 2011.

[8] Economic and Social Council, "Progress towards the sustainable development goals," Tech. Rep., United Nations, 2016.

[9] Arturo J Carrillo, "Having your cake and eating it too? zero-rating, net neutrality and international law," *Stanford Technology Law Review, Forthcoming*, 2016.

[10] Liang Zhang and Dan Wang, "Sponsoring content: Motivation and pitfalls for content service providers," in *Computer Communications Workshops (INFOCOM WKSHPS), 2014 IEEE Conference on*. IEEE, 2014, pp. 577–582.

[11] Nicholas Economides, "The economics of the internet backbone," *NYU, Law and Economics Research Paper*, , no. 04-033, pp. 04–23, 2005.

[12] John Musacchio and Dohoon Kim, "Network platform competition in a two-sided market: implications to the net neutrality issue," in *TPRC: Conference on Communication, Information, and Internet Policy*, 2009.

[13] Nicholas Economides and Benjamin E Hermalin, "The economics of network neutrality," *The RAND Journal of Economics*, vol. 43, no. 4, pp. 602–629, 2012.

[14] Frago Kourandi, Jan Krämer, and Tommaso Valletti, "Net neutrality, exclusivity contracts, and internet fragmentation," *Information Systems Research*, vol. 26, no. 2, pp. 320–338, 2015.

[15] Hassan Habibi Gharakheili, Arun Vishwanath, and Vijay Sivaraman, "An economic model for a new broadband ecosystem based on fast and slow lanes," *IEEE Network*, vol. 30, no. 2, pp. 26–31, 2016.

[16] Mark Andrews, Glenn Bruns, and Hyoseop Lee, "Calculating the benefits of sponsored data for an individual content provider," in *Information Sciences and Systems (CISS), 2014 48th Annual Conference on*. IEEE, 2014, pp. 1–6.

QUALITY AND STANDARDIZATION IN TECHNOLOGY-ENHANCED LEARNING

Irina Tal[1], Eva Ibarrola[2], Gabriel-Miro Muntean[3]

[1]Dublin City University, irina.tal2@mail.dcu.ie
[2]University of the Basque Country-UPV/EHU, eva.ibarrola@ehu.eus
[3]Dublin City University, gabriel.muntean@dcu.ie

ABSTRACT

Education and technological advances have enabled digital learning technologies to become a key pillar holding up one of the main United Nations Sustainable Development Goals (SDGs): Quality Education. Technology-enhanced learning strategies have led to innovative ICT applications for the development of new improved learning and teaching practices aiding to guarantee inclusive and equitable quality education and promote opportunities for all. Nonetheless, there is still a big challenge to get hold of all the capabilities of these technology-enhanced learning strategies: improving the learner quality of experience (QoE). This paper presents an innovative technology-enhanced learning initiative that aims to attract students to STEM education and improve access for students with disabilities through the analysis of the learner QoE. Moreover, in this paper some of the required standards and specifications to be used for developing this initiative are identified. Another key contribution of the paper is that it explores the necessity of new related standards and introduces some novel proposals for standardization in this area.

Keywords— Technology-enhanced learning, QoE, mulsemedia, standardization.

1. INTRODUCTION

The convergence of education and technological innovations are motivating digital learning to become one of the most significant models in all the teaching and learning environments. In addition, the rapid advance of new multidimensional applications and services, with new and modern functional facilities, like multisensory dimensions or augmented reality, drive growth in new technology-enhanced learning (TEL) strategies with significant capabilities to improve the quality and outcomes of teaching and learning. However, in order to make a whole of all the features and capabilities of these novel digital learning systems, it is crucial to analyze and enhance the learner QoE/.

A lot of studies and research have been carried out during the past years to define QoE evaluation methods [1, 2] and also to define and validate related standards [3, 4]. Nevertheless, as stated in [5]: "*With the growing level of functional sophistication of services and systems, quality evaluation has become progressively more complex, notably due to the exponentially mounting number of dimensions involved. While some services and systems involve a small number of technologies, functional capabilities, sensory dimensions and consuming paths, other applications are an explosive cocktail of technologies, capabilities, navigation courses and sensations. These growing application scenarios involve larger user susceptibility, not only to the degree of usability of the overall system but also to the content itself*". Therefore, the target of evaluating QoE to enhance the learner experience when using, for example, new mulsemedia (Multiple Sensorial Media Advances and Applications) technology-enhance learning applications [6, 7] may become a really hard task.

Even though some standardization activities related to general technology-enhanced learning have been defined [4] (in particular for mobile technologies and video), there is still a need to fulfill the gap of novel standards to define frameworks and procedures for the QoE evaluation of these new multimedia and mulsemedia TEL applications and systems.

In this paper, it is presented an initiative, NEWTON project [8], that aims to develop, integrate and disseminate innovative TEL methods and tools. The main goal of this project is to create new or inter-connect existing state-of-the art teaching labs and to build a pan-European learning network platform. Next generation TEL methods are aimed to be integrated in the NEWTON platform that relate to gamification, augmented reality, mulsemedia, and adaptive multimedia. The project will be tested and validated in real life pilots across Europe, using a network of secondary schools, vocational establishments and third level institutes. One of the initial pilots that will be carried out in the context of this project is presented in more details in this paper. This pilot study will be carried out in November 2016, in the University of the Basque Country (UPV/EHU).

Since the platform aims to integrate also others already existing teaching labs, it is a key target of the project to develop the system complying with the existing standards in this area. For that reason, a revision of the most important related standards that should be taken into account when designing, testing and deploying the platform

is also included in the paper. Moreover, some novel proposals to fill the gap of standardization related to the analysis and evaluation of the learner QoE in these TEL systems and scenarios are presented.

The structure of the paper is as follows. In section 2 a review of recent literature on TEL strategies and related standardization activities is included. Section 3 presents a general description of NEWTON, the project that aims to build a novel TEL-based learning platform. Novel related standardization proposals for the enhancement of the QoE evaluation in this kind of platforms will be described in section 4. In section 5, the proposed pilot test bed to be deployed at the University of the Basque Country (UPV/EHU) and the expected results will be detailed. In the last section, main conclusions and contributions of this work are summarized.

2. BACKGROUND

One of the goals of the pan-European learning network platform proposed in NEWTON project is to be developed in a way that makes it possible integrate and interconnect different teaching platforms, existing teaching platforms or newly developed in order to support fast dissemination of STEM (Science, Technology, Engineering and Mathematics) learning content to a large audience and also to improve the access to the education for the people with disabilities in particular. Harmonization, alignment and compliance with related standards will promote this interconnection and facilitate integration of other TEL systems in the platform.

Therefore, in this section, a revision of some of the European initiatives in technology-enhanced learning TEL methods and tools is included. Some of them could be considered for integration in the NEWTON platform. On the other hand, the most important related standards in this area to be complied with are also identified and described as well as in the field of the QoE evaluation for TEL.

2.1. Innovative initiatives in TEL platforms

A lot of studies and research have been conducted over the past few years focusing on the analysis and benefits of TEL [9, 10]. Nevertheless, few of them were successful in developing real functional platforms to fulfill concrete objectives, like the ones proposed in NEWTON project (encouraging young students to engage to STEM education or improving access for students with disabilities). Even fewer proposals can be found that relates to enhancing the learner QoE to reach those objectives.

One of the most remarkable initiatives in this context is the the Go-Lab Project (Global Online Science Labs for Inquiry Learning at School) [11]. In this project the goal is to promote student´s engagement to science topics through the access to online laboratories. In this way, students may conduct their own experiments and analysis learning from their own experience and motivating them to make an STEM career in the future. For this purpose, the Go-Lab Portal [12] (Figure 1) has been developed "*to provide access to a set of online labs from worldwide renowned research organizations, such as European Space Agency (ESA, the Netherlands), European Organisation for Nuclear Research (CERN, Switzerland), Núcleo Interactivo de Astronomia (NUCLIO, Portugal), as well as multiple universities and institutions. These online labs can be used by universities, schools, instructors, students and lifelong learners to extend regular learning activities with scientific experiments that can be conducted not only by teachers as a demonstration, but also by students themselves giving them real experience of scientific work.*" [12]

Another interesting initiative in this line is the Physics Education Technology (PhET) project [13] which provides simulations tools in the STEM areas over a free website: https://phet.colorado.edu/. Under this project, more than 80 interactive simulations have been developed covering different topics in Physics and also real-world applications, such as the greenhouse effect and lasers. In [14], more proposals of remote virtual laboratories with collaborative roles for learning environments are also described. It is demonstrated that this kind of virtual laboratories enhance the creativity and the natural dispositions of young people to experiment with ICT, as stated in [15], and may encourage them also to pursue their studies and a career in a STEM field.

Nevertheless, none of the above mentioned platforms makes use of the next generation of TEL strategies, like gamification, mulsemedia or interactive augmented reality teaching. In addition, in most of these platforms there is no analysis on the learner QoE and that is something really important in order to have the required feedback on the contribution of these TEL strategies and methods on enhancing the learning and teaching processes.

In this regard, some motivating studies were carried out [2, 16] to analyze how to improve the learner QoE, by means of different adaptive techniques. However, these studies did not consider the aforementioned next generation TEL strategies and methods.Very recently, new research on the QoE analysis when using this next generation strategies has been raised [17, 18] and very promising applications to enhance learning have been foreseen.

Nevertheless, there is a need to join efforts to integrate all these new advances and develop new network platforms based on innovative TEL methods and tools. This is the one of the goals of the NEWTON project.

Figure 1. The Go-Lab portal [12]

2.2. Standards for technology-enhanced learning

As mentioned in the beginning of this section, NEWTON platform aims to integrate and interconnect other existing teaching platforms. For that reason, one of the targets of the project is to be developed complying the most recent and applicable standards.

From the point of view of TEL, there is a website from the ITU (International Telecommunication Union) [19] and a report on it (see figure 2) [4] which covers in its section 5 a review of specific standardization activities for TEL. In this report it is stated that "*Learning technology standards do not include instructional design, pedagogical norms, cultural conventions, learning objectives or specific learning content"*. However, it does consider the issues that are referred in Figure 3.

In this report, it is also highlighted the importance of the Recommendation ITU-T F.742 [20] which describes application scenarios of distance learning and deduces general requirements to be met by distance learning services, and other important initiatives and groups working on the standardization of LTE. Some examples are the "ITU's Focus Group on Innovation is maintaining a living list of emerging ICT products and services" [21], the "ISO/IEC JTC1 Subcommittee 36: Information technology for learning, education and training"[22] or the "The IEEE Learning Technology Standards Committee (LTSC)"[23]. It is important also to mention, when talking about next generation technologies, the "*IMS Global Learning Consortium (IMS GLC)" [24]* a community of educational institutions, government organizations and equipment vendors which aims is to bring advances in technology in order to improve educational participation. It is also remarkable the work that is being done in LTSC related to "*Augmented Reality Learning Experience Model***".** This new LTSC working group is developing a standard model for defining AR-based learning activities. Also of interest in this context, is the "*ISO/IEC 19796-1: Learning, education and training- Quality management, assurance and metrics"*[25] defined by the ISO/IEC JTC1 Subcommittee 36.

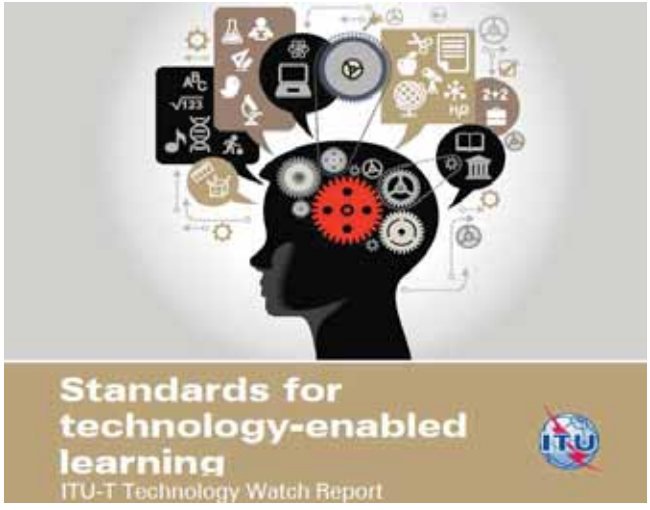


Figure 2. Standards for technology-enabled learning [4]

Figure 3. Items addressed for TEL standards [4]

In the context of more specific technical standards related to mulsemedia and other next generation TEL strategies, there are some important standardization initiatives. One example is the media context and control MPEG-V standard (ISO/IEC 23005) [26, 27] , capable of supporting mulsemedia applications, providing an architecture and specifications to enable the interoperability between virtual worlds (digital content providers of a virtual world, gaming, simulation, etc.) with the real world (sensors, actuators, vision and rendering, robotics, etc.). The multimedia content description MPEG-7 standard (ISO/IEC 15938) [28] is also to be mentioned in this context.

2.3. Standardization for QoE evaluation in TEL

Although some promising research related to the evaluation of the learner experience when using digital learning has been done [2, 16, 29], no standardization attempts for defining frameworks and procedures, as to have different comparable results, have been published so far. Likewise, remarkable efforts can be identified in the research literature that relates to the evaluation of the QoE when using mulsemedia (figure 4) and other next generation TEL strategies [6, 7, 17, 30-32], but still neither general nor specific standards have been developed in this area.

However, the interest for the standardization that relates to QoE in general us er applications is well evidenced by the increasing number of dedicated standards [33-42], whereas the need for specific standardization for QoE evaluation in TEL is underlined in [43]: "*The absence of a standard, comprehensive approach to evaluating technology-enhanced learning (TEL) limits the utility of individual evaluations, and impedes the integration and synthesis of results across studies."*

Therefore, future work and novel proposals are needed to standardize a framework and procedures for next generation TEL strategies to better understand how new generation learners would react to such experiences. Furthermore, it would be of high value if such efforts had also spread out to define and progress standards for unified architectures of mulsemedia, gamification, augmented reality and other next generation strategies for interoperability between real and virtual worlds. For a plus of value, these proposals could be accompanied with the definition of QoE metrics (KQI and KPI) for the evaluation of the quality of service (QoS) in these new emerging scenarios.

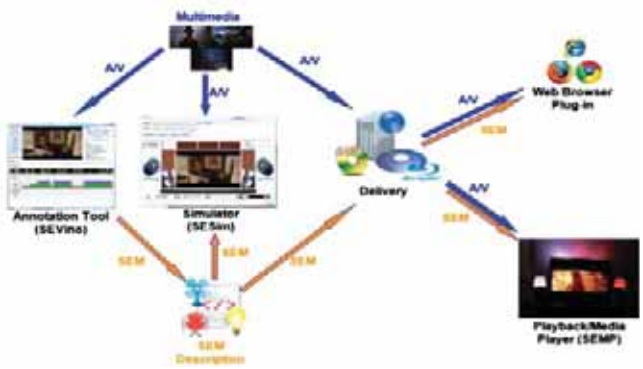


Figure 4. Capturing end-to-end QuaSE [7]

Figure 5. Horizon 2020 NEWTON Project and Partners

3. INNOVATIVE TECHNOLOGY-ENHANCED LEARNING INITIATIVE: NEWTON PROJECT

Networked Labs for Training in Sciences and Technologies for Information and Communication (NEWTON[1]) project is a Horizon 2020 funded project that brings together academia and industry partners from 7 different European countries (Figure 5). NEWTON aims to provide a pan-European learning platform that facilitates the delivery of STEM subjects to learners from a variety of backgrounds: secondary and vocational schools, third level education, people with disabilities. This pan-European platform will integrate a set of distributed labs: existing state-of-the-art teaching labs (e.g. FabLabs[2]) and newly created teaching labs as a result of the project. Moreover, the platform will be designed to be open to extension: the platform will allow for an easy integration of other new teaching labs created by different third-parties. The purpose is to facilitate the access of a large audience, including people with disabilities, to a large database of learning content that will be experienced in novel manners in order to enhance their STEM education.

Consequently, one of the main focuses of NEWTON project is to develop innovative TEL methods that will be integrated in the platform. The purpose of employing these novel methods is to increase learner QoE, improve learning process and increase learning outcome.

Such TEL methods that are aimed to be developed in the context of NEWTON project relates to:

- learner model-based personalisation
- gamification
- self-directed and independent learning
- augmented reality
- multimedia and multi-sensorial (mulsemedia) content delivery
- adaptation of content delivery to learner operational environment: variation of network conditions, user device characteristics and user profile.

The last two listed innovative technologies are some of the main novelties of NEWTON project and represent the focus in this particular work. NEWTON aims to employ adaptive solutions for the delivery of multimedia content. This involves monitoring and content adjustment based on the combined effect of dynamic variation of network conditions, user device characteristics and user profile.

Mulsemedia is considered a new type of multimedia that unlike classic multimedia that usually involves two senses (audio/video), involves three or more human senses (olfactory, haptic, etc.). NEWTON project aims to provide the learners with mulsemedia experience in order to enhance their learning experience. Additionally a highly innovative solution for multiple-sensorial content adaptation is envisaged to be proposed and used.

NEWTON platform will be validated through European-wide pilots carried out in secondary and vocational schools and universities from 5 European countries, including learners with special needs.

4. NOVEL STANDARDIZATION PROPOSALS FOR QOE EVALUATION IN TEL

As identified in the review presented in section 2, there is an important lack of standardized methodologies for the evaluation of the learner QoE, in TEL context. Nowadays, TEL can be supported by a large plethora of technologies, from traditional digital technologies, like web-based technologies or streaming multimedia systems, to other next generation technologies, like mulsemedia, gamification or augmented reality. Therefore, defining a framework and methodology for the evaluation of the learner QoE in such a complex scenario is a big challenge.

In this section, a novel proposal for the QoE evaluation and standardization in TEL context is introduced that aims to do some steps forward towards achieving the aforementioned goal. However, this work is in early stages and, although evaluating QoE has also theoretical dimensions, is the later practical validation that will confirm the validity of the approach. At this point, some early guidelines to be followed in order to define the proposed standards will be described, only. These guidelines were defined based on general QoE frameworks [44].

[1] http://www.newtonproject.eu/
[2] https://www.fablabs.io/

As a first step, it will be important to identify the factors influencing QoE in TEL. ITU-T Rec. G.1031 [38] may be a good example of this approach. As referred in this recommendation, QoE influence factors are usually grouped in three main categories (#1 to #3 in Figure 6):

- User influence factors
- Context influence factors
- System influence factors

In the first category, some key factors like the learner expectations, previous good or bad experiences in TEL and the learner expertise with the subject and all the cultural aspects should be identified (student segment A or B in Figure 6). The context, in which the learning is developed, like location, type of task, level of interaction, must be also considered when analyzing the QoE of the learner (#2 in Figure 6). Finally, the system influence factors (#3 in Figure 6) must be determined. Being aware that the sort of system influence factors may vary a lot depending of the strategies that are being implemented in the TEL process (just web browsing, streaming, gamification, mulsemedia, etc.), our proposal is that the standard for determining the QoE factors in TEL should describe the different Key Quality Indicators (KQI) and the related Key Performance Indicators (KPI) that must be considered when using the different technologies that can be implemented (#4 to #6 in Figure 6). Separate standardization documents about subjective testing for each of the different technologies or strategies will be necessary to be developed and the global standard would only include the reference to each of them.

Once, the standard for the "QoE factors in TEL" is defined, the next step would be the definition of the standard that describes the "Framework for the QoE evaluation methodology for TEL" (#7 to #9 in Figure 6).The subjective model should be defined in the standard and some generalized proposals for satisfaction models found in literature [45] (#7 to #9 in Figure 6) could be used.

Figure 6 illustrates the importance of the evaluation of the learner experience as an input used in the update and enhancement of TEL systems in order to improve the student´s learning process. Another thing that we want to highlight in relation to Figure 6 is the fact that the update in student´s expectations should be taken into account along the new generation of students. In addition, it must be considered that other factors that differ from the TEL management system factors may influence in the QoE results.

Nevertheless, as mentioned before, this entire proposal should be validated and this goal is intended to be achieved through the network of real life pilots to be developed across Europe in the NEWTON project that will give us the opportunity to verify how other context´s and student´s factors may have influence in the results.

In the next section one of the fist testing experiments to be developed in the University of the Basque Country (UPV/EHU) in Spain will be presented.

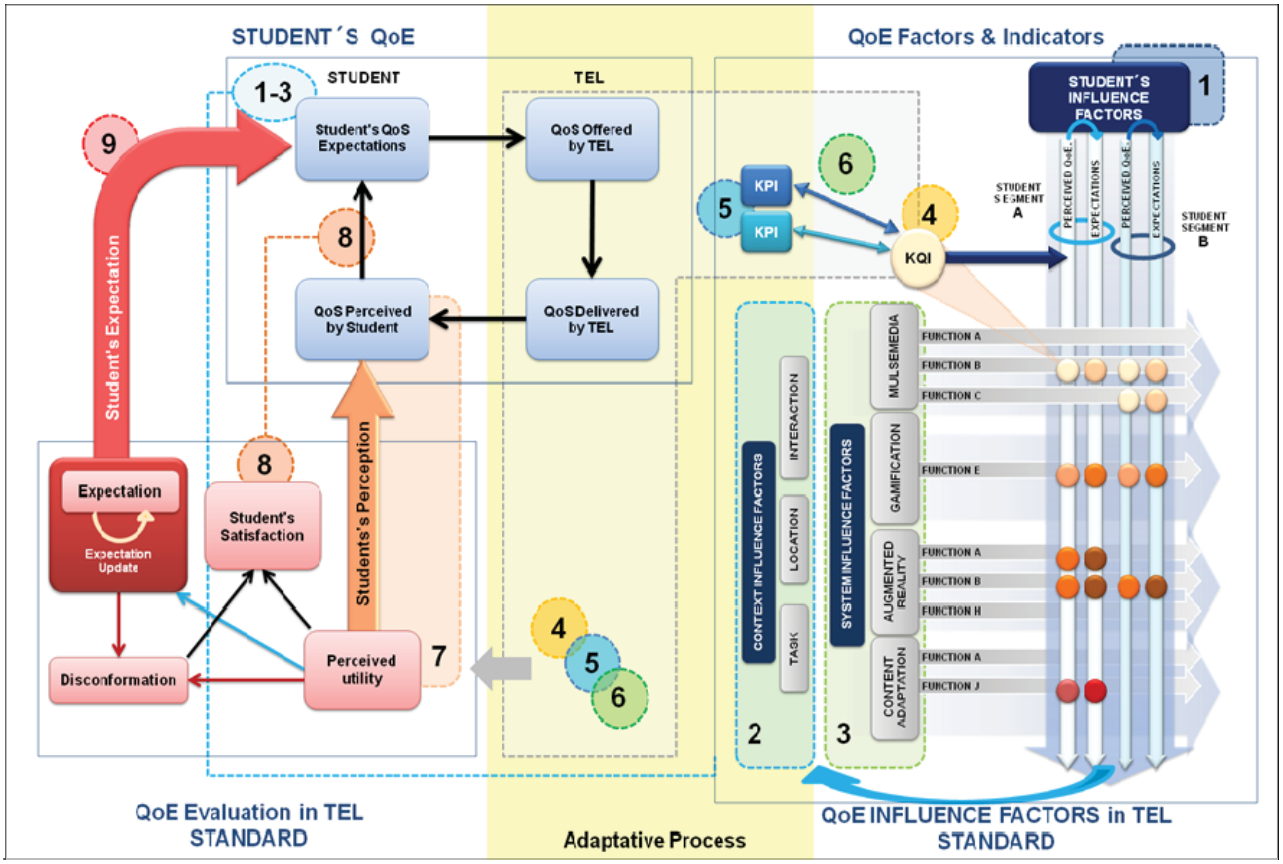


Figure 6. Proposals for the QoE evaluation and standardization process in TEL

5. PILOT STUDY

The Research Group NQaS (Networking Quality and Security) of the University of the Basque Country in Spain has a large experience in the definition and validation of QoS/QoE standards [46-55]. They have also participated as rapporteurs and editors in the ITU-T Study Group 11 being one of the first academia members to be actively involved in the standardization process developed in this SDO. In fact, recently, it has been approved one of the most innovative Recommendations in the area of Internet QoE/QoS measurements that was, indeed, edited by two members of this university [56].

For that reason, one of the first pilot studies that aims to bring advances in both NEWTON project in general and the research on the learner experience in TEL environments in particular, will be developed through the cooperation of this research group with DCU (one of the partners of the NEWTON project).The proposed study targets the students of the second year of the official Master in Telecommunication Engineering, attending the "*Performance on Telecommunications Networks*" course. This course is offered in the first semester of the year so it is intended that the pilot testing will be carried out sometime in November 2106, when some of the devices and testing platforms of the NEWTON project will be already ready for testing.

The experience and knowledge in the telematics area of the students that will participate in this validation will also assist to define the capabilities of NEWTON platform related to one of its target: attracting students to STEM higher education.

The general proposal is to provide certain parts of the "*Performance on Telecommunications Networks*" syllabus using both next generation TEL and traditional learning and analyze how the students learning experience is enhanced by TEL strategies.

The pilot will focus on two main innovative technologies that are brought by NEWTON project in the general context of learning: adaptive multimedia and mulsemedia content delivery.

The main goals of the study are to measure the following aspects that are highly important in the context of NEWTON project in particular and any TEL-related project in general:

- The influence of the employed innovative technologies on increasing learner QoE
- The influence of the employed innovative technologies on improving learning process
- The influence of the employed innovative technologies on increasing learning outcome.

In this purpose the proposed framework (Figure 6) will be used. In addition, in measuring the influence of TEL on improving the learning process we will also employ a questionnaire developed in collaboration with experts in Psychopedagogy from University of Bucharest, a partner of NEWTON project.

Moreover, in order to measure the influence of TEL on increasing the learning outcome, the knowledge of the students related to the material presented in the 2-hours course will be assessed via a test. The aim is to compare the students' performance in the questions related to the course aspects that were not presented employing TEL elements against the students' performance in the questions related to the course aspects that were presented employing TEL. A pre-test will be also given to the students to make sure that the results analysis is not influenced by pre-existent knowledge.

As mentioned before, the participants to the study will be master students attending their usual course that this time will be delivered employing the aforementioned technology advances.

The specific part of the course syllabus that will be carried out with the proposed TEL framework will focus on network performance metrics, quality of service, QoE and quality of perception concepts.

Parts of the course will be delivered using multimedia adaptive techniques. Short videos related to different network performance metrics will be delivered to the students. Note that NEWTON platform will not be available at the time of this pilot study, and therefore the delivery of these videos will not be done via the platform itself, but in a test-bed like manner (Figure 7).

A network emulator will be employed to simulate different network conditions. Some of the videos will be delivered in an adaptive manner, some in a non-adaptive manner in order for the students to be able to note the difference. All these aspects will be captured in the questionnaire.

Mulsemedia examples will be delivered as support for a better understanding of some of the taught concepts. In this particular case, mulsemedia will not be used to deliver the de-facto content of the course. Students will get to experience mulsemedia on short movies for a better understanding of the quality of perception metric. As depicted in Figure 8, different devices will be used to deliver mulsemedia content (e.g. air device – fan, haptic device – haptic vest, etc.).

Same pilot is planned to be carried also in Dublin, Ireland in Dublin City University (DCU), at the master level as well. Results obtained in both institutions, UPV/EHU and DCU, will be analyzed and compared.

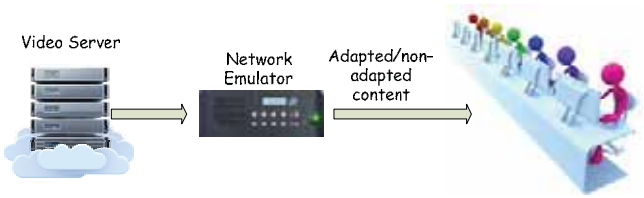


Figure 7. Delivery of course content using adaptive multimedia

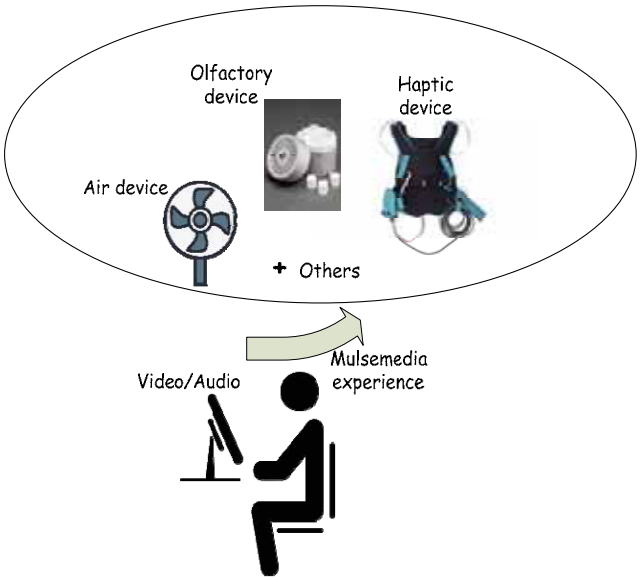


Figure 8. Mulsemedia experience during pilot study

6. CONCLUSIONS

This paper presents a novel initiative to create a network of new and existing teaching labs in the aim to develop a pan-European learning platform targeting to attract students to STEM studies and to support the learning process of people with special needs through the enhancement of the student learning process by means of the analysis of the learner QoE.

This platform will be developed using next generation TEL tools and strategies (mulsemedia, gamification, augmented reality, etc.) and the platform will be tested and validated in real life pilots across Europe, using a network of secondary schools, third level institutes and some universities. One of the first pilot's studies will be carried out in November 2016 and will target master students of the University of the Basque Country (UPV/EHU). This pilot is described in details in the paper.

A parallel initiative related to this teaching platform, and based on the achieved results, aims to define and develop new standards to unify criteria on the QoE evaluation on TEL platforms for comparable results. This paper presents some initial steps in the context of this research initiative that is in early development stages. As such, in this paper it is introduced a novel proposal for the QoE evaluation and standardization in TEL context. In future works that will include NEWTON pilots studies, this proposal will be validated on the basis of the results of these studies.

ACKNOWLEDGEMENTS

This work was supported by European Union's Horizon 2020 Research and Innovation programme under Grant Agreement no. 688503 for the NEWTON project. (http://www.newtonproject.eu) and the Spanish Ministerio de Economia y Competitividad (MINECO) under grant TEC2013-46766-R: "QoEverage - QoE-aware optimization mechanisms for next generation networks and services".

REFERENCES

[1] A. R. Sebastian Möller, *QoE. Advanced Concepts, Applications and Methods*. Switzerland: Springer International Publishing, 2014.

[2] C. H. Muntean and J. McManis, "End-User QoE Oriented Adaptive E-learning System," *Journal of Digital Information*, vol. 7, 2006.

[3] E. Ibarrola, E. Liberal, I. Taboada, and R. Ortega, "Web QoE Evaluation in Multi-agent Networks: Validation of ITU-T G.1030," presented at the Proceedings of the 2009 Fifth International Conference on Autonomic and Autonomous Systems Valencia, 2009, pp. 289-294..

[4] M. Adolph, "Standards for technology-enabled learning," ITU, Geneve, Switzerland 2012.

[5] Qualinet, "Qualinet white paper on definitions of QoE," Qualinet- European Network on QoE in Multimedia Systems and Services (COST Action IC 1003), Lausanne, Switzerland,2012.

[6] Z. Yuan, S. Chen, G. Ghinea, and G.-M. Muntean, "User QoE of Mulsemedia Applications," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 11, pp. 1-19, 2014.

[7] G. Ghinea, C. Timmerer, W. Lin, and S. R. Gulliver, "Mulsemedia: State of the Art, Perspectives, and Challenges," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 11, pp. 1-23, 2014.

[8] NEWTON. (2015). *Networked Labs for Training in Sciences and Technologies for Information and Communication*. H2020-EU.2.1.1.IL.-ICT Available: http://cordis.europa.eu/project/rcn/199183_en.html

[9] S. T. L. B. University, "To What Extent Do Academic Staff See An E-learning Framework As Being Effective In Supporting Technology Enhanced Learning (TEL) Discussions And Activities?," *Journal of Perspectives in Applied Academic Practice*, vol. Vol 4, pp. 2-11, 2016.

[10] S. Bayne, "What's the matter with 'technology-enhanced learning'?," *Learning, Media and Technology*, vol. 40, pp. 5-20, 2015/01/02 2015.

[11] T. de Jong, S. Sotiriou, and D. Gillet, "Innovations in STEM education: the Go-Lab federation of online labs," *Smart Learning Environments*, vol. 1, pp. 1-16, 2014.

[12] G.-L. Project. (2013). *The Go-Lab Project*. Available: http://www.go-lab-project.eu

[13] C. E. Wieman, W. K. Adams, and K. K. Perkins, "PhET: Simulations that enhance learning," *Science*, vol. 322, 2008.

[14] L. F. Z. Rivera and M. M. Larrondo-Petrie, "Models of remote laboratories and collaborative roles for learning environments," in *2016 13th International Conference on Remote Engineering and Virtual Instrumentation (REV)*, 2016, pp. 423-429.

[15] OECD, *Connected minds: Tecnology and Today´s learners*: OECD publishing 2012.

[16] C. H. Muntean, "Improving learner QoE by content adaptation based on network conditions," *Computers in Human Behavior*, vol. 24, pp. 1452-1472, 2008.

[17] N. Murray, B. Lee, Y. Qiao, and G. Miro-Muntean, "The influence of human factors on olfaction based mulsemedia QoE," in *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*, 2016, pp. 1-6.

[18] Z. Yuan, G. Ghinea, and G. M. Muntean, "Beyond Multimedia Adaptation: QoE-Aware Multi-Sensorial Media Delivery," *IEEE Transactions on Multimedia*, vol. 17, pp. 104-117, 2015.

[19] *G.1040: Network contribution to transaction time*, 2006.

[20] *ITU-T. F.742: Service description and requirements for distance learning services*, ed. Geneve, Switzerland, 2005.

[21] ITU-T. (2012). *ITU-T Focus Group Bridging the Gap: From Innovation to Standards*. Available: http://itu.int/en/ITU-T/focusgroups/innovation/

[22] ISO/IEC. (1999). *ISO/IEC JTC1 Subcommittee 36: Information technology for learning, education and training*. Available: http://www.iso.org/iso/iso_technical_committee%3Fcommid%3D45392

[23] IEEE. (2016). *The IEEE Learning Technology Standards Committee*. Available: https://ieee-sa.imeetcentral.com/ltsc/

[24] I. GLC. (1995). *IMS Global Learning Consortium (IMS GLC)*. Available: https://www.imsglobal.org/

[25] ISO, "ISO/IEC 19796-1: Information technology -- Learning, education and training -- Quality management, assurance and metrics -- Part 1: General approach," 2005.

[26] ISO/IEC, "ISO/IEC 23005: Information technology -- Media context and control (MPEG-V)," 2014.

[27] K. Yoon;, S.-K. Kim;, J. J. Han;, S. Han;, and M. Preda, *MPEG-V: Bridging the Virtual and Real World*: Todd Green, 2015.

[28] ISO/IEC, "ISO/IEC 15938: Information technology -- Multimedia content description interface (MPEG-7)," 2002.

[29] R. Huang, Y. Hu, and J. Yang, "Improving Learner Experience in the Technology Rich Classrooms," in *Ubiquitous Learning Environments and Technologies*, Kinshuk and R. Huang, Eds., ed Berlin, Heidelberg: Springer Berlin Heidelberg, 2015, pp. 243-258.

[30] Z. Yuan, T. Bi, G. M. Muntean, and G. Ghinea, "Perceived Synchronization of Mulsemedia Services," *IEEE Transactions on Multimedia*, vol. 17, pp. 957-966, 2015.

[31] Z. Yuan, G. Ghinea, and G. M. Muntean, "QoE study for multiple sensorial media delivery," in *2014 International Wireless Communications and Mobile Computing Conference (IWCMC)*, 2014, pp. 1142-1146.

[32] N. Murray, B. Lee, Y. Qiao, and G. Miro-Muntean, "The Impact of Scent Type on Olfaction-Enhanced Multimedia QoE," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. PP, pp. 1-13, 2016.

[33] *ITU-T. P.911: Subjective audiovisual quality assessment methods for multimedia applications*, 1998.

[34] *ITU-T. P.10/G.100 (2006) Amendment 1 (01/07): New Appendix I - Definition of QoE (QoE)* 2007.

[35] *ITU-T. G.1080: QoE requirements for IPTV services*, 2008.

[36] *ITU-T. P.910: Subjective video quality assessment methods for multimedia applications*, 2008.

[37] *ITU-T. G.1011: Reference guide to QoE assessment methodologies*, 2011.

[38] *ITU-T. G.1031: QoE factors in web-browsing* 2014.

[39] *ITU-T. P.1501: Subjective testing methodology for web browsing*, 2014.

[40] *ITU-T. P.912: Subjective video quality assessment methods for recognition tasks*, 2016.

[41] *ITU-T. P.913: Methods for the subjective assessment of video quality, audio quality and audiovisual quality of Internet video and distribution quality television in any environment*, 2016.

[42] *ITU-T. P.915: Subjective assessment methods for 3D video quality*, 2016.

[43] D. A. Cook; and R. H. Ellaway, "Evaluating technology-enhanced learning: A comprehensive framework," *Med Teach*, 2015.

[44] ITU-T, G.1000: *Communications quality of service: A framework and definitions*, 2001.

[45] E. W. Anderson and M. W. Sullivan, "The Antecedents and Consequences of Customer Satisfaction for Firms " *Marketing Science*, vol. 12, pp. 125-143, Spring, 1993 1993.

[46] UPV/EHU, "TD 676 (GEN/11): Invited presentation from UPV/EHU to advance in the draft Recommendation ITU-T Q.Int_speed_test," in *Q15/11*, vol. SG-11, Fidel Liberal and Eva Ibarrola, Eds., ed. Geneve, Switzerland: ITU-T, 2015, p. 1.

[47] UPV/EHU, "COM 11 – C 272 Rev.1 – E: Proposal to advance the Draft recommendation Q.Int_speed_test "Unified methodology of Internet speed quality measurement usable by end-users on the fixed and mobile networks"," in *Q.15/11*, vol. SG-11, Fidel Liberal, Eduardo Saiz, Eneko Atxuregi, and Eva Ibarrola, Eds., ed. Geneve, Switzerland: ITU-T, 2015, pp. 1-22.

[48] UPV/EHU, "COM 11 – C 150 – E: Proposal to advance the Draft Recommendation Q.QMS "The Framework of the QoS/QoE Monitoring System"," in *Q.10/11*, vol. SG-11, Eva Ibarrola and Armando Ferro, Eds., ed. Geneve, Switzerland: ITU-T, 2014, pp. 1-17.

[49] UPV/EHU, "COM 12 – C 214 – E: Proposal for updating ITU-T Rec. G.1000," in *Q.2/12*, vol. SG-12, Eva Ibarrola, Ed., ed. Geneve, Switzerland: ITU-T, 2014, pp. 1-15.

[50] UPV/EHU, "COM 11 – C 165 Rev.1– E: Proposal to advance the Draft recommendation Q.Int_speed_test "Unified methodology of Internet speed quality measurement usable by end-users on the fixed and mobile networks"," in *Q.15/11*, vol. SG-11, Eva Ibarrola and Fidel Liberal, Eds., ed. Geneve, Switzerland: ITU-T, 2014, pp. 1-13.

[51] UPV/EHU, "COM 11 – C 24 – E: Proposals for the collaborative QoE/QoS testing platform to advance in "Q.QMS" work item," in *Q.10/11*, vol. SG-11, Eva Ibarrola, Armando Ferro, and Jose Luis Jodra, Eds., ed. Geneve, Switzerland: ITU-T, 2013, pp. 1-5.

[52] UPV/EHU, "COM 11 – C 99 – E: Proposals to advance in "Q.QMS" work item," in *Q.10/11, Q.11/11*, vol. SG-11, Eva Ibarrola and Armando Ferro, Eds., ed. Geneve, Switzerland: ITU-T, 2013, pp. 1-3.

[53] UPV/EHU, "COM 12 – C 30 – E: Proposal on enhancement of G.1000 framework," in *Q.2/12*, vol. SG-12, Eva Ibarrola, Ed., ed. Geneve, Switzerland: ITU-T, 2013, pp. 1-13.

[54] UPV/EHU, "COM 12 – C 124 – E: Test results and new proposals for enhancement of Draft P.STMWeb," in *Q.7/12*, vol. SG-12, Eva Ibarrola and Eduardo Saiz, Eds., ed. Geneve, Switzerland: ITU-T, 2013, pp. 1-11.

[55] UPV/EHU, "COM 11 – C 283 Rev.1 – E: Collaborative testing platform for extensive QoS experimental scenarios," in *Q.11/11*, vol. SG-11, Eva Ibarrola, Armando Ferro, and Luis Zabala, Eds., ed. Geneve, Switzerland: ITU-T, 2012, pp. 1-7.

[56] *ITU-T. Q.3960: Framework for Internet related performance measurements*, 2016.

SESSION 3

SPECTRAL EFFICIENCY IN WIRELESS NETWORKS

S3.1 Space division multiplexing technology: next generation optical communication strategy.*

S3.2 Resource allocation for device-to-device communications in multi-cell LTE-advanced wireless networks with C-RAN architecture.*

S3.3 PAPR reduction in SC-FDMA via a novel combined pulse-shaping scheme.

S3.4 Accelerating the introduction of spectrum sharing using market-based mechanisms.

SPACE DIVISION MULTIPLEXING TECHNOLOGY: NEXT GENERATION OPTICAL COMMUNICATION STRATEGY

Kazuhide NAKAJIMA, Takashi MATSUI, Kotaro SAITO, Taiji SAKAMOTO, and Noriyuki ARAKI

Access Network Service Systems Laboratories, NTT Corporation
1-7-1, Hanabatake, Tsukuba, Ibaraki, 305-0805, Japan

ABSTRACT

Space division multiplexing (SDM) is expected to be a key technology both for dealing with the future capacity crunch facing traditional single-mode fibre (SMF) and for realizing a sustainable optical network that can accommodate the various data streams originating from, for example, future 5G communication, the Internet of things (IoT), and machine to machine (M2M) networks. This paper describes the potential of SDM as regards optical fibre and cable technology. We focus on the potential of multi-core fibre (MCF), and investigate the reality of MCF based SDM optical wiring as the first example of an SDM application taking the latest research and development into consideration. Finally, we show that MCF based SDM optical fibre cable will be a promising technology for next generation optical networks, and the key technology behind MCF based SDM optical wiring is ready for discussion as the near future standard.

Keywords— Space division multiplexing (SDM); Optical fibre and cable; Multi-core fibre (MCF); Optical connector

1. INTRODUCTION

Data capacity demand has been increasing rapidly against the background of the worldwide spread of fibre to the home (FTTH) and mobile networks. The imminent introduction of 5G, the Internet of Things (IoT), and machine to machine (M2M) communications not only requires further capacity but also demands a flexible and sustainable interface with a high-speed optical fibre link.

Over the past ten years, space division multiplexing (SDM) technology has been investigated intensively mainly with the aim of increasing the maximum transmission capacity in an optical fibre. This is because the capacity demand in the 2020s will potentially exceed the limitation imposed by conventional single-mode fibre (SMF) [1]. SDM could be the key to realizing both ultra-large capacity transmission and to solving various problems with current and future optical networks.

In this paper, we investigate the potential of SDM technology as a new strategy for realizing next generation optical networks. In section 2, we will briefly review the relationship between SDM technology and various objectives that must be met by current and future optical networks. We then investigate SDM technology in relation to optical fibre cable. Here, we clarify the potential of a multi-core fibre (MCF) as an SDM medium. After that we discuss key stages in the deployment of MCF technology. We also introduce some key technologies described in the latest reports. Finally, we show that MCF based SDM optical wiring is ready for discussion as the next generation optical communication standard for the near future

2. WHY SDM?

The worldwide spread of FTTH and mobile communication has increased data capacity at a rate of a few tens of percent per year. The upcoming 5G service requires a wider bandwidth, and the growing communication network has created various data streams including IoT, M2M, sensor networks, and connected cars. Therefore, it is true that the continuing increase in data transmission will cause a capacity crunch in existing SMF since its maximum transmission capacity is limited to around 100 Tbit/s because of the low loss and optically amplified transmission window [1]. In general terms, the approaching capacity crunch requires either additional optical infrastructure or an innovative transmission strategy. SDM can be considered a fifth physical dimension that can be used with the traditional dimensions of time, wavelength, frequency and polarization multiplexing technology [2]. Thus, it is easy to imagine that SDM can be used as a candidate technology for overcoming the future capacity crunch.

Figure 1. Schematic image of growing communication network. FTTH, mobile, 5G, IoT, M2M, sensor network, connected car, and so on increase the data capacity continuously, and require additional interfaces with a high-speed optical link.

In addition to the capacity crunch, SDM can be used to realize such functions as the effective operation and/or management of the optical infrastructure, and an optical transmission system with a lower power consumption. The worldwide expansion of the network scale and its divergence has led to a rapid increase in the number of optical fibres particularly in central offices and data centres. Thus, an urgent problem for both network operators and service providers is to obtain an effective tool for solving the optical fibre conjunction in their facilities. The SDM of the optical physical infrastructure potentially provides a direct solution to this problem in terms of space saving. On the other hand, the latest transmission system enables us to realize a rate of 40 Gbit/s or more by employing digital signal processing technology and a powerful error correction scheme. And this technological progress generally increases the power consumption per bit particularly in metro and core optical transmission systems. Here, SDM technology also enables us to reduce power consumption by sharing some of the active components used in the system. For example, Ref. [3] showed that power efficient ultra-high capacity long-haul transmission system can be designed by sharing pump light for optical amplifiers with multiple spatial channels. Therefore, it can be said that SDM technology will be a key transmission strategy for future optical communication, and it should be used effectively for updating existing optical communication systems in a sustainable manner. Then, we discuss SDM in relation to optical fibre cable technology

3. SDM IN OPTICAL FIBRE CABLE

3.1. Additional Spatial Dimension

Figure 2 shows historical progress made on communication cable density. The black, blue and red circles show 400-pair copper, slot-type 400 fibre ribbon, and 400 rollable fibre ribbon [4] cables, respectively. The solid green line shows the numerical limit when we assume a hexagonally packed 250 μm fibre bundle with a 2 mm thick cable sheath. Figure 2 reveals that the communication cable density has been increased greatly by employing optical fibre and a ribbon structure. A cable density of 4.1 mm^{-2} is obtained with the latest rollable ribbon as shown in the inset photograph, and the value is approaching the geometrical limit. Thus, Fig. 2 shows that we need another dimension if we want to realize a higher density beyond the current geometrical limit. Here, we should note the possibility of a higher count fibre bundle rather than a higher cable density. It is true that we can overcome the future capacity crunch simply by introducing additional SMFs. However, this approach intrinsically needs additional space, and it results in an additional cost for constructing/renewing the physical infrastructure.

Historically, optical fibre is tailored using the two spatial dimensions of mode and core as shown in Fig. 3. Thus, there is the potential to overcome the current geometrical limit of traditional optical fibre cable by employing mode and/or core multiplexing adequately. The latest research has proved that we can obtain more than 100 spatial channels by employing mode and core multiplexing

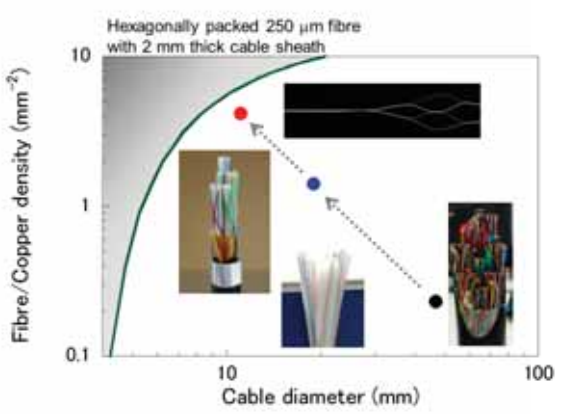


Figure 2. Historical progress on communication cable density. Black, blue and red corresponds to 400-pair copper, slot-type 400 fibre ribbon and 400 rollable fibre ribbon [4] cables, respectively.

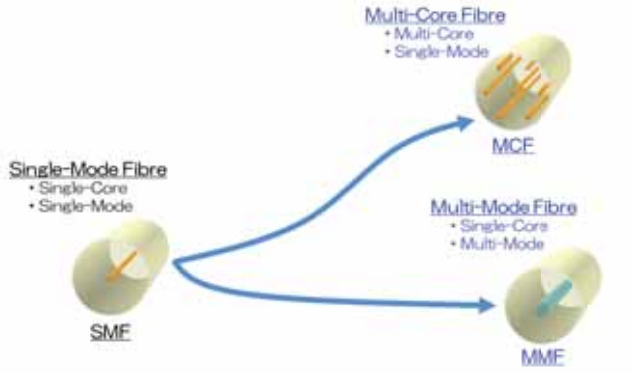


Figure 3. Schematic image of two representative spatial dimensions of mode and core in an optical fibre.

simultaneously [5]. However, mode division multiplexing intrinsically needs a complex transmission strategy because of the variation in the mode coupling and/or transmission characteristics between modes in an optical fibre. At the same time, MCF has been continuously investigated, and we can find example pioneering studies in Refs. [6] and [7]. Moreover, MCF with a single-mode core can easily employ the latest single-mode technology. In this paper, we focus on MCF technology and investigate the potential of MCF as an SDM transmission medium.

3.2. Potential of MCF

MCF has three key geometrical parameters; (I) cladding diameter D, (II) cladding thickness t, and (III) core pitch Λ as shown schematically in Fig. 4. The cladding diameter D is closely related to the mechanical reliability of the optical fibre. Generally speaking, a larger D degrades the failure probability when there is a small bend [8], although a larger D value is useful for increasing the core number in MCF.

Figure 5 shows an example of the calculated relative failure probability as a function of D. Red and blue solid lines correspond to proof levels of 1 and 2%, respectively.

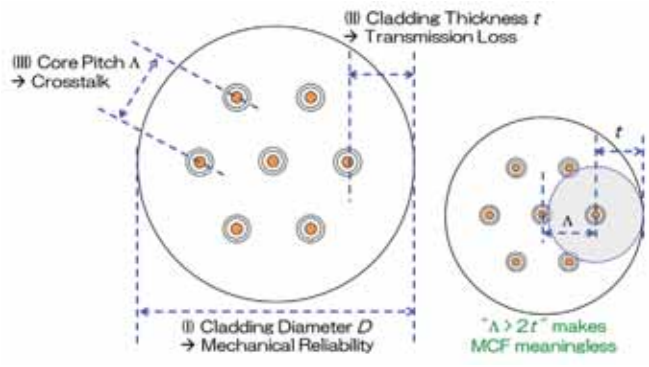


Figure 4. Three key geometrical parameters in MCF. Right schematic shows an example relationship between core pitch Λ and cladding thickness *t*.

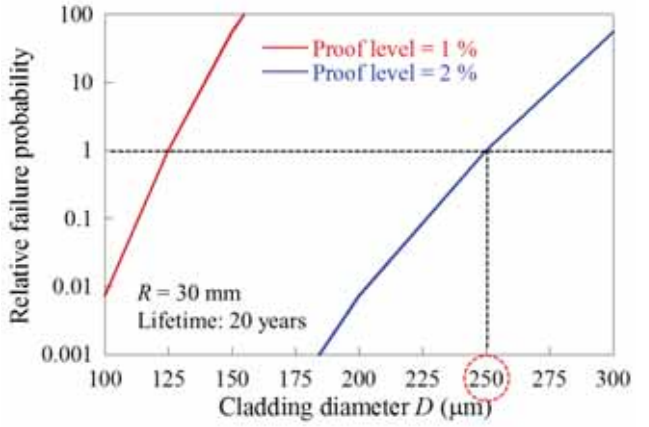


Figure 5. Calculated failure probability as a function of cladding diameter *D* when we assumed a 30 mm bending radius *R* and 20 years lifetime.

These values are commonly used in the current fabrication process. The bending radius *R* and lifetime are assumed to be 30 mm and 20 years, respectively. Here, the 30 mm bending radius corresponds to the minimum allowable value, which is standardized in Recommendation G.652 (i.e. standard SMF). The vertical axis is normalized with the value when the proof levels and *D* are 1% and 125 μm, respectively. A 125 μm cladding diameter corresponds to the typical value for a conventional SMF. In this calculation, traditional power law theory [9] was used and typical stress corrosion parameter obtained with conventional SMF was assumed. Figure 5 reveals that the failure probability increases greatly with a larger *D*. Figure 5 also confirms that we can maintain a relative failure probability of one even at *D* = 250 μm by using a 2% proof level. These results show that we can increase the *D* value of MCF up to 250 μm in order to allocate multiple cores while maintaining a feasible mechanical reliability.

As regards the *t* value, a smaller *t* degrades the transmission loss. Roughly speaking, the minimum *t* value is almost equivalent to the allowable cladding diameter in a thin cladding fibre as shown by the blue dotted line in the schematic on the right in Fig. 4. Although the minimum *t* is also related to an effective core area, we can reduce *t* from the conventional value of 62.5 μm to 35-40 μm [10] while maintaining a feasible transmission loss.

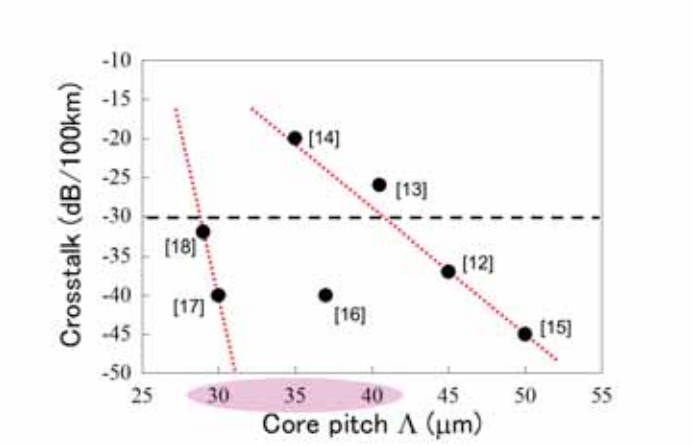


Figure 6. Example relationship between crosstalk and core pitch Λ found in some fabricated MCFs [12]-[18].

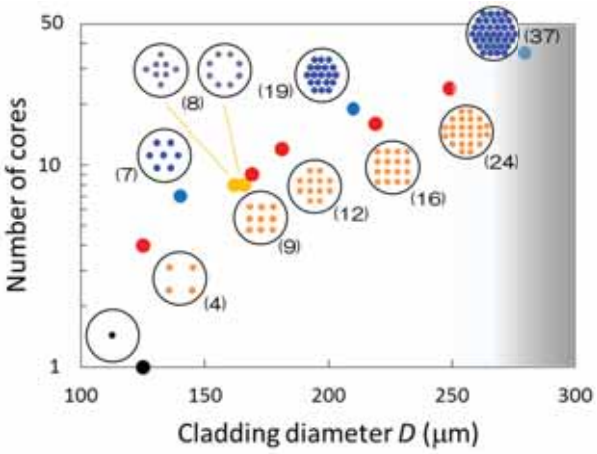


Figure 7. Numerical relationship between number of cores and cladding diameter *D* when we set both the Λ and *t* values at 35 μm. Red and blue circles represent the square lattice and hexagonal core arrangements and the values in brackets show the number of cores.

The minimum core pitch Λ is determined by the allowable crosstalk level between neighbouring cores. The relationship between crosstalk and Λ can be estimated numerically by considering various core arrangements [11]. In the past ten years, various MCFs have been fabricated taking the above numerical guideline into consideration. Figure 6 shows an example relationship between crosstalk and Λ found in some fabricated MCFs [12]-[18]. Figure 6 confirms that the crosstalk degrades as Λ decreases, and it seems there are two boundaries as shown by the dashed red lines. The MCFs around the right boundary have homogeneous cores. On the other hand, the two MCFs around the left boundary have a heterogeneous core arrangement to improve the crosstalk characteristic. Thus, these results reveal that we can manage the crosstalk in

MCF in terms of Λ, and we can use a Λ value of 30-40 μm if we set -30 dB/100km as the allowable crosstalk level. Here, it should be noted that an MCF becomes meaningless in terms of an SDM medium if the core pitch Λ is larger than $2\times t$. This is because $\Lambda = 2t$ is equivalent to bundled thin cladding fibre. However, the previous discussion makes it clear that we can obtain sufficient SDM efficiency with a well-designed MCF since the minimum Λ is expected to be satisfactorily smaller than $2\times t$ as shown by the schematic on the right in Fig. 4.

Figure 7 shows the numerical relationship between the number of cores and the cladding diameter *D*. On the basis of the above discussion, we assumed that the Λ and *t* values were both 35 μm. The red and blue circles show the results we obtained when we used a square lattice and a hexagonal core arrangement, respectively, as shown by the inset schematic. The value in brackets shows the number of cores. Two orange symbols also show the results when we assumed an eight-core fibre using the modified square lattice or circular core arrangements. Figure 7 confirms that the cores in the MCF can be increased by using a larger *D* value. However, we also found that a 37-core MCF with a hexagonal core arrangement may require additional care as regards mechanical reliability because the *D* value exceeds 250 μm. As a result, we confirmed that MCF can potentially provide a few tens of cores for SDM transmission if we optimise three key geometrical parameters simultaneously.

4. STANDARDIZATOIN OF MCF TECHNOLOGY

4.1. Example Milestone in MCF Technology

In this section, we investigate an example milestone in MCF technology taking two aspects into account, namely widespread use and consistency to the current fibre fabrication process. Figure 8 shows schematically the relationship between the potential application area of MCF and its penetration phases. Generally, it is difficult to consider the metro/core network as the first deployment area of MCF technology since it requires all the network elements other than the MCF cable. By contrast, a central office and/or data centre can be more easily considered as the first application area of MCF technology since these areas are managed and/or optimised individually by the well-skilled operators and it is easier to replace the existing network element taking the latest trend into consideration. We will then spread MCF technology into a point-to-point (P2P) and/or parallel transmission system in its second deployment period. Here, it should be noted that a long-haul transmission system may have more applicability compared with the metro/core transmission system since the long-haul transmission systems, such as optical submarine transmission systems, have been individually optimised using the latest technology.

In terms of a consistent fabrication process, we can also consider a three-tier approach shown in Fig. 9. In fact, we need more time to realize a reliable and cost-effective mass-production process for MCF with a larger cladding diameter. This is because any increase in the cladding diameter

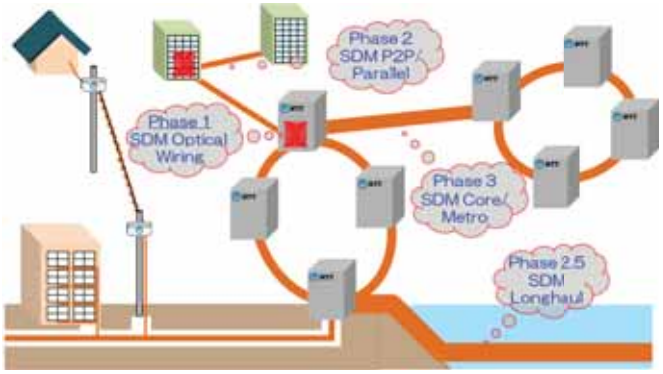


Figure 8. Schematic relationship between considerable application area of MCF and its spreading time.

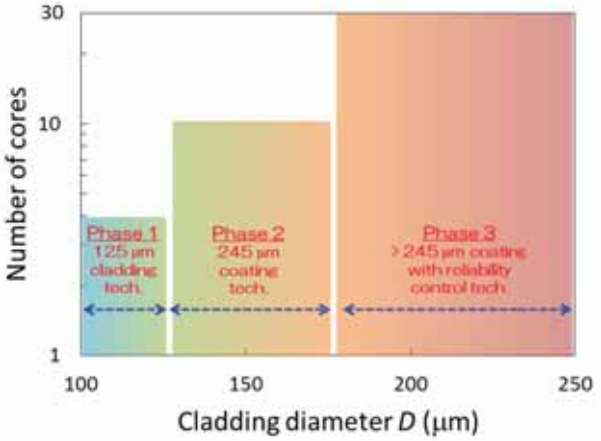


Figure 9. Example relationship between MCF design and fabrication technology.

directly limits the maximum fibre length that can be drawn from the same size preform. Moreover, we must undertake additional study to realize a feasible reliability control. By contrast, we can easily use the current fibre fabrication process if we consider a 125 μm cladding diameter as the first MCF application. Moreover, we can easily move to the second phase by considering a coating diameter of 245 μm. This is because a 245 μm coating diameter is consistent with the current dimension, and we also have some knowledge of thin coating technology such as that used for 200 μm coated optical fibre. This knowledge can be applied to an MCF with a larger cladding diameter and a smaller coating thickness. We can then easily expand the application area and available core number step by step as shown in Fig. 9. As a result, we can imagine reaching a feasible milestone in MCF technology by considering an adequate application area and fabrication level.

4.2. Key Technology for MCF Wiring

When we focus on the first application area of SDM optical wiring, a 125 μm cladding MCF and splicing technology are essential. In this section, we review recent progress on these

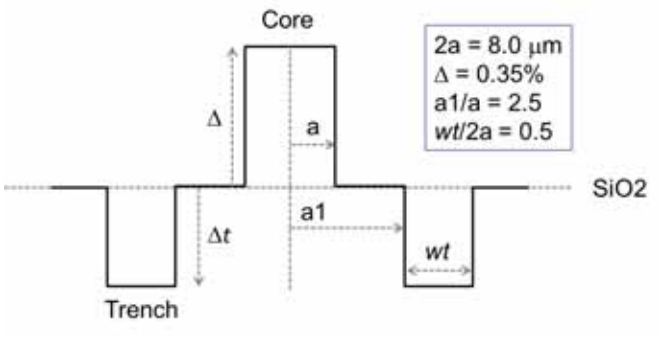


Figure 10. Refractive index profile considered in a 125 μm cladding MCF [19].

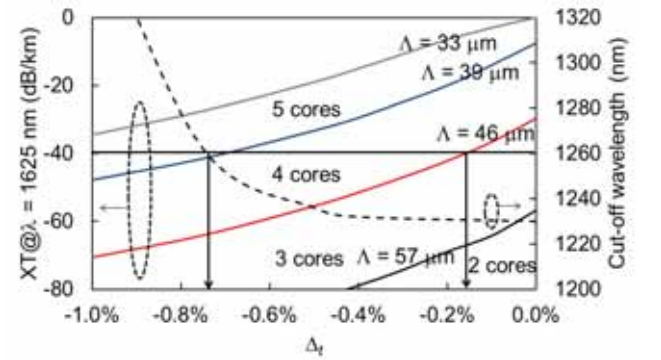


Figure 11. Calculated Δ*t* dependence of crosstalk (XT) and cutoff wavelength [19]. Black, red, blue and gray solid lines show the XT characteristics when the Λ values are set at 57, 46, 39 and 33 μm, respectively. The dashed line shows the cutoff wavelength characteristic.

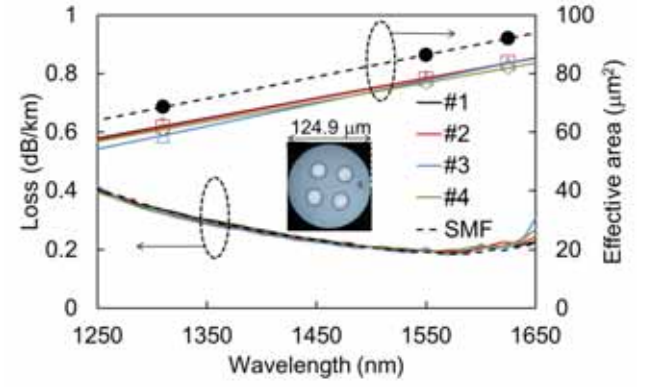


Figure 12. Measured loss and effective area characteristics of the fabricated four-core MCF with a 124.9 μm cladding diameter.

technologies. An optimum design for a 125 μm cladding MCF is described in Ref. [19]. A trench assisted refractive index profile, as shown in Fig. 10, was considered in this study. The centre core and inner cladding radius are assumed to be a and a1, respectively. The relative index difference of the centre core and trench are defined as Δ and Δ*t*, respectively. *wt* represents the trench width. In this study, we assumed optical consistency with conventional G.652 fibre in terms of cut-off wavelength, mode-field diameter (MFD), zero-dispersion wavelength and bending loss characteristics. These requirements result in an example refractive index of 2a = 8.0 μm, Δ = 0.35%, a1/a = 2.5 and *wt*/2a = 0.5. Here, a and Δ values are mainly determined by the MFD, zero-dispersion wavelength, and bending loss requirements. Roughly speaking, a1/a, *wt*/2a and Δ*t* values are relating with the cut-off wavelength and crosstalk (XT) requirements.

Figure 11 shows the calculated Δ*t* dependence of the XT and cut-off wavelength. The solid black, red, blue and grey lines show the XT characteristics when the Λ values were set at 57, 46, 39 and 33 μm, respectively. These Λ values correspond to the maximum limit for supporting 2, 3, 4 and 5 core arrangements with a 125 μm cladding diameter. The dashed line shows the cut-off wavelength characteristic. It is confirmed from Fig. 11 that we can arrange four cores by setting Δ*t* at -0.5% to -0.74% while maintaining optical compatibility with G.652 fibre as expected in Fig. 7.

Figure 12 shows the wavelength dependence of the loss (left axis) and effective area (right axis) measured when using a fabricated four-core MCF with a 124.9 μm cladding diameter. The four solid lines correspond to the individual core in the MCF, and the dashed lines show example characteristics of conventional G.652 fibre. Figure 12 confirms that the fabricated MCF successfully achieved full compliance with conventional G.652 fibre, and Ref. [19] also confirmed that the fabricated MCF can be used for a 100 Gbit/s parallel transmission and a 40 Gbit/s DWDM transmission using the O and C+L bands. These results reveal that a 125 μm cladding MCF fully compliant with existing G.652 fibre is ready to be discussed in terms of real applications.

An MCF connector has also been studied intensively, and MCF splicing intrinsically requires rotational angle alignment in addition to conventional lateral offset alignment. Nagase realized an MU type MCF connector as shown in Fig. 13 (a) [20]. The fabricated MU connector successfully realized a low average loss of 0.13 dB by using an Oldham's coupling for rotational angle alignment. Reference [21] also describes an SC type MCF connector as shown in Fig. 13 (b). In this study, the angle alignment has been achieved by introducing a V-groove into a conventional ferrule. This study also proposed a rotatable mechanism by using a ferrule with four V-grooves as shown in Fig. 13 (c), and realized a pluggable add/drop module shown as Fig. 13 (d) combined with MCF and planar lightwave circuit technologies. These considerations are beneficial for opening up a new application area of SDM technology.

Fusion splicing for MCF has also been investigated. For example, Saito proposed aligning the rotational angle of the MCF simply by using a side-view image [22]. He revealed the validity of the proposed technique using four and eight core MCFs as shown in Fig. 14 (a). Here, the cladding diameters of four and eight core MCFs were 125 and 175 μm, respectively. In this technique, the average brightness shown in Figs. 14 (b) or (c), respectively obtained with 4-

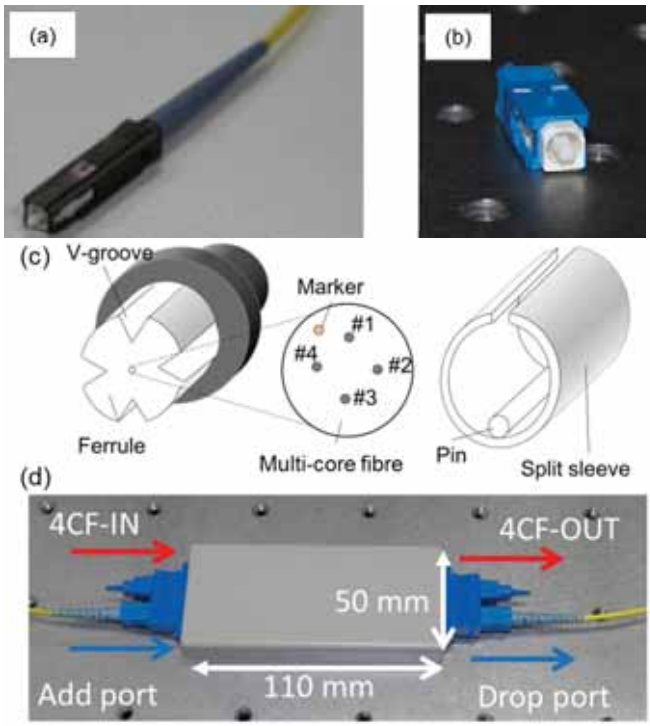


Figure 13. Latest research on MCF connectors [20], [21]. (a) and (b) are photographs of MU and SC type MCF connectors, respectively. (c) is a schematic image of a rotatable ferrule. (d) is a photograph of a pluggable add/drop module using the MCF and a planar lightwave circuit.

core and 8-core MCFs, is analysed to derive the correlation coefficient *ZNCC*. Then, rotational angle alignment is performed to maximise the correlation coefficient as shown in Fig. 14 (d). Here, it should be noted that the proposed algorithm can be used not only for aligning the multiple cores but also for aligning the marker used in the MCF. These results show that the side-view alignment technique can potentially be used for realizing an automatic fusion splice for MCF. Thus, it can be said that the key technology for MCF splicing has already been established, and this is important in terms discussing the reality of random connection using the multi-vendor solution.

As a result, SDM optical wiring is becoming a real technology, and it is a good time to start discussing the standardization of these new physical layer technologies in order to open up the next generation optical communication strategy and new future markets.

5. CONCLUSION

This paper investigated the potential of SDM technology as a new strategy for next generation optical networks. We pointed out that the SDM concept can potentially be used not only for overcoming the future capacity crunch in conventional G.652 fibre but also for providing a key solutions to such issues as achieving effective infrastructure operation/management and lower power consumption. We

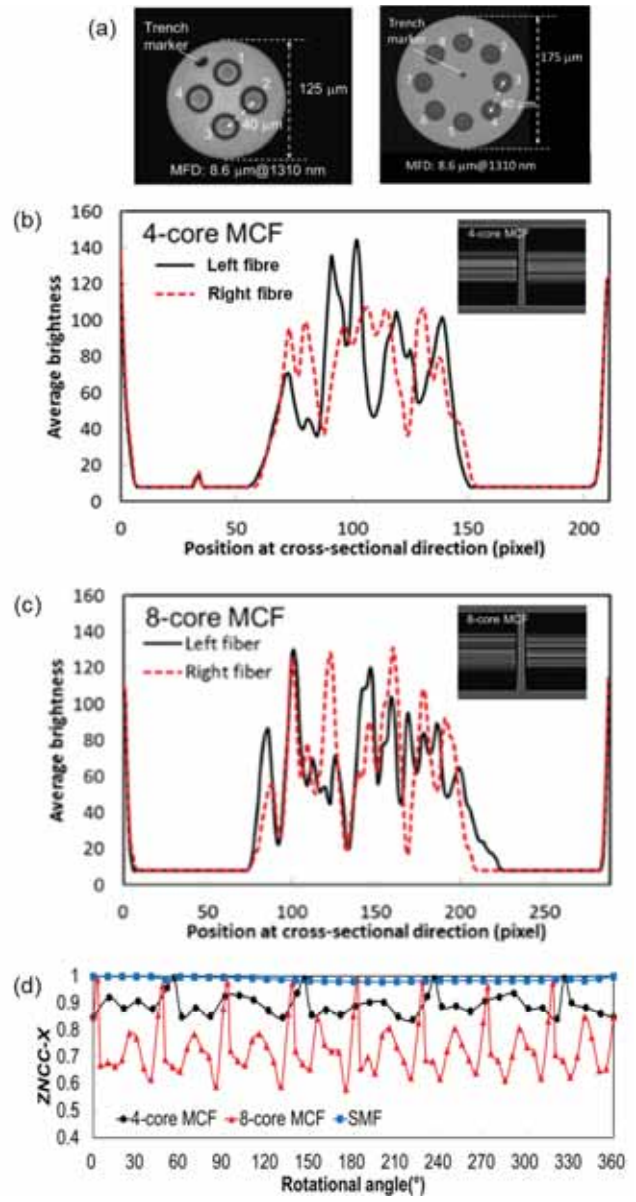


Fig. 14 Side-view based rotational angle alignment technique [22]. (a) shows a photograph of sample MCFs. (b) shows example average brightness characteristics as a function of the cross sectional direction obtained with 4-core MCF. (c) shows example average brightness characteristics as a function of the cross sectional direction obtained with 8-core MCF. (d) shows an example relationship between the correlation coefficient *ZNCC* and the rotational angle.

also proposed a tier approach that takes the application area and mass-producibility into consideration and which is beneficial for realizing the smooth and effective spread of MCF technology. Finally, we showed that 125 μm cladding MCF technology including a connector and fusion splicing technique is now ready thus allowing the discussion of its use for real applications. We believe that it is a good time to start a discussion on the standardization of SDM optical wiring so that we can open up the next generation optical communication strategy and new future markets.

REFERENCES

[1] T. Morioka, “New generation optical infrastructure technologies: “EXAT Initiative” towards 2020 and beyond,” in Proc. OECC’09, FT4, 2009.

[2] P. J. Winzer, “Optical networking beyond WDM,” IEEE Photon. J., vol. 4, no. 2, pp. 647-651, 2012.

[3] A. Turukhin et al., “105.1 Tb/s power-efficient transmission over 14,350 km using a 12-core fiber,” in Proc. OFC’16, Th4C.1, 2016.

[4] K. Hogari et al., “Novel optical fiber cables with ultrahigh density,” J. Lightwave Technol., vol. 26, no. 17, pp. 3104-3109, 2008.

[5] T. Sakamoto et al., “Low-loss and low-DMD few-mode multi-core fiber with highest core multiplicity factor,” in Proc. OFC'16, Th5A.2, 2016.

[6] S. Inao et al., “High density multicore-fiber cable,” in Proc. IWCS’79, p. 370-384, 1979.

[7] G. Le. Noane et al., “Ultra high density cables using a new concept of bunched multicore monomode fibers: A key for the future FTTH networks,” in Proc. IWCS’94, p. 203-210, 1994.

[8] M. Tachikura et al., “Improved theoretical estimation of mechanical reliability of optical fibers,” in Proc. SPIE, vol. 5623, pp. 622-629, 2005.

[9] IEC TR 62048, “Optical fibre – Reliability – Power law theory,” Ed. 3, Oct. 2014.

[10] Y. Goto et al., “Influence of cladding thickness on transmission loss and its relationship with multicore fiber structure,” J. Lightwave Technol., vol. 33, no. 23, pp. 4942-4949, 2015.

[11] M. Koshiba et al., “Multi-core fiber design and analysis: coupled-mode theory and coupled-power theory,” Opt. Express, vol. 19, no. 26, pp. B102-B111, 2011.

[12] T. Hayashi et al., “Low-crosstalk and low-loss multi-core fiber utilizing fiber bend,” in Proc. OFC'11, OWJ3, 2011.

[13] S. Matsuo et al., “Large-effective-area ten-core fiber with cladding diameter of about 200mm,” Opt. Lett., vol. 36, no. 23, pp. 4626-4628, 2011.

[14] J. Sakaguchi et al., “19-core fiber transmission of 19x100x172-Gb/s SDM-WDM-PDM-QPSK signals at 305Tb/s,” in Proc. OFC'12, PDP5C.1, 2012.

[15] M.J. Li et al., “Multicore fiber for optical interconnect applications,” in Proc. OECC'12, 5E4-2, 2012.

[16] S. Matsuo et al., “12-core fiber with one ring structure for extremely large capacity transmission,” Opt. Express, vol. 20, no. 27, pp. 28398-28408, 2012.

[17] Y. Amma et al., “High-density multicore fiber with heterogeneous core arrangement,” in Proc. OFC'15, Th4C.4, 2015.

[18] T. Mizuno et al., “32-core dense SDM unidirectional transmission of PDM-16QAM signals over 1600 km using crosstalk-managed single-mode heterogeneous multicore transmission line,” in Proc. OFC'16, Th5C.3, 2016.

[19] T. Matsui et al., “Design of 125 mm cladding multi-core fiber with full-band compatibility to conventional single-mode fiber,” in Proc. ECOC’15, We.1.4.5, 2015.

[20] R. Nagase et al., “MU-type multicore fiber connector,” IEICE Electron., vol. E96-C, no. 9, pp. 1173-1177, 2013.

[21] K. Saito et al., “Multi-core fiber based pluggable add/drop link using rotational connector,” in Proc. OFC’15, M2B.2, 2015.

[22] K. Saito et al., “Side-view based angle alignment technique for multi-core fiber,” in Proc. OFC’16, M3F.3, 2016.

RESOURCE ALLOCATION FOR DEVICE-TO-DEVICE COMMUNICATIONS IN MULTI-CELL LTE-ADVANCED WIRELESS NETWORKS WITH C-RAN ARCHITECTURE

Sajjad Mehri Alamouti and Ahmad R. Sharafat

Faculty of Electrical and Computer Engineering, Tarbiat Modares University, Tehran, Iran

ABSTRACT

Device-to-device (D2D) communications underlaying LTE-Advanced wireless networks reuse cellular frequency spectrum to establish direct links between users without traversing base stations or the cellular network. In this paradigm, there is a need to optimally allocate resources with a view to maximizing the utility, e.g., the total throughput, and mitigating the interference caused by sharing the same spectrum between cellular users (CUs) and D2D pairs. This paper proposes a scheme for optimally allocating transmit power levels and channels to maximize the total number of active D2D pairs and reused channels while minimizing the aggregate transmit power pertaining to CUs and D2D pairs. We consider a multi-cell scenario in which the transmitter and the receiver of each D2D pair can be in the same cell or in two different cells, and each user can simultaneously transmit over multiple reused channels. The optimization is done via a centralized baseband processing in the cloud radio access network (C-RAN) architecture. Simulations show that via our proposed scheme, more users (both cellular users and D2D pairs) can simultaneously communicate and the total system throughput is also significantly increased.

Keywords— D2D, spectrum sharing, LTE-A, C-RAN, multi-cell.

1. INTRODUCTION

Global mobile data traffic will increase nearly eightfold and grow at a compound annual growth rate (CAGR) of 53 percent from 2015 to 2020, reaching 30.6 exabytes per month by 2020. Moreover, there will be 11.6 billion mobile-connected devices (including machine-to-machine modules) by 2020, exceeding the world's projected population at that time (7.8 billion) [1]. In order to cope with these tremendous growth, new technologies have been introduced to use the limited resources (frequency spectrum, network infrastructure, mobile device battery, etc.) more efficiently. Device to device (D2D) communications [2] and cloud radio access network (C-RAN) architecture [3] are among such technologies.

In next generation cellular networks, nearby users may set up a direct D2D link with the help and control of cellular infrastructure, and use the licensed cellular spectrum or the unlicensed spectrum to communicate via the D2D link without traversing a base station. D2D links underlaying a cellular network may use the same channels also used by cellular users (CUs), which would cause interference [4]. In order to moderate these interferences, different resource allocation and interference management schemes have been proposed [5–12]. In [5–11], static power allocation to D2D pairs is considered, and in [12], dynamic power control for all D2D pairs sharing a common channel orthogonal to the cellular channels in a single cell is proposed.

In general, existing schemes assume a single cell and ignore inter-cell interference, except for [5, 6]. It is also generally assumed that each D2D pair is situated in one cell, and uses only one channel. In [5], a heuristic near-far interference avoidance scheme is proposed to establish a D2D link between two users in adjacent cells. In doing so, the interference from D2D users on CUs is ignored, and the interference from CUs on D2D pairs is mitigated by base stations. In [6], a resource allocation scheme for cross-cell D2D users is proposed, where the interference from D2D users on CUs is ignored; and a heuristic scheme that utilizes information from base stations is devised to identify a channel to be utilized by the D2D pair.

A distributed power control scheme is proposed in [9], where D2D pairs can opportunistically reuse cellular uplink frequency spectrum only when their interference on base stations is within an acceptable margin. This margin forces CUs to increase their transmit power levels, which may be very undesirable.

Establishing D2D links can be facilitated via C-RAN, which is a novel architecture that can reduce network cost and energy consumption, and increase spectral efficiency. In this architecture, the radio unit, called the remote radio head (RRH), is separated from the baseband unit (BBU), and BBUs are pooled together in a cloud environment. The first commercial deployment of C-RAN includes 1000 servers (BBUs) in one BBU Pool, where each server manages 144 base stations [13].

We propose a novel resource allocation scheme for D2D pairs underlaying a LTE-A network with C-RAN architecture. Specifically, the central unit in the C-RAN architecture determines the channels and the transmit power levels for each D2D pair with a view to maximizing the number of active D2D pairs and reused channels while minimizing the aggregate system uplink transmit power. This paper is an extension of [7], and our contributions are as follows. We

This work was supported in part by Tarbiat Modares University, Tehran, Iran.

consider a multi-cell network with inter-cell interferences; assume a D2D link can be established between two users that are situated in different cells, and assume further that each user can transmit on more than one channel at the same time. In addition to offloading backhaul of cellular networks, this scheme is energy efficient for the D2D pairs at the cell boundaries by avoiding the high transmit power in the uplink to the base station at the center of the cell. In this paper, we will not consider many simplifying assumptions that are considered in the literature; and instead focus on real cases. In doing so, we focus on relatively static end-users (i.e., no high speed movement).

The rest of this paper is organized as follows. System model is given in Section 2, followed by formulating the problem in Section 3. The algorithm for optimal resource allocation is proposed and analyzed in Section 4. Simulation results are in Section 5, and conclusions are in Section 6.

2. SYSTEM MODEL

2.1. Infrastructure and User Model

Consider L CUs, M D2D pairs, N orthogonal uplink channels for CUs, and U cells in a LTE-A network with C-RAN architecture. The set of uplink channels is $\mathcal{N} = \{1, \cdots, N\}$, the set of CUs is $\mathcal{C} = \{1, \cdots, L\}$ and the set of D2D pairs is $\mathcal{D} = \{1, \cdots, M\}$. The RRHs are connected to a BBU pool, as shown in Fig. 1.

Each CU uses at least one channel. Each D2D pair can use idle channels or reuse occupied channels to establish direct links. Since single-carrier orthogonal frequency-division multiple access (SC-FDMA) is used for the uplink in LTE-A, CUs do not interfere with each other in a cell, but inter-cell interference exists. The transmitter of a D2D pair (D_Tx) and its receiver (D_Rx) are not required to be in the same cell, i.e., they can be in any of the U cells. Resources (channels and transmit power levels) for CUs and D2D pairs are allocated by the BBU pool.

Carrier aggregation is permitted in LTE-A, and we assume each user can simultaneously transmit on multiple channels. Let $\bar{I}_l^{\mathrm{c}}$ be the maximum number of channels simultaneously used by CU l, and $\bar{I}_m^{\mathrm{d}}$ be the maximum number of channels simultaneously used by D2D pair m. Quality of Service (QoS) is stated in terms of signal-to-interference-plus-noise ratio (SINR) for CUs and D2D pairs. Let $\xi_{l,n}^{\mathrm{c}}$, $\hat{\xi}_{l,n}^{\mathrm{c}}$, $P_{l,n}^{\mathrm{c}}$, and $\bar{P}_{l,n}^{\mathrm{c}}$ be the actual SINR, the required SINR, the actual transmit power and the maximum transmit power of CU l on channel n, respectively. Also, let P_l^{c} and $\bar{P}_l^{\mathrm{c}}$ be the actual aggregate transmit power and the maximum aggregate transmit power of CU l on all channels, respectively. The set of uplink channels simultaneously used by CU l is $\mathcal{N}_l^{\mathrm{c}}$, where $\mathcal{N}_l^{\mathrm{c}} \subset \mathcal{N}$. We have

$$\bar{P}_{l,n}^{\mathrm{c}} = \bar{P}_l^{\mathrm{c}} - \sum_{j \in \mathcal{N}_l^{\mathrm{c}}, j \neq n} P_{l,j}^{\mathrm{c}}, \quad P_l^{\mathrm{c}} = \sum_{n \in \mathcal{N}_l^{\mathrm{c}}} P_{l,n}^{\mathrm{c}}. \quad (1)$$

Let $\xi_{m,n}^{\mathrm{d}}$ and $\hat{\xi}_{m,n}^{\mathrm{d}}$ be the actual SINR and the required SINR

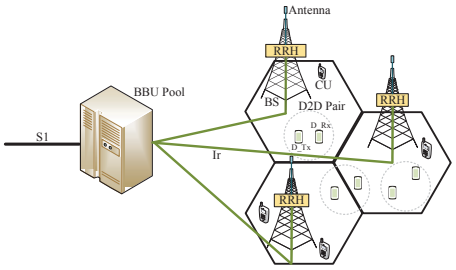


Fig. 1. System model for D2D Links in LTE-A with C-RAN.

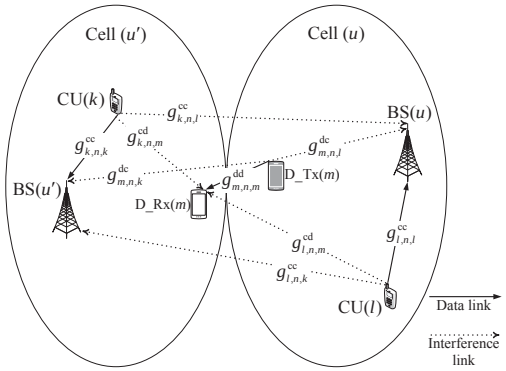


Fig. 2. D2D links underlaying LTE-A network.

of D2D pair m on channel n; and $P_{m,n}^{\mathrm{d}}$ and $\bar{P}_{m,n}^{\mathrm{d}}$ be the actual transmit power and the maximum transmit power of D2D pair m on channel n. Also, let P_m^{d} and $\bar{P}_m^{\mathrm{d}}$ be the actual aggregate transmit power and the maximum aggregate transmit power of D2D pair m on all channels, respectively. The set of uplink channels simultaneously used by D2D pair m is $\mathcal{N}_m^{\mathrm{d}}$, where $\mathcal{N}_m^{\mathrm{d}} \subset \mathcal{N}$. We have

$$\bar{P}_{m,n}^{\mathrm{d}} = \bar{P}_m^{\mathrm{d}} - \sum_{j \in \mathcal{N}_m^{\mathrm{d}}, j \neq n} P_{m,j}^{\mathrm{d}}, \quad P_m^{\mathrm{d}} = \sum_{n \in \mathcal{N}_m^{\mathrm{d}}} P_{m,n}^{\mathrm{d}}. \quad (2)$$

2.2. Channel Model

Let D_Tx(m) and D_Rx(m) be the transmitter and the receiver of D2D pair m, respectively. In Fig. 2, assume CU l is transmitting to base station u on channel n, and simultaneously, CU k is transmitting to base station u' on the same channel. These transmitters cause interference on non-targeted base stations. In this figure, D2D pair m is transmitting on channel n and causing interference on base stations. Besides, the two CUs cause interference on the receiver of D2D pair m. Also in Fig. 2, $g_{m,n,m}^{\mathrm{dd}}$ is the channel gain between the transmitter and the receiver of D2D pair m on channel n, $g_{m,n,l}^{\mathrm{dc}}$ is the channel gain between the transmitter of D2D pair m and the receiver of CU l (i.e., its base station) on channel n, and $g_{l,n,m}^{\mathrm{cd}}$ is the channel gain between CU l and the receiver of D2D pair m on channel n. For simplicity, other users' transmissions are not shown in Fig. 2.

In addition to the distanced-based path loss, both fast fading due to multi-path propagation and slow fading due to shad-

owing are considered. As an example, the channel gain between CU k and the receiver of CU l (i.e., the base station to which CU l is communicating) on channel n is

$$g^{\mathrm{cc}}_{k,n,l} = K\beta_{k,n,l}\zeta_{k,n,l}L^{-\alpha}_{k,n,l}, \tag{3}$$

where K is a constant that depends on system parameters, $\beta_{k,n,l}$ is the fast fading gain with exponential distribution, and $\zeta_{k,n,l}$ is the slow fading gain with log-normal distribution between CU k and the receiver of CU l on channel n. Also, α is the path loss exponent and $L_{k,n,l}$ is the distance between CU k and the receiver of CU l on channel n.

We assume additive white Gaussian noise in each channel. Noise power at the receiver of CU l in channel n is $\sigma^{\mathrm{c}}_{l,n}$, and at the receiver of D2D pair m in channel n is $\sigma^{\mathrm{d}}_{m,n}$.

3. RESOURCE ALLOCATION PROBLEM

When the required SINRs of a D2D pair and CUs using the same channel can be satisfied with bounded transmit power levels, the D2D pair is admissible and the channel is a candidate for reuse. Let $\mathcal{D}'$ ($\mathcal{D}' \subseteq \mathcal{D}$) be the set of all admissible D2D pairs. Each channel is reused by at most one D2D pair. This constraint eliminates mutual interference between different D2D pairs, and simplifies the problem as well. If D2D pair m reuses channel n, then $\rho^{\mathrm{d}}_{m,n}$ is 1, otherwise it is 0. We wish to maximize the number of admissible D2D pairs and the number of reused channels while minimizing the total transmit power for all users (P^{sum}). The problem is

$$\text{Determine} \begin{cases} \rho^{\mathrm{d}}_{m,n}, & \forall m \in \mathcal{D},\ \forall n \in \mathcal{N}, \\ P^{\mathrm{d}}_{m,n}, & \forall m \in \mathcal{D},\ \forall n \in \mathcal{N}, \\ P^{\mathrm{c}}_{l,n}, & \forall l \in \mathcal{C},\ \forall n \in \mathcal{N}, \end{cases} \tag{4a}$$

$$\text{To Maximize} \sum_{m\in\mathcal{D}}\sum_{n\in\mathcal{N}} \rho^{\mathrm{d}}_{m,n}, \tag{4b}$$

$$\text{To Minimize} \sum_{l\in\mathcal{C}}\sum_{m\in\mathcal{D}}\sum_{n\in\mathcal{N}} (P^{\mathrm{c}}_{l,n} + \rho^{\mathrm{d}}_{m,n}P^{\mathrm{d}}_{m,n}), \tag{4c}$$

Subject to:

$$\xi^{\mathrm{c}}_{l,n} = \frac{g^{\mathrm{cc}}_{l,n,l}P^{\mathrm{c}}_{l,n}}{\sigma^{\mathrm{c}}_{l,n} + \sum\limits_{\substack{k\in\mathcal{C}\\k\neq l}} g^{\mathrm{cc}}_{k,n,l}P^{\mathrm{c}}_{k,n} + \sum\limits_{m\in\mathcal{D}} \rho^{\mathrm{d}}_{m,n}g^{\mathrm{dc}}_{m,n,l}P^{\mathrm{d}}_{m,n}}$$
$$\geq \hat{\xi}^{\mathrm{c}}_{l,n}, \qquad \forall l \in \mathcal{C},\ \forall n \in \mathcal{N}, \tag{4d}$$

$$\xi^{\mathrm{d}}_{m,n} = \frac{g^{\mathrm{dd}}_{m,n,m}P^{\mathrm{d}}_{m,n}}{\sigma^{\mathrm{d}}_{m,n} + \sum\limits_{k\in\mathcal{C}} g^{\mathrm{cd}}_{k,n,m}P^{\mathrm{c}}_{k,n}}$$
$$\geq \hat{\xi}^{\mathrm{d}}_{m,n}, \qquad \forall m \in \mathcal{D}',\ \forall n \in \mathcal{N}, \tag{4e}$$

$$\rho^{\mathrm{d}}_{m,n} \in \{0,1\}, \qquad \forall m \in \mathcal{D},\ \forall n \in \mathcal{N}, \tag{4f}$$

$$\sum_{m\in\mathcal{D}} \rho^{\mathrm{d}}_{m,n} \leq 1, \qquad \forall n \in \mathcal{N}, \tag{4g}$$

$$1 \leq \sum_{n\in\mathcal{N}} \rho^{\mathrm{d}}_{m,n} \leq \bar{I}^{\mathrm{d}}_{m}, \qquad \forall m \in \mathcal{D}', \tag{4h}$$

$$0 \leq P^{\mathrm{c}}_{l,n} \leq \bar{P}^{\mathrm{c}}_{l,n}, \qquad \forall l \in \mathcal{C},\ \forall n \in \mathcal{N}, \tag{4i}$$

$$0 \leq P^{\mathrm{d}}_{m,n} \leq \bar{P}^{\mathrm{d}}_{m,n}, \qquad \forall m \in \mathcal{D},\ \forall n \in \mathcal{N}, \tag{4j}$$

$$0 \leq P^{\mathrm{c}}_{l} \leq \bar{P}^{\mathrm{c}}_{l}, \qquad \forall l \in \mathcal{C}, \tag{4k}$$

$$0 \leq P^{\mathrm{d}}_{m} \leq \bar{P}^{\mathrm{d}}_{m}, \qquad \forall m \in \mathcal{D}. \tag{4l}$$

When a CU does not use an uplink channel, its required SINR and transmit power on that channel are zero. Constraints (4d) and (4e) guarantee the required SINR for each CU and for the admissible D2D pairs, respectively. Constraint (4f) makes $\rho^{\mathrm{d}}_{m,n}$ a binary variable. Constraint (4g) ensures that each channel is reused by at most one D2D pair. Constraint (4h) guarantees at least one and at most $\bar{I}^{\mathrm{d}}_{m}$ channels are assigned to each admissible D2D pair. Constraints (4i) and (4j) guarantee that transmit power levels do not exceed their upper bounds. Finally, constraints (4k) and (4l) guarantee that the aggregate transmit power of each user does not exceed its upper bound.

This problem is a mixed integer linear programming (MILP) problem, which is difficult to solve directly. In what follows, we present our approach for solving it.

4. OPTIMAL RESOURCE ALLOCATION

We divide the optimization problem into two sub-problems, solve each sub-problem separately, and combine the results via our proposed algorithm.

In the first sub-problem, we determine if D2D pair m is admissible to reuse channel n. We also determine the transmit power levels of the transmitter of D2D pair m and those CUs that use the same channel n with a view to minimize the total transmit power of all transmitters that use channel n (the D2D pair and CUs).

In the second sub-problem, we identify the optimal reuse channels for all admissible D2D pairs among all candidate reuse channels so that the total transmit power of all users is minimized.

4.1. D2D Admissibility and Optimal Power Control

The D2D pair m can reuse channel n if (4d), (4e), (4i), (4j), (4k), and (4l) are satisfied. In this case we have

$$\begin{cases} \rho^{\mathrm{d}}_{m,n} = 1,\ P^{\mathrm{d}}_{m,n} \neq 0, & \text{(5a)} \\ \rho^{\mathrm{d}}_{i,n} = 0,\ \ P^{\mathrm{d}}_{i,n} = 0, \qquad \forall i \in \mathcal{D},\ i \neq m. & \text{(5b)} \end{cases}$$

Since $\bar{P}^{\mathrm{c}}_{l,n}$ and $\bar{P}^{\mathrm{d}}_{m,n}$ are obtained by (1) and (2), respectively, when (4i) and (4j) are satisfied, (4k) and (4l) are satisfied as well. Hence, (4d), (4e), (4i), (4j), (4k), and (4l) become

$$\begin{cases} \xi^{\mathrm{c}}_{l,n} = \dfrac{g^{\mathrm{cc}}_{l,n,l}P^{\mathrm{c}}_{l,n}}{\sigma^{\mathrm{c}}_{l,n} + \sum\limits_{\substack{k\in\mathcal{C}\\k\neq l}} g^{\mathrm{cc}}_{k,n,l}P^{\mathrm{c}}_{k,n} + g^{\mathrm{dc}}_{m,n,l}P^{\mathrm{d}}_{m,n}} \\ \qquad \geq \hat{\xi}^{\mathrm{c}}_{l,n}, \qquad \forall l \in \mathcal{C}, & \text{(6a)} \\ \xi^{\mathrm{d}}_{m,n} = \dfrac{g^{\mathrm{dd}}_{m,n,m}P^{\mathrm{d}}_{m,n}}{\sigma^{\mathrm{d}}_{m,n} + \sum\limits_{k\in\mathcal{C}} g^{\mathrm{cd}}_{k,n,m}P^{\mathrm{c}}_{k,n}} \geq \hat{\xi}^{\mathrm{d}}_{m,n}, & \text{(6b)} \\ 0 \leq P^{\mathrm{c}}_{l,n} \leq \bar{P}^{\mathrm{c}}_{l,n}, \qquad \forall l \in \mathcal{C}, & \text{(6c)} \\ 0 \leq P^{\mathrm{d}}_{m,n} \leq \bar{P}^{\mathrm{d}}_{m,n}. & \text{(6d)} \end{cases}$$

Since $\hat{\xi}^{\mathrm{d}}_{m,n} > 0$, constraints (6b) and (6d) are satisfied only when $P^{\mathrm{d}}_{m,n} \neq 0$; and (6a) and (6b) can be reformulated as

$$\begin{cases} (g^{\mathrm{cc}}_{l,n,l} P^{\mathrm{c}}_{l,n} - \sum\limits_{\substack{k\in\mathcal{C}\\ k\neq l}} \hat{\xi}^{\mathrm{c}}_{l,n} g^{\mathrm{cc}}_{k,n,l} P^{\mathrm{c}}_{k,n}) - \hat{\xi}^{\mathrm{c}}_{l,n} g^{\mathrm{dc}}_{m,n,l} P^{\mathrm{d}}_{m,n} \\ \quad \geq \hat{\xi}^{\mathrm{c}}_{l,n} \sigma^{\mathrm{c}}_{l,n}, \qquad \forall l \in \mathcal{C}, \qquad (7a) \\ -\sum\limits_{k\in\mathcal{C}} \hat{\xi}^{\mathrm{d}}_{m,n} g^{\mathrm{cd}}_{k,n,m} P^{\mathrm{c}}_{k,n} + g^{\mathrm{dd}}_{m,n,m} P^{\mathrm{d}}_{m,n} \\ \quad \geq \hat{\xi}^{\mathrm{d}}_{m,n} \sigma^{\mathrm{d}}_{m,n}. \qquad (7b) \end{cases}$$

The power vector $\mathbf{p}_{m,n}$ that contains transmit power levels of all CUs and the transmitter of D2D pair m on channel n is

$$\mathbf{p}_{m,n} = \begin{bmatrix} P^{\mathrm{c}}_{1,n} & P^{\mathrm{c}}_{2,n} & \cdots & P^{\mathrm{c}}_{L,n} & P^{\mathrm{d}}_{m,n} \end{bmatrix}^{\mathrm{T}}. \qquad (8)$$

The vector $\boldsymbol{\mu}_{m,n}$ that contains the product of the required SINRs and AWGN power levels in channel n at the receivers of CUs (i.e., their base stations) and the receiver of D2D pair m is

$$\boldsymbol{\mu}_{m,n} = \begin{bmatrix} \hat{\xi}^{\mathrm{c}}_{1,n}\sigma^{\mathrm{c}}_{1,n} & \cdots & \hat{\xi}^{\mathrm{c}}_{L,n}\sigma^{\mathrm{c}}_{L,n} & \hat{\xi}^{\mathrm{d}}_{m,n}\sigma^{\mathrm{d}}_{m,n} \end{bmatrix}^{\mathrm{T}}. \qquad (9)$$

Now, (7a) and (7b) can be combined in a matrix form as

$$\mathbf{A}_{m,n}\mathbf{p}_{m,n} \geq \boldsymbol{\mu}_{m,n}, \qquad (10)$$

where $\mathbf{A}_{m,n}$ that contains all coefficients in (7a) and (7b) is

$$\mathbf{A}_{m,n} = \begin{bmatrix} g^{\mathrm{cc}}_{1,n,1} & \cdots & -\hat{\xi}^{\mathrm{c}}_{1,n} g^{\mathrm{cc}}_{L,n,1} & -\hat{\xi}^{\mathrm{c}}_{1,n} g^{\mathrm{dc}}_{m,n,1} \\ \vdots & \ddots & \vdots & \vdots \\ -\hat{\xi}^{\mathrm{c}}_{L,n} g^{\mathrm{cc}}_{1,n,L} & \cdots & g^{\mathrm{cc}}_{L,n,L} & -\hat{\xi}^{\mathrm{c}}_{L,n} g^{\mathrm{dc}}_{m,n,L} \\ -\hat{\xi}^{\mathrm{d}}_{m,n} g^{\mathrm{cd}}_{1,n,m} & \cdots & -\hat{\xi}^{\mathrm{d}}_{m,n} g^{\mathrm{cd}}_{L,n,m} & g^{\mathrm{dd}}_{m,n,m} \end{bmatrix}. \qquad (11)$$

In matrix form, the power constraints in (6c) and (6d) are

$$\mathbf{0} \leq \mathbf{p}_{m,n} \leq \bar{\mathbf{p}}_{m,n}, \qquad (12)$$

where $\bar{\mathbf{p}}_{m,n}$ is

$$\bar{\mathbf{p}}_{m,n} = \begin{bmatrix} \bar{P}^{\mathrm{c}}_{1,n} & \bar{P}^{\mathrm{c}}_{2,n} & \cdots & \bar{P}^{\mathrm{c}}_{L,n} & \bar{P}^{\mathrm{d}}_{m,n} \end{bmatrix}^{\mathrm{T}}. \qquad (13)$$

Now, the first sub-problem is

$$\text{Minimize} \quad \mathbf{1}^{\mathrm{T}}_{L+1}\mathbf{p}_{m,n}, \qquad (14a)$$

$$\text{Subject to} \begin{cases} \mathbf{A}_{m,n}\mathbf{p}_{m,n} \geq \boldsymbol{\mu}_{m,n}, & (14b) \\ \mathbf{0} \leq \mathbf{p}_{m,n} \leq \bar{\mathbf{p}}_{m,n}, & (14c) \end{cases}$$

where $\mathbf{1}_{L+1}$ is a vector of $L+1$ ones; and its solution is

$$\mathbf{p}^{*}_{m,n} = \begin{bmatrix} P^{\mathrm{c}*}_{1,n} & P^{\mathrm{c}*}_{2,n} & \cdots & P^{\mathrm{c}*}_{L,n} & P^{\mathrm{d}*}_{m,n} \end{bmatrix}^{\mathrm{T}}. \qquad (15)$$

When (14b) and (14c) are satisfied, the D2D pair m is admissible; and the minimum total transmit power of D2D pair m and CUs on channel n is the sum of elements in $\mathbf{p}^{*}_{m,n}$, i.e., $P^{\mathrm{sum}}_{m,n} = \mathbf{1}^{\mathrm{T}}_{L+1}\mathbf{p}^{*}_{m,n}$.

The sub-problem (14) is a linear programming (LP) problem, and can be solved by the Simplex, the Active-Set or the Interior-Point algorithm [14–16]. By solving this sub-problem, all candidate reuse channels for each D2D pair and transmit power levels are obtained. Let $\mathcal{R}_m$ be the set of candidate reuse channels for the D2D pair m. The D2D pair m is admissible and $m \in \mathcal{D}'$ if and only if $\mathcal{R}_m \neq \emptyset$. When there is no candidate reuse channel for the D2D pair m, this D2D pair is starved, and we have

$$\begin{cases} \mathcal{R}_m = \emptyset, & \\ \rho^{\mathrm{d}}_{m,n} = 0, & \forall n \in \mathcal{N}, \\ P^{\mathrm{d}}_{m,n} = 0, & \forall n \in \mathcal{N}. \end{cases} \qquad (16)$$

When there is only one admissible D2D pair m in all cells, its optimal reuse channel is the one that has the smallest $P^{\mathrm{sum}}_{m,n}$, and when there are multiple admissible D2D pairs, the optimal reuse channel for each admissible D2D pair is determined in such a way to minimize the total transmit power of all users.

4.2. Resource Allocation for Admissible D2D Pairs

When only CUs use channel n and no D2D pair reuses it, i.e., when $\rho^{\mathrm{d}}_{m,n} = 0$, $\forall m \in \mathcal{D}$, the problem of obtaining transmit power levels of CUs on channel n that minimize their total transmit power subject to their required SINRs and maximum transmit power levels is also a LP problem [17]. In the sequel, this problem is defined.

When $\rho^{\mathrm{d}}_{m,n} = 0$, $\forall m \in \mathcal{D}$, in order to guarantee the required SINRs without violating the maximum transmit power levels of CUs on channel n, the constraints (4d), (4i) and (4k) for channel n must be satisfied. Since $\bar{P}^{\mathrm{c}}_{l,n}$ is obtained by (1), when (4i) is satisfied, (4k) is satisfied as well. Hence, (4d), (4i), and (4k) become

$$\begin{cases} \xi^{\mathrm{c}}_{l,n} = \dfrac{g^{\mathrm{cc}}_{l,n,l} P^{\mathrm{c}}_{l,n}}{\sigma^{\mathrm{c}}_{l,n} + \sum\limits_{\substack{k\in\mathcal{C}\\ k\neq l}} g^{\mathrm{cc}}_{k,n,l} P^{\mathrm{c}}_{k,n}} \geq \hat{\xi}^{\mathrm{c}}_{l,n}, \qquad \forall l \in \mathcal{C}, \quad (17a) \\ 0 \leq P^{\mathrm{c}}_{l,n} \leq \bar{P}^{\mathrm{c}}_{l,n}, \qquad \forall l \in \mathcal{C}. \quad (17b) \end{cases}$$

When CU l does not use channel n, the required SINR of that CU, i.e., $\hat{\xi}^{\mathrm{c}}_{l,n}$, and its transmit power on that channel, i.e., $P^{\mathrm{c}}_{l,n}$ are zero. Constraint (17a) can be reformulated as

$$g^{\mathrm{cc}}_{l,n,l} P^{\mathrm{c}}_{l,n} - \sum_{k\in\mathcal{C},\ k\neq l} \hat{\xi}^{\mathrm{c}}_{l,n} g^{\mathrm{cc}}_{k,n,l} P^{\mathrm{c}}_{k,n} \geq \hat{\xi}^{\mathrm{c}}_{l,n}\sigma^{\mathrm{c}}_{l,n}, \qquad \forall l \in \mathcal{C}. \qquad (18)$$

Hence, when no D2D pair reuses channel n, the matrix $\mathbf{A}_{0,n}$ contains all coefficients in (17a), the vector $\mathbf{p}_{0,n}$ contains transmit power levels of all CUs on channel n, and the vector $\boldsymbol{\mu}_{0,n}$ contains the product of required SINRs of all CUs and AWGN power levels in channel n at their respective receivers (i.e., their base stations). In this case, the minimum aggregate transmit power of CUs in channel n is the sum of elements in vector $\mathbf{p}^{*}_{0,n}$, i.e., $P^{\mathrm{sum}}_{0,n} = \mathbf{1}^{\mathrm{T}}_{L+1}\mathbf{p}^{*}_{0,n}$.

In LTE, when SINR for a CU on a channel is below a given threshold, that channel is not assigned to the CU [18]. Hence,

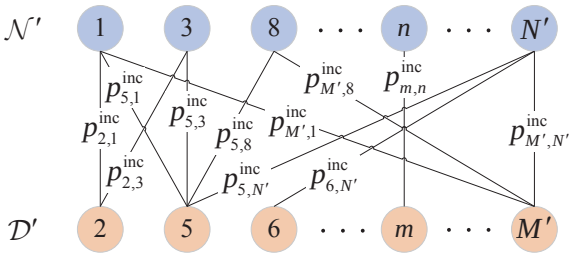


Fig. 3. A bipartite graph for channel assignment problem.

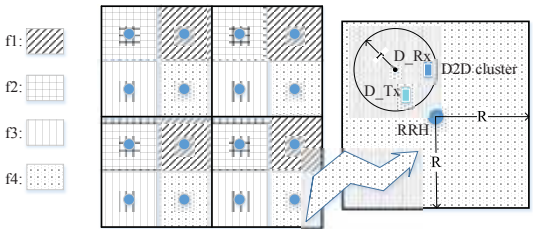


Fig. 4. Cellular layout in simulations.

in the absence of D2D pairs (i.e., no D2D interference), channels are assigned to CUs and (14) has a solution.

When the transmitter of D2D pair m reuses channel n already in use by CUs, it causes interference, which in turn causes CUs to increase their transmit power in order to maintain their required SINRs. In this case, we have $\mathbf{p}^*_{m,n} \geq \mathbf{p}^*_{0,n}$ and the increase in the aggregate transmit power of CUs and the transmitter of D2D pair m on channel n is

$$P^{\text{inc}}_{m,n} = P^{\text{sum}}_{m,n} - P^{\text{sum}}_{0,n}. \tag{19}$$

When there is only one admissible D2D pair m in all cells, its optimal reuse channel is the one that has the smallest $P^{\text{sum}}_{m,n}$ or $P^{\text{inc}}_{m,n}$. This optimal channel can be found via

$$n^*_m = \arg\min_{n \in \mathcal{R}_m} P^{\text{inc}}_{m,n}. \tag{20}$$

When there are multiple admissible D2D pairs, the problem of finding the optimal reuse channel for each admissible D2D pair is the minimum weighted matching problem on the bipartite graph, or in general, is an assignment problem [19, Ch. 10]. This is our second sub-problem, formulated as

$$\min_{\rho^{\text{d}}_{m,n}} \left\{ \sum_{n \in \mathcal{N}'} \sum_{m \in \mathcal{D}'} \rho^{\text{d}}_{m,n} P^{\text{inc}}_{m,n} \right\}, \tag{21a}$$

$$\text{Subject to} \begin{cases} \rho^{\text{d}}_{m,n} \in \{0,1\}, & \text{(21b)} \\ \sum_{m \in \mathcal{D}'} \rho^{\text{d}}_{m,n} \leq 1, \quad \forall n \in \mathcal{N}', & \text{(21c)} \\ \sum_{n \in \mathcal{N}'} \rho^{\text{d}}_{m,n} = 1, \quad \forall m \in \mathcal{D}', & \text{(21d)} \end{cases}$$

where $\mathcal{N}'$ is the union of all candidate reuse channels for all D2D pairs, i.e., $\mathcal{N}' = \mathcal{R}_1 \cup \mathcal{R}_2 \cup \cdots \cup \mathcal{R}_M$.

Fig. 3 shows an example of the assignment problem in (21), where $\mathcal{D}'$ and $\mathcal{N}'$ are the two groups of vertices in the bipartite graph. If channel n is a reuse candidate for the D2D pair m, the vertex m connects to vertex n by an edge mn with the weight of $P^{\text{inc}}_{m,n}$.

The Hungarian algorithm can be used to solve (21). This algorithm is an efficient bipartite assignment scheme whose computational complexity is at most O(n^3), where n is the number of vertices in one group of the symmetric bipartite graph [20, Ch. 3]. The reuse channels for admissible D2D pairs that minimize the aggregate transmit power of CUs and the D2D pairs can be obtained via using the Hungarian algorithm to solve (21). In other words, $\rho^{\text{d}}_{m,n}, \forall m \in \mathcal{D}'$ is obtained, and the transmit power levels are determined via (15). In this way, one cellular channel is assigned to each admissible D2D pair. When assigning more than one channel to each D2D pair is desired, the following algorithm is used.

4.3. Optimal Resource Allocation Algorithm

By solving (14) and (21), each admissible D2D pair is assigned one channel. If more channels are needed by each D2D pair, (14) and (21) should be solved again. Algorithm 1 solves the resource allocation problem (4).

In Algorithm 1, Step 1 solves the first sub-problem, i.e., (14), and Step 2 solves the second sub-problem, i.e., (21). After that, we remove those channels that are assigned to the D2D pairs from the set of available channels, and also remove the D2D pairs to whom the maximum allowable number of simultaneous channels $\bar{I}^{\text{d}}_m$ are assigned to, from the set of D2D pairs requesting channels. These steps iterate until there is no other channel to assign to D2D pairs or the set of D2D pairs is emptied. Simulation shows that our algorithm is convergent. This is because we only consider admissible D2D pairs, and both the simplex and the Hungarian algorithms are convergent [19].

Our proposed resource allocation scheme that utilizes a centralized BBU pool with C-RAN reduces backhaul delay and synchronization errors as compared to the conventional schemes where resources are allocated by base stations and need backhaul links between base stations. Those D2D pairs that have better channels are prioritized, but each admissible D2D pair is allocated at least one and at most $\bar{I}^{\text{d}}_m$ channels.

5. SIMULATION RESULTS

As shown in Fig. 4, we consider 16 square-shape cells each with the width of $2R$. The frequency reuse factor is 4. CUs are uniformly distributed in all cells. As in [9], we adopt the clustered distribution model for D2D pairs, in which the transmitter (D_Tx) and the receiver (D_Rx) of each D2D pair are uniformly distributed in a cluster with radius r; and clusters are uniformly distributed in all cells so that (D_Tx) and (D_Rx) may be situated in different cells.

Algorithm 1 Optimal Resource Allocation Algorithm

$\mathcal{C}$: The set of active CUs
$\mathcal{D}$: The set of D2D pairs
$\mathcal{R}_m$: The set of candidate reuse channels for D2D pair m
$\mathcal{N}$: The set of uplink channels
$\mathcal{N}_m^{\mathrm{d}}$: The set of assigned channels to D2D pair m
Initialization: $\begin{cases} \rho_{m,n}^{\mathrm{d}} = 0, \forall n \in \mathcal{N}, \forall m \in \mathcal{D}, \\ \mathcal{N}_m^{\mathrm{d}} = \emptyset, \forall m \in \mathcal{D}, \\ \mathcal{R}_m = \emptyset, \forall m \in \mathcal{D}. \end{cases}$
while $\mathcal{N} \neq \emptyset$ & $\mathcal{D} \neq \emptyset$ **do**
    Calculate $\bar{P}_{l,n}^{\mathrm{c}}, \forall n \in \mathcal{N}, \forall l \in \mathcal{C}$,
    Calculate $\bar{P}_{m,n}^{\mathrm{d}}, \forall n \in \mathcal{N}, \forall m \in \mathcal{D}$,
    Step 1
    for $\forall m \in \mathcal{D}$ **do**
        for $\forall n \in \mathcal{N}$ **do**
            Calculate $\mathbf{p}_{m,n}^*$ by solving 1st sub-problem
            if 1st sub-problem has a solution **then** $n \in \mathcal{R}_m$
            end if
        end for
        if $\mathcal{R}_m = \emptyset$ **then** $\mathcal{D} = \mathcal{D} - m$
        end if
    end for
    $\mathcal{N} = \mathcal{R}_1 \cup \mathcal{R}_2 \cup \cdots \cup \mathcal{R}_M$
    end Step 1
    Step 2
    for $\forall m \in \mathcal{D}$ **do**
        for $\forall n \in \mathcal{R}_m$ **do**
            Calculate $P_{m,n}^{\mathrm{inc}}$
        end for
    end for
    if $|\mathcal{D}| = 1$ **then** $\begin{cases} n_m^* = \arg\min_{n \in \mathcal{R}_m} P_{m,n}^{\mathrm{inc}} \\ \rho_{m,n_m^*}^{\mathrm{d}} = 1 \\ \mathcal{N}_m^{\mathrm{d}} = \mathcal{N}_m^{\mathrm{d}} + n_m^* \end{cases}$
    else Use the Hungarian algorithm to get $n_m^*, \forall m \in \mathcal{D}$,
        & then $\begin{cases} \rho_{m,n_m^*}^{\mathrm{d}} = 1, \forall m \in \mathcal{D} \\ \mathcal{N}_m^{\mathrm{d}} = \mathcal{N}_m^{\mathrm{d}} + n_m^*, \forall m \in \mathcal{D} \end{cases}$
    end if
    end Step 2
    for $\forall m \in \mathcal{D}$ **do**
        $\mathcal{R}_m = \emptyset$,
        $\mathcal{N} = \mathcal{N} - n_m^*$,
        if $\sum_{n \in \mathcal{N}_m^{\mathrm{d}}} \rho_{m,n}^{\mathrm{d}} = \bar{I}_m^{\mathrm{d}}$ **then** $\mathcal{D} = \mathcal{D} - m$
        end if
    end for
end while

When no D2D link is active, each CU can reach its required SINR with its constrained transmit power. We consider the worst-case interference, where all channels are assigned to CUs and there is no idle channel for D2D pairs. However, D2D pairs may reuse these channels that are in use by CUs. We assume that each CU is assigned one channel and each D2D pair is able to reuse at most 3 channels. Parameter values for simulations are given in Table 1.

Simulation metrics, each averaged for 200 realizations are:

1. Channel reuse ratio: the number of channels reused by D2D pairs divided by the total number of channels.
2. D2D access ratio: the number of admissible D2D pairs divided by the total number of D2D pairs.

Table 1: Simulation Parameters

Parameter	Value
Cell radius (R)	50, 100 m
Channel bandwidth	1 MHz
AWGN power (σ)	-114 dBm
Pathloss exponent (α)	3
Pathloss constant (K)	10^{-2}
Max. CU aggregate power ($\bar{P}_l^{\mathrm{c}}$)	20 dBm
Max. D_Tx aggregate power ($\bar{P}_m^{\mathrm{d}}$)	20 dBm
Req. SINR for a CU ($\hat{\xi}_{l,n}^{\mathrm{c}}$)	Uniform distribution in [0,20] dB
Req. SINR for a D2D pair ($\hat{\xi}_{m,n}^{\mathrm{d}}$)	Uniform distribution in [0,20] dB
Max. number of a CU's channels ($\bar{I}_l^{\mathrm{c}}$)	1
Max. number of a D2D pair's channels ($\bar{I}_m^{\mathrm{d}}$)	3
D2D cluster radius (r)	10, 30, 50, $\cdots$, 90 m
Number of cellular channels (N)	32, 64
Number of cellular users (L)	32, 64
No. of D2D pairs (M)	0.25, 0.4375, $\cdots$, 1 of N
Fast fading gain (β)	Exponential distribution with unit mean
Slow fading gain (ζ)	Log-normal distribution with unit mean and standard deviation of 8 dB
SINR margin (k)	2 dB

3. The increase in the total system uplink throughput when D2D links are allowed as compared to the case in which D2D links are not permitted.

In what follows, we compare the performance of our proposed scheme with that in [9] which assumes a margin k in each CU's required SINR to take into account the interference caused by D2D transmitters. In this comparison, we use the Hungarian algorithm to determine the optimal channels to be reused by admissible D2D pairs.

Figs. 5 to 7 compare the performance of our proposed scheme with that of [9] for different values of r and R. Note that when r increases, the channel gain between D2D pair is decreased, requiring a higher transmit power to guarantee D2D's required SINR. This increases the interference caused by the D2D pair on CUs, which in turn forces CUs to increase their transmit power levels to maintain their required SINRs. Hence, increasing the D2D cluster radius reduces the channel reuse ratio, the D2D access ratio, and the total system uplink throughput. Moreover, when R is increased, the distance from D2D users to CUs and base stations can increase and so, D2D users and CUs may cause less interference on each other. Hence, increasing the cell size increases the channel reuse ratio, the D2D access ratio, and the total system uplink throughput.

Figs. 8 to 10 compare the performance of our proposed scheme with that of [9] for different values of N and M/N. Not that for a given N, increasing M/N is equivalent to increasing M. When M is increased, more D2D pairs may reuse uplink channels, which means that the channel reuse ratio, and the total system uplink throughput can increase. However, the D2D access ratio does not change because both its denominator and numerator are increased. On the other hand, for a given M/N, when N is increased, each D2D pair has more options for choosing its reuse channels, resulting in higher values for all metrics.

Simulations show that via our proposed scheme, the underlay D2D communications in cellular networks enables more

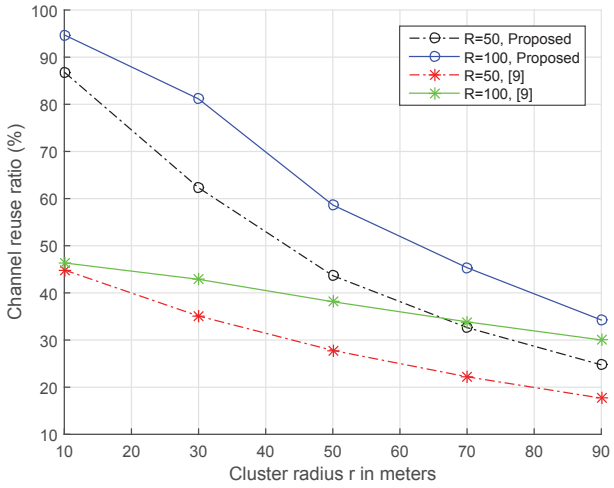


Fig. 5. Channel reuse ratio for $N = 32$ and $M = 15$.

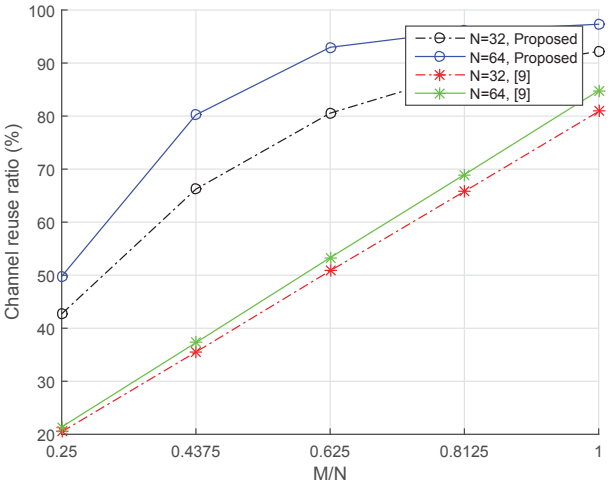


Fig. 8. Channel reuse ratio for $R = 50$ m and $r = 25$ m.

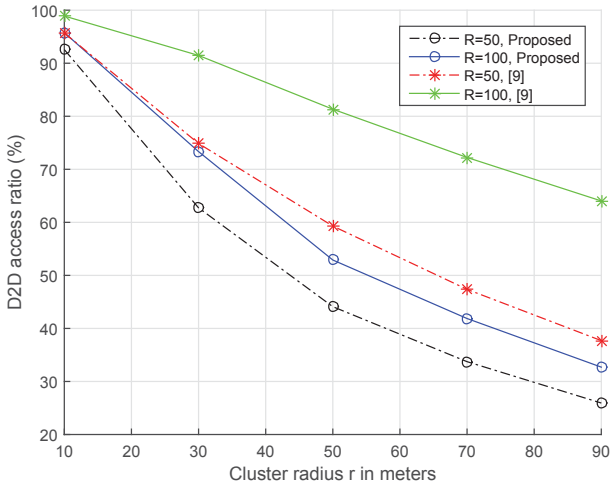


Fig. 6. D2D access ratio for $N = 32$ and $M = 15$.

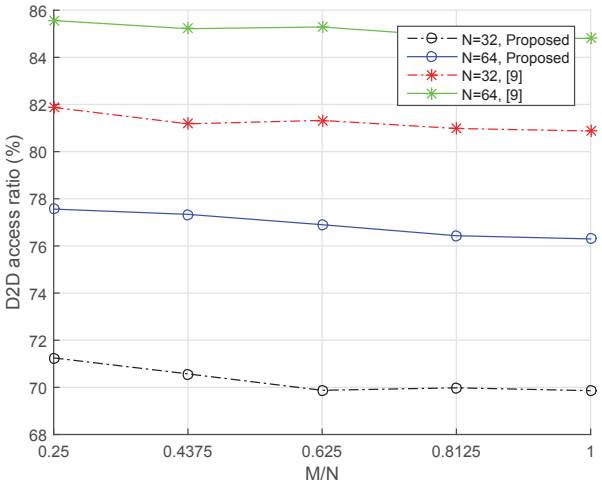


Fig. 9. D2D access ratio for $R = 50$ m and $r = 25$ m.

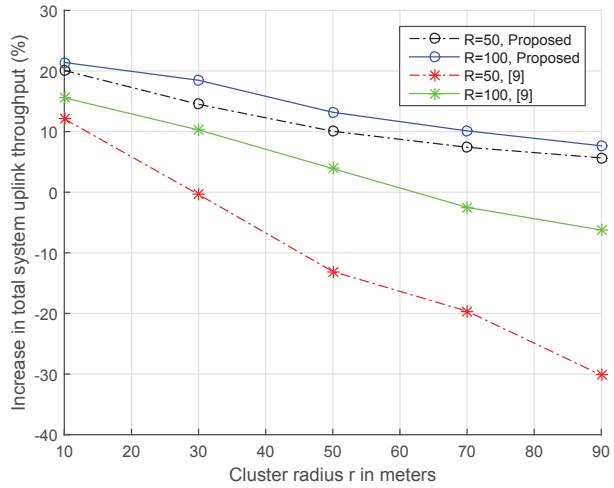


Fig. 7. Increase in total system uplink throughput for $N = 32$ and $M = 15$.

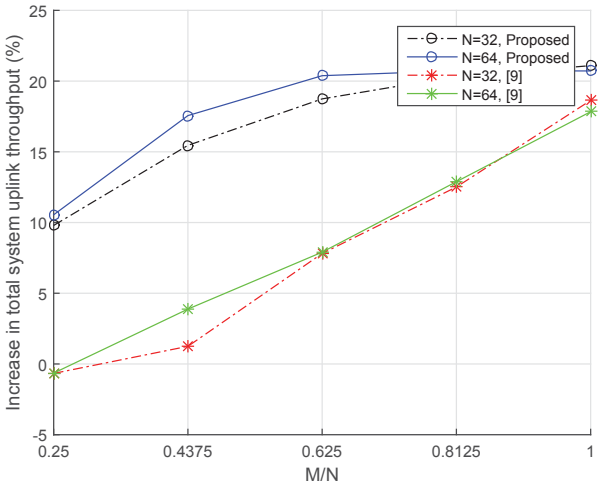


Fig. 10. Increase in total system uplink throughput for $R = 50$ m and $r = 25$ m.

users to communicate and increases the total system uplink throughput. As shown in Figs. 5-10, our proposed scheme outperforms the scheme in [9] except for the D2D access ratio. The access ratio in [9] is higher than our algorithm because maintaining a margin k in each CU's SINR provides the D2D pair with more choices for channels, i.e., more D2D pairs can be admitted, resulting in a higher D2D access ratio. However, maintaining a margin k in each CU's SINR [9] means higher levels of uplink transmit power.

6. CONCLUSIONS

We proposed a novel optimal resource allocation scheme for D2D users in a multi-cell LTE-A network with C-RAN architecture, which increases the total capacity of the system. The scheme maintains the required quality of service in terms of SINR for all users, considers both intracell and intercell interference, permits the D2D transmitter and its receiver to be situated in different cells, and allows each D2D pair to simultaneously utilize multiple channels. We divided the optimization problem into two sub-problems, solved each subproblem separately, and combined the results via our proposed algorithm. Simulation results demonstrate significant improvements in system performance when our scheme is used as compared to [9].

REFERENCES

[1] Cisco, *Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2015-2020 White Paper*, February 2016.

[2] P. Mach, Z. Becvar, and T. Vanek, "In-band device-to-device communication in OFDMA cellular networks: A survey and challenges," *IEEE Commun. Surveys & Tutorials*, vol. 17, no. 4, pp. 1885-1922, June 2015.

[3] A. Checko, H. L. Christiansen, Y. Yan, L. Scolari, G. Kardaras, M. S. Berger, and L. Dittmann, "Cloud RAN for mobile networks - a technology overview," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 1, pp. 405-426, March 2015.

[4] K. Doppler, M. Rinne, C. Wijting, C. B. Ribeiro, and K. Hugl, "Device-to-device communication as an underlay to LTE-Advanced networks," *IEEE Communnications Magazine*, vol. 47, no. 12, pp. 42-49, December 2009.

[5] S. Xu, H. Wang and T. Chen, "Effective interference cancellation mechanisms for D2D communication in multi-cell cellular networks," *Proceedings of 75th IEEE Vehicular Technology Conference*, May 2012, pp. 1-5.

[6] N. Dan, B. Li, B. Lan, and C. JunRen, "Resource allocation over cooperation for cross-cell D2D communication underlaying LTE network," *Proceedings of TENCON 2013*, October 2013, pp. 1-4.

[7] S. M. Alamouti and A. R. Sharafat, "Resource allocation for energy-efficient device-to-device communication in 4G networks," *Proceedings of 7th International Symposium on Telecommunications (IST 2014)*, September 2014, pp. 1058-1063.

[8] N. Lee, X. Lin, J. G. Andrews, and R. W. Heath, "Power control for D2D underlaid cellular networks: Modeling, algorithms, and analysis," *IEEE Journal on Selected Areas in Communnications*, vol. 33, no. 1, pp. 1-13, January 2015.

[9] B. Kaufman, J. Lilleberg, and B. Aazhang, "Spectrum sharing scheme between cellular users and ad-hoc device-to-device users," *IEEE Transactions on Wireless Communications*, vol. 12, no. 3, pp. 1038-1049, March 2013.

[10] R. Yin, C. Zhong, G. Yu, Z. Zhang, K. K. Wong, and X. Chen, "Joint spectrum and power allocation for D2D communications underlaying cellular networks," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 4, pp. 2182-2195, April 2016.

[11] T. D. Hoang, L. B. Le, and T. Le-Ngoc, "Energy-efficient resource allocation for D2D communications in cellular networks," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 9, pp. 6972-6986, September 2016.

[12] W. Wang, F. Zhang, and V. K. N. Lau, "Dynamic power control for delay-aware device-to-device communications," *IEEE Journal on Selected Areas in Communications*, vol. 33, no. 1, pp. 14-27, January 2015.

[13] Rethink Wireless, *Korea Telecom Plans World's First Commercial Cloud-RAN*, December 2011.

[14] G. B. Dantzig, A. Orden, and P. Wolfe, "The generalized simplex method for minimizing a linear form under linear inequality restraints," *Pacific Journal of Mathematics*, vol. 5, no. 2, pp. 183-195, October 1955.

[15] S. Mehrotra, "On the implementation of a primal-dual interior point method," *SIAM Journal on Optimization*, vol. 2, no. 4, pp. 575-601, November 1992.

[16] Y. Zhang, "Solving large-scale linear programs by interior-point methods under the MATLAB environment," *Optimization Methods and Software*, vol. 10, no. 1, pp. 1-31, January 1998.

[17] F. Bock and B. Ebstein, "Assignment of transmitter powers by linear programming," *IEEE Transactions on Electromagnetic Compatibility*, vol. 6, no. 2, pp. 36–44, July 1964.

[18] H. Holma and A. Toskala, *LTE for UMTS - OFDMA and SC-FDMA Based Radio Access*, Wiley, 2009.

[19] M. S. Bazaraa, J. J. Jarvis, and H. D. Sherali, *Linear Programming and Network Flows*, Wiley, 2010.

[20] K. G. Murty, *Network Programming*, Prentice Hall, 1992.

PAPR REDUCTION IN SC-FDMA VIA A NOVEL COMBINED PULSE-SHAPING SCHEME

Naser Ahmadi Moghaddam and Ahmad R. Sharafat

Faculty of Electrical and Computer Engineering, Tarbiat Modares University, Tehran, Iran

ABSTRACT

Peak-to-average-power-ratio (PAPR) is an important parameter that affects the cost of end-user devices in next generation wireless networks. When PAPR is high, the end-user power amplifier's dynamic range should also be high, resulting in costly power amplifiers. Single-carrier frequency-division-multiple-access (SC-FDMA) is used as the air-interface in LTE-Advanced, and this paper proposes a novel and efficient technique for PAPR reduction via pulse shaping for interleaved-FDMA (IFDMA) subcarrier mapping in SC-FDMA. By way of simulations, we show that PAPR can be reduced by 2.11 dB for our novel pulse shaping compared to raised cosine (RC) pulse shaping with QPSK modulation.

Keywords— Peak-to-Average Power Ratio (PAPR), pulse shaping, single-carrier FDMA (SC-FDMA).

1. INTRODUCTION

The rapid proliferation of wireless equipment, including smart phones and other end-user devices, together with widespread provisioning of high speed data and multimedia services and applications that require high spectrum bandwidth have made carrier aggregation an indispensable technology in LTE-Advanced and future generations of wireless networks and services. It is estimated that by 2020, more than 50 billion mobile devices will be connected to cellular wireless networks [1]. LTE-Advanced utilizes Multiple-Input Multiple-Output (MIMO) Orthogonal Frequency Division Multiplexing (OFDM) in the physical layer to substantially increase wireless users' data rates in a manner compatible with TCP/IP and UDP/IP.

However, a significant disadvantage of MIMO OFDM is that the dynamic range of end users' power amplifiers' should be much higher than those in single carrier modulations. This is because in MIMO OFDM with N sub-carriers, the peak of summed sub-carriers can be N times higher than the peak of each sub-carrier. Each sub-system (e.g., the power amplifier) has a finite dynamic range, and can go into saturation if the input exceeds the corresponding dynamic range; resulting in non-linear distortions. This leads to a significant performance degradation and higher error. While an expanded dynamic range is very desirable for avoiding distortions and non-linearities, it significantly increases the hardware cost. One approach for keeping the dynamic range reasonable is to utilize advanced signal processing techniques to reduce the signal's Peak-to-Average Power Ratio (PAPR) and use Single Carrier Frequency Division Multiple Access (SC-FDMA), also called Linear Precoded-OFDM (LP-OFDM).

There are two distinct variants for sub-carrier mapping in SC-FDMA, namely the localized FDMA (LFDMA) and the distributed FDMA (DFDMA). A special case of DFDMA is equally spaced (interleaved) modulation symbols over all subcarriers, called the interleaved FDMA (IFDMA). SC-IFDMA has the advantage of maintaining the input time symbols in each sample, whereas SC-LFDMA and SC-DFDMA have more complicated time samples due to the complex-weighted sum of the input symbols. This implies that the PAPR is higher for SC-LFDMA and SC-DFDMA as compared to SC-IFDMA. Our aim is to develop a pulse shaping scheme for SC-IFDMA to further reduce its PAPR.

PAPR reduction methods can be divided into linear or non-linear techniques. Linear techniques cause no distortion in the time domain (i.e., do not affect the frequency domain) but require extensive calculations. On the other hand, non-linear techniques cause distortion in the time domain (i.e., spread the signal's bandwidth and increase out-of-band emissions), but require less calculations as compared to linear methods. Linear techniques include tone reservation [2], tone injection [2], selective level mapping (SLM) [3-5], partial transmit sequence [6-12] and convex optimization [13-15]. Non-linear techniques include amplitude clipping [16], amplitude clipping and filtering [17,18], and peak cancelation [19].

The PAPR in single-carrier FDMA is much less than the PAPR in multi-carrier systems [20-24]. In [25], pulse shaping is proposed to limit the bandwidth of transmitted signals. However, pulse shaping increases the PAPR in SC-FDMA. Hence, there is a trade-off between reducing PAPR and reducing the bandwidth. In the impulse response of the pulse-shaping filter, the two largest sidelobes have the biggest impact on PAPR; meaning that a pulse shaping filter with a reduced tail size is needed to reduce the PAPR of reduced-bandwidth modulated signals [25-27].

We propose a novel pulse shaping scheme which is a linear combination of Nyquist-I pulse shaping [28] for PAPR reduction in SC-FDMA; and compare the performance of our scheme with other existing pulse shaping techniques, namely those that utilize raised cosine (RC) [28], root raised cosine (RRC) [29], parametric linear pulses (PLP) [29], parametric exponential pulses (PEP) [30], convex [31], concave [31], and parametric linear combination pulses (PLCP) [22].

This work was supported in part by Tarbiat Modares University, Tehran, Iran.

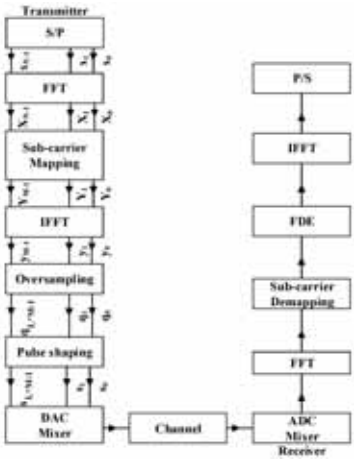


Figure 1: Basedband System model of SC-FDMA

The rest of this paper is organized as follows. In Section 2, SC-FDMA is described, followed by an overview of Nyquist-I pulse shaping in Section 3. In Section 4, our novel pulse shaping scheme is presented. Simulation results are given in Section 5, and conclusions are in Section 6.

2. SC-FDMA

The transmitter of a SC-FDMA system converts a binary input signal into a sequence of modulated sub-carriers. The blocks that comprise a SC-FDMA transmitter are shown in Fig. 1. Note that the FFT and sub-carrier mapping blocks are additional blocks that do not exist in OFDMA systems. The binary input can be modulated using either quaternary phase shift keying (QPSK), 16 level quadrature amplitude modulation (16-QAM) or 64-QAM.

Prior to transmitting a string of symbol, the transmitter first adds a cyclic prefix (CP) to the end of IFFT output to prevent inter-symbol interference (ISI) between consecutive strings, and then applies pulse shaping (linear filtering) in order to reduce out-of-band emissions [20,21,23].

The continuous-time signal can be reconstructed using a low pass filter from N symbols at the receiver end. The maxima of continuous-time signal does not necessarily occur at the maxima of the corresponding binary string. This means that in order to estimate the PAPR of continuous time signal from its samples, one should perform oversampling; which reduces ISI as well. Oversampling the output is formulated by

$$\overline{Y} = \left\{ y_1, \underbrace{0, \cdots, 0}_{L-1}, y_2, \underbrace{0, \cdots, 0}_{L-1}, \ldots, y_{M-1}, \underbrace{0, \cdots, 0}_{L-1} \right\} \tag{1}$$

where L is the oversampling factor. In [32], it is shown that with $L = 4$, an accurate estimation of PAPR can be achieved. The PAPR for SC-FDMA is [33]

$$PAPR = \frac{\max\limits_{0 \le k \le M \times L-1} |s_k|^2}{E\left\{|s_k|^2\right\}}, \tag{2}$$

where s_k is the output of pulse-shaping block in Fig. 1.

In sub-carrier mapping, FFT outputs are mapped to a subset of sub-carriers. This can be done either into a subset of localized or distributed sub-carriers [20,21,23]. In the localized sub-carrier mode, FFT outputs are mapped to a subset of consecutive sub-carriers. In the distributed mapping, the FFT outputs are dispersed in different sub-carriers that are not contiguous and zero amplitude for unused sub-carriers. A special case of distributed SC-FDMA (SC-DFDMA) is the interleaved SC-FDMA (SC-IFDMA), in which, the occupied sub-carriers are equally spaced over the entire bandwidth.

3. NYQUIST-I PULSE SHAPING

The Frequency response of Nyquist-I pulse shaping is [28]

$$H(f) = \begin{cases} T & 0 \le f < B(1-\alpha) \\ TG\left(f - B(1-\alpha)\right) & B(1-\alpha) \le f < B \\ T\left(1 - G\left(B(1+\alpha) - f\right)\right) & B \le f < B(1+\alpha) \\ 0 & f > B(1+\alpha) \end{cases} \tag{3}$$

where B is the signal bandwidth, $T = \frac{1}{2B}$ is the symbol repetition rate, and f denoted frequency. The function $G(f)$ should be differentiable at $f = 0$, and $G(0) = 1$. The roll-off factor α takes values from $[0, 1]$.

The frequency response of the modified Nyquist-I pulse shaping [29] is stated by (4) at the top of next page, where $\gamma_n = \frac{\gamma_0}{\alpha^n B^n}$, $\gamma_0 = G^{-1}(0.5)$, and n is a positive integer. By increasing the value of n, the bandwidth of modified Nyquist-I pulse shaping does not change, i.e., there is no ISI. In what follows, we present different versions of modified Nyquist-I pulse shaping that depend on the choice of $G(f)$.

3.1. Raised Cosine (RC)

For RC pulse shaping, $G(f)$ is

$$G_{\mathrm{RC}}(f) = \cos(f). \tag{5}$$

The frequency response of RC pulse shaping is as in (6) at the top of next page, and its impulse response is

$$h_{\mathrm{RC}}(t) = \mathrm{sinc}\left(\frac{t}{T}\right) \times \frac{\cos\left(\frac{\pi\alpha t}{T}\right)}{1 - \frac{4\alpha^2 t^2}{T^2}}. \tag{7}$$

In [31], convex and concave pulse shaping methods are proposed, which are modified versions of RC pulse shaping. This modification is based on simply allowing any multiple or fraction of cosine cycles to be fitted in the transition region. The frequency response of convex pulse shaping is as in (8) at the top of next page, where $q = \cos^{-1}\left(\frac{\pi}{2}\frac{(d-1)}{d}\right)$ is an amplitude normalization factor for making the frequency

$$H_n(f) = \begin{cases} T & 0 \le f < B(1-\alpha) \\ TG\left(\gamma_n(f - B(1-\alpha))^n\right) & B(1-\alpha) \le f < B \\ T\left(1 - G\left(\gamma_n(B(1+\alpha) - f)^n\right)\right) & B \le f < B(1+\alpha) \\ 0 & f > B(1+\alpha) \end{cases} \quad (4)$$

$$H_{\mathrm{RC}}(f) = \begin{cases} T & 0 \le f < B(1-\alpha) \\ \frac{T}{2}\left\{1 + \cos\left(\frac{\pi}{2B\alpha}\left(f - B\left(1-\alpha\right)\right)\right)\right\} & B(1-\alpha) \le f < B(1+\alpha) \\ 0 & f > B(1+\alpha) \end{cases} \quad (6)$$

$$H_{\mathrm{CV}}(f) = \begin{cases} T & 0 \le f < B(1-\alpha) \\ \frac{T}{2}\left\{1 + q\cos\left(\frac{\pi}{2B\alpha d}\left(f - B\left(1-\alpha d\right)\right)\right)\right\} & B(1-\alpha) \le f < B(1+\alpha) \\ 0 & f > B(1+\alpha) \end{cases} \quad (8)$$

responses to be continuous across the borders between different regions, and its impulse response is

$$h_{\mathrm{cv}}(t) = h_1(t) + h_2(t) \quad (9)$$

where

$$h_1(t) = \mathrm{sinc}(t)\cos(\pi\alpha t) \quad (10)$$

$$h_2(t) = q\frac{\alpha}{2}\sin(\pi t)\left(\mathrm{sinc}(\alpha t - \frac{1}{2\mathrm{d}}) - \mathrm{sinc}(\alpha t + \frac{1}{2\mathrm{d}})\right) \quad (11)$$

The frequency response of the convex pulse shaping is as in (12) at the top of the next page, and its impulse response is

$$h_{\mathrm{cc}}(t) = h_1(t) - h_2(t) + h_3(t) \quad (13)$$

where

$$h_3(t) = 2\sin(t)\left(\mathrm{sinc}(\alpha t) - \cos(\pi\alpha t)\right) \quad (14)$$

3.2. Root Raised Cosine (RRC)

For RRC pulse shaping, $G_{\mathrm{RRC}}(f)$ is

$$G_{\mathrm{RRC}}(f) = \cos^2(f). \quad (15)$$

The frequency response of the RRC pulse shaping is as in (16) at the top of the next page, and its impulse response is

$$h_{\mathrm{RRC}}(t) = \frac{\sin\left(\pi\frac{t}{T}\left(1-\alpha\right)\right) + 4\alpha\frac{t}{T}\cos\left(\pi\frac{t}{T}\left(1+\alpha\right)\right)}{\pi\frac{t}{T}\left(1 - \left(4\alpha\frac{t}{T}\right)^2\right)}. \quad (17)$$

3.3. Parametric Linear Pulses (PLP)

For PLP pulse shaping, $G_{\mathrm{PLP(n=1)}}(f)$ is

$$G_{\mathrm{PLP(n=1)}}(f) = 1 - f. \quad (18)$$

The frequency response of PLP pulse shaping is as in (19) at the top of next page, and its impulse response is

$$h_{\mathrm{PLP}(n=1)}(t) = \mathrm{sinc}\left(\frac{t}{T}\right) \times \mathrm{sinc}\left(\frac{\alpha t}{T}\right). \quad (20)$$

For the modified PLP pulse shaping, $G_{\mathrm{PLP(n)}}(f)$ is as in (18), its frequency response is as in (21) at the top of next page, and its impulse response is

$$h_{\mathrm{PLP}(n)}(t) = \mathrm{sinc}\left(\frac{t}{T}\right) \times \mathrm{sinc}^n\left(\frac{\alpha t}{T}\right). \quad (22)$$

In Section 5, we will show that PAPR is reduced by increasing the value of n in modified PLP pulse shaping, but the signal's bandwidth is spread, resulting in higher ISI and more out-of-band emissions. This means that by choosing the value of n, one can control PAPR reduction and out-of-band emissions.

3.4. Parametric Linear Combination Pulses (PLCP)

The impulse response of parametric linear combination pulses (PLCP) pulse shaping [22] is

$$h_{\mathrm{PLCP}}(t) = \mu h_{\mathrm{PLP}(n=1)}(t) + (1-\mu)\,h_{\mathrm{PLP}(n=2)}(t), \quad (23)$$

where μ is a positive random value.

3.5. Parametric Exponential Pulses (PEP)

For PEP pulse shaping, $G_{\mathrm{PEP}}(f)$ is

$$G_{\mathrm{PEP}}(f) = \exp(f). \quad (24)$$

The frequency response of PEP pulse shaping is as in (25) at the top of next page, and its impulse response is

$$h_{\mathrm{PEP}}(t) = \frac{\mathrm{sinc}\left(\frac{t}{T}\right) \times \left(\frac{2\lambda t}{T}\sin\left(\frac{\pi\alpha t}{T}\right) + 2\cos\left(\frac{\pi\alpha t}{T}\right) - 1\right)}{\left(\frac{\lambda t}{T}\right)^2 + 1}, \quad (26)$$

where $\lambda = \frac{\pi\alpha}{ln2}$.

4. PROPOSED PULSE SHAPING SCHEME

In [34], combination of two different pulse shaping schemes is proposed, and in [22] it is shown that this combination

$$H_{\mathrm{CC}}(f)=\begin{cases} T & 0\le f<B(1-\alpha) \\ \frac{T}{2}\left(1+\frac{2}{\alpha}\left(1-2Tf\right)-q\cos\left(\frac{\pi}{2B\alpha d}\left(f-B\left(1-\alpha d\right)\right)\right)\right) & B(1-\alpha)\le f<B(1+\alpha) \\ 0 & f>B(1+\alpha) \end{cases} \tag{12}$$

$$H_{\mathrm{RRC}}(f)=\begin{cases} \sqrt{T} & 0\le f<B(1-\alpha) \\ \sqrt{\frac{T}{2}}\sqrt{1+\cos\left(\frac{\pi}{2B\alpha}\left(f-B\left(1-\alpha\right)\right)\right)} & B(1-\alpha)\le f<B(1+\alpha) \\ 0 & f>B(1+\alpha) \end{cases} \tag{16}$$

$$H_{\mathrm{PLP(n=1)}}(f)=\begin{cases} T & 0\le f<B(1-\alpha) \\ T(1-\frac{1}{2B\alpha}(f-B(1-\alpha))) & B(1-\alpha)\le f<B \\ T(\frac{1}{2B\alpha}(B(1+\alpha)-f))) & B\le f<B(1+\alpha) \\ 0 & f>B(1+\alpha) \end{cases} \tag{19}$$

$$H_{\mathrm{PLP(n)}}(f)=\begin{cases} T & 0\le f<B(1-\alpha) \\ T(1-\frac{1}{2\alpha^n B^n}(f-B(1-\alpha))^n)) & B(1-\alpha)\le f<B \\ T(\frac{1}{2\alpha^n B^n}(B(1+\alpha)-f)^n))) & B\le f<B(1+\alpha) \\ 0 & f>B(1+\alpha) \end{cases} \tag{21}$$

$$H_{\mathrm{PEP}}(f)=\begin{cases} T & 0\le f<B(1-\alpha) \\ T\exp\left(\frac{\ln 2}{\alpha B}\left(f-B\left(1-\alpha\right)\right)\right) & B(1-\alpha)\le f<B \\ T\left(1-\exp\left(\frac{\ln 2}{\alpha B}\left(B\left(1+\alpha\right)-f\right)\right)\right) & B\le f<B(1+\alpha) \\ 0 & f>B(1+\alpha) \end{cases} \tag{25}$$

reduces PAPR. Our proposed scheme for PAPR reduction is to combine K pulse shaping schemes instead of combining just two schemes.

The impulse response of our pulse shaping scheme is

$$h\left(t\right)=\sum_{i=1}^{K}a_i h_i\left(t\right) \tag{27}$$

where a_i is a coefficient to be optimized in order to reduce PAPR subject to $\sum_{i=1}^{K}a_i=1$, and $h_i(t)$ is the impulse response for each pulse shaping scheme. As an example, for $K=3$ in (27) , and by combining PEP, PLP ($n=1$) and modified PLP ($n=2$) pulse shaping schemes, we get

$$h\left(t\right)=\mu h_{\mathrm{PEP}}\left(t\right)+\nu h_{\mathrm{PLP(2)}}\left(t\right)+\left(1-\mu-\nu\right)h_{\mathrm{PLP(1)}}\left(t\right) \tag{28}$$

where μ and ν are non-negative parameters corresponding to a_1 and a_2, and $a_3=1-\mu-\nu$ in (27).

In order to minimize PAPR, one should obtain optimal values of μ and ν that minimize the absolute values of the impulse response at its first minima (at t_1 in Fig. 2) and the subsequent maxima (at t_2 in Fig. 2) subject to the condition that the absolute value of the impulse response at its minima be higher than that of the subsequent maxima. The objective is to minimize $|h\left(t_1\right)|$ and $|h\left(t_2\right)|$ when $|h\left(t_1\right)|>|h\left(t_2\right)|$. If we minimize $|h\left(t_1\right)|\times|h\left(t_2\right)|$ when $|h\left(t_1\right)|>|h\left(t_2\right)|$, we can fulfill our goal. The following optimization problem captures the above statement.

Table 1: Simulation Parameters

Parameter	Value
No. of subcarriers	512
No. of used subcarriers	128
Sampling frequency	10 MHz
Oversampling factor	4
Roll-off factor (α)	0.22
Sub-carrier mapping	interleaved

$$\begin{aligned} \underset{\mu,\nu}{\text{minimize}} \quad & |h\left(t_1\right)|\times|h\left(t_2\right)| \\ \text{subject to} \quad & |h\left(t_1\right)|>|h\left(t_2\right)| \end{aligned} \tag{29}$$

We utilize numerical methods to obtain the optimal value of $\mu=1$. The complement of cumulative distribution function (CCDF) of PAPR denotes the probability that PAPR of a data block exceeds a given threshold. To find the optimal value of ν, the CCDF diagram of PAPR is drawn for different values of ν. The optimal value of $\nu=2$ minimizes the PAPR.

5. SIMULATION RESULTS

Simulations are performed using MATLAB on a PC with a single-core processor, 2.6 GHz clock frequency, and 3 G Bytes of RAM. Simulation parameters are given in Table 1.

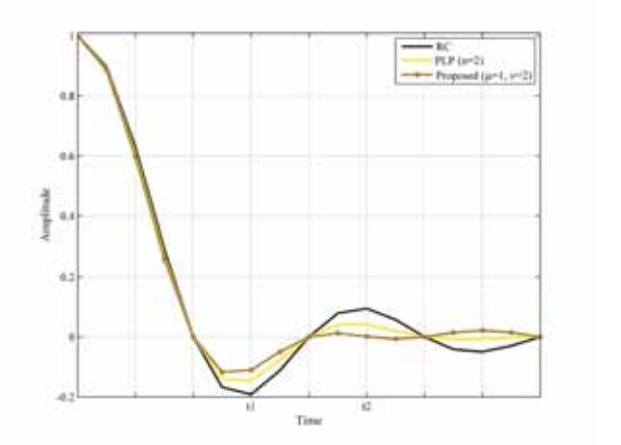


Figure 2: Impulse response of RC, modified PLP ($n = 2$) and our proposed pulse shaping ($\mu = 1.0$ and $\nu = 2.0$) schemes.

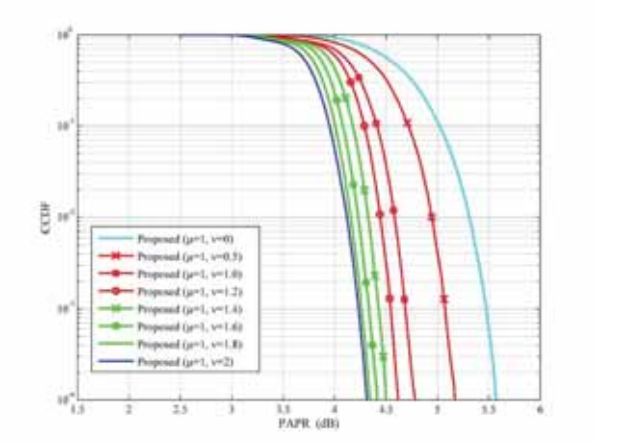


Figure 3: CCDF of PAPR for SC-IFDMA with QPSK modulation via our proposed pulse shaping scheme for $\mu = 1$ and $\nu \in [0, 2]$.

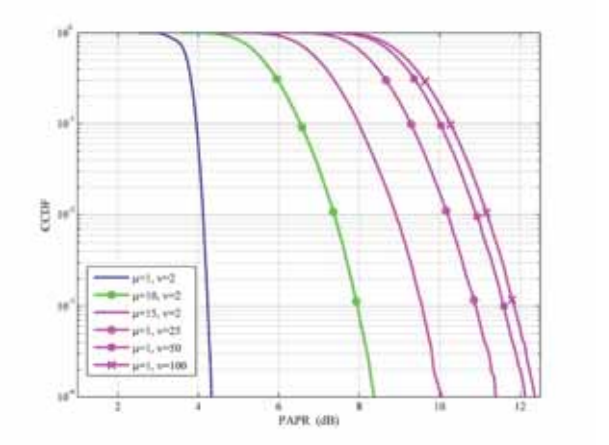


Figure 4: CCDF of PAPR for SC-IFDMA with QPSK modulation via our pulse shaping scheme for $\mu = 1$ and $\nu \in [2, 100]$

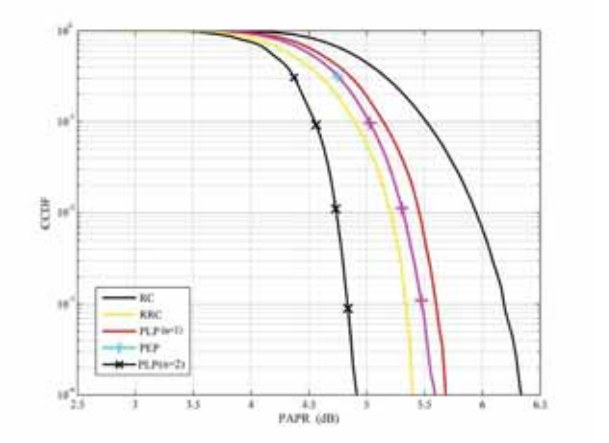


Figure 5: CCDF of PAPR for SC-IFDMA with QPSK via RC, RRC, PLP ($n = 1$), PEP and modified PLP ($n = 2$) pulse shaping schemes.

Fig. 2 compares the impulse response of our proposed pulse shaping scheme with those of RC and modified PLP ($n = 2$). Note lower sidelobes in the impulse response of our scheme.

Figs. 3 and 4 show the CCDF of PAPR for different values of ν. Note that $\nu = 2$ maximizes PAPR reduction.

Fig. 5 compares CCDF of PAPR for RC, RRC, PLP ($n = 1$), PEP and modified PLP ($n = 2$) pulse shaping schemes. It is evident that modified PLP ($n = 2$) pulse shaping has a lower PAPR for any given value of CCDF.

Fig. 6 shows the impulse response of RC, PLP ($n = 1$), PEP and modified PLP ($n = 2$) pulse shaping schemes. It is shown in [25-27] that when the sidelobe maxima is reduced, PAPR is also reduced.

Fig. 7 shows the CCDF of PAPR for modified PLP pulse shaping scheme for different values of n. Note for any given value of CCDF, PAPR is lower for a higher value of n.

Figs. 8 and 9 show the impulse response and the frequency response of the modified PLP for different values of n, respectively. Note in Fig. 9 that out-of-band emissions are higher for higher values of n.

Fig. 10 compares PAPR reduction for RC, PLCP, convex, concave, modified PLP ($n = 2$), and our proposed pulse shaping schemes. Our scheme reduces PAPR by about 2 dB for CCDF of 0.01 as compared to RC pulse shaping.

Fig. 11 shows the frequency response of for RC, PLCP, convex, concave, modified PLP ($n = 2$), and our proposed pulse shaping schemes. Note that out-of-band emissions in our scheme are higher, but acceptable as compared to RC pulse shaping.

In Table 2, the required time for generating a transmit string using different pulse shaping schemes is given. Our proposed and PLCP pulse shaping schemes pass through 3 and 2 parallel filters, respectively; and a weighted sum of the respective filters is obtained for each scheme. If parallel filters and their weighted sum in our scheme are combined into a single block, the time to generate a transmit string in our proposed scheme is comparable to other schemes as shown in Table 3.

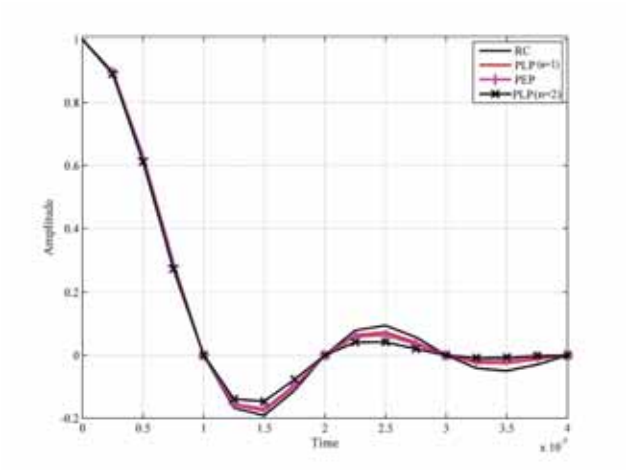


Figure 6: Impulse response of RC, PLP, PEP, and modified PLP schemes.

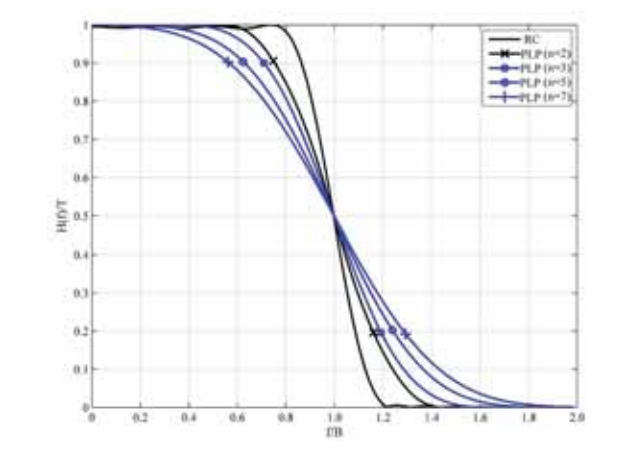


Figure 9: Frequency response of RC and modified PLP schemes.

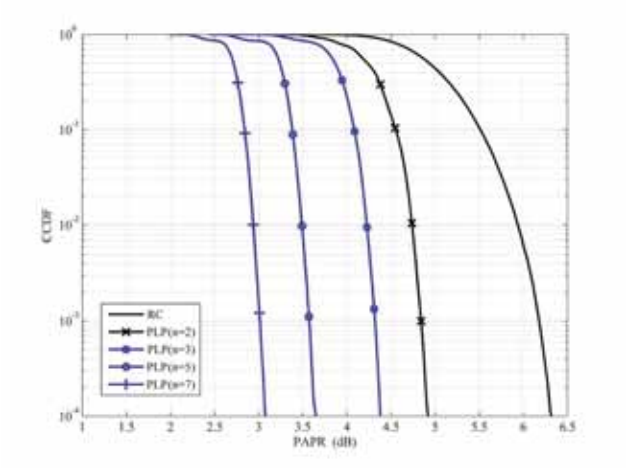


Figure 7: CCDF of PAPR for SC-IFDMA with QPSK via RC and modified PLP schemes.

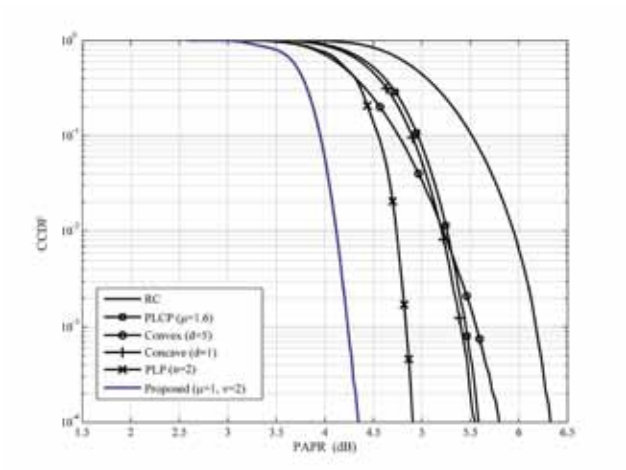


Figure 10: CCDF of PAPR for SC-IFDMA with QPSK via RC, PLCP, convex, concave, modified PLP and our schemes.

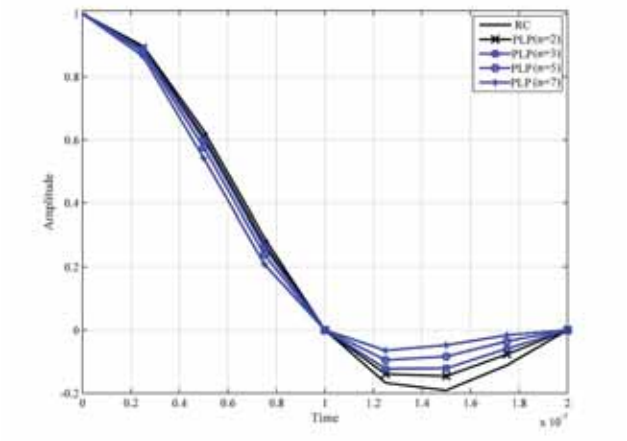


Figure 8: Impulse response of the RC and modified PLP schemes.

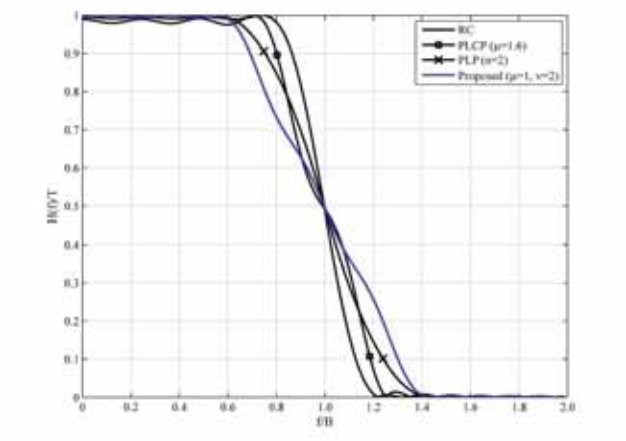


Figure 11: Frequency response of RC, PLCP, modified PLP, and our schemes.

Table 2: Required time to generate a transmit string in different pulse shaping schemes (parallel filters)

Pulse Shaping	SC-IFDMA	
	QPSK(μs)	16QMA(μs)
RC	643.74	720.79
RRC	644.73	722.58
PLP	637.06	718.92
PEP	643.56	717.96
PP ($n = 2$)	637.44	719.50
PLCP ($\mu = 1.6$)	687.09	755.12
Proposed ($\mu = 1$ and $\nu = 2$)	710.26	774.42

Table 3: Required time to generate a transmit string in different pulse shaping schemes (combined filters)

Pulse Shaping	SC-IFDMA	
	QPSK(μs)	16QMA(μs)
RC	643.74	720.79
RRC	644.73	722.58
PLP	637.06	718.92
PEP	643.56	717.96
PP ($n = 2$)	637.43	719.50
PLCP ($\mu = 1.6$)	637.39	719.23
Proposed ($\mu = 1$ and $\nu = 2$)	645.31	720.59

To compare different pulse shaping schemes, the average (β) and the variance (σ^2) of PAPR in dB are shown in Table IV for $\alpha = 0.35$ and sampling frequency of 20 MHz. Note that our proposed scheme has the least values of PAPR average and variance compared to other pulse shaping schemes. Specifically, the average PAPR in our scheme is 2.11 dB, 1.08 dB and 0.67 dB less than those in RC pulse shaping for QPSK, 16-QAM and 64-QAM respectively.

6. CONCLUSION

In this paper, we proposed a novel pulse shaping scheme to reduce PAPR in SC-FDMA systems, and compared its performance with other existing schemes via simulation. Our scheme reduces PAPR more than existing schemes, where its computational cost is higher, but acceptable. The PAPR in our scheme is 2.11 dB, 1.08 dB, and 0.67 dB less than those in RC pulse shaping for QPSK, 16-QAM and 64-QAM respectively.

Table 4: Average values and variances of PAPR for different pulse shaping schemes

Pulse Shaping	QPSK		16QAM		64QAM	
	β	σ^2	β	σ^2	β	σ^2
RC	4.45	0.11	5.49	0.32	5.76	0.32
RRC	3.53	0.05	5.02	0.14	5.55	0.14
PLP	3.93	0.07	5.21	0.25	5.54	0.25
PEP	3.77	0.07	5.12	0.24	5.48	0.24
PP ($n = 2$)	3.10	0.04	4.81	0.15	5.27	0.18
PLCP ($\mu = 1.6$)	3.70	0.08	5.09	0.23	5.45	0.23
Convex ($d = 5$)	3.90	0.16	4.99	0.23	5.39	0.21
Concave ($d = 1$)	3.64	0.08	5.04	0.25	5.42	0.22
Proposed	2.34	0.02	4.41	0.08	5.09	0.10

REFERENCES

[1] L. M. Ericsson, "More than 50 billion connected devices," *White Paper*, Feb. 2011.

[2] T. Wattanasuwakull and W. Benjapolakul, "PAPR reduction for OFDM transmission by using method of tone reservation and tone injection," *5th International Conference on information, Communications and Signal Processing*, Bangkok, pp. 273-277, 2005.

[3] S. H. Muller and J. B. Huber, "A comparison of peak power reduction Schemes for OFDM," *IEEE Global Telecommunications Conference (GLOBECOM)*, Phoenix, vol. 1, pp. 1-5, Nov. 3-8, 1997.

[4] R. W. Bauml, R. F. H. Fisher, and J. B. Huber, "Reducing the peak-to-average power ratio of multicarrier modulation by selected mapping," *Electronics Letters*, vol. 32, no. 22, pp. 2056-2057, Oct. 1996.

[5] H. Breiling, S. H. Muller-Weinfurtner, and J. B. Huber, "SLM peak-power reduction without explicit side information," *IEEE Communications Letters*, vol. 5, no. 6, pp. 239-2041, Jun. 2001.

[6] S. H. Muller and J. B. Huber, "OFDM with reduced peak-to-average power ratio by optimum combination of partial transmit sequences," *Electronics Letters*, vol. 33, no. 5, pp. 368-369, Feb. 1997.

[7] S. H. Muller and J. B. Huber, "A novel peak power reduction scheme for OFDM," *The 8th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, Helsinki,vol. 4, pp. 1090-1094, Sep. 1-4, 1997.

[8] A. D. S. Jayalath and C. Tellambura, "Adaptive PTS approach for reduction of peak-to-average power ratio of OFDM signal," *Electronics Letters*, vol. 36, no. 14, pp. 1226-1228, Jul. 2000.

[9] L. J. Cimini and N. R. Sollenberger, "Peak-to-average power ratio reduction of an OFDM signal using partial transmit sequences," *IEEE Communication Letters*, vol. 4, no. 3, pp. 86-88, Mar. 2000.

[10] C. Tellambura, "Improved phase factor computation for the PAR reduction of an OFDM signal using PTS," *IEEE Communication Letters*, vol. 5, no. 4, pp. 135-137, Apr. 2001.

[11] S. H. Han and J. H. Lee, "PAPR reduction of OFDM signals using a reduced complexity PTS technique," *IEEE Signal Processing Letters*, vol. 11, no. 11, pp. 887-890, Nov. 2004.

[12] S. H. Müller and J. B. Huber, "A comparison of peak power reduction schemes for OFDM," *IEEE Global Telecommunications Conference (GLOBECOM)*, Phoenix, vol. 1, pp. 1-5, Nov. 3-8, 1997.

[13] B. Gayathri and V. Nandalal, "PAPR reduction in OFDM systems using customized convex optimization in differential scaling," *Latest Trends in Circuits, Systems, Signal Processing and Automatic Control*, Salerno, Jun. 3-5, 2014.

[14] A. Aggarwal and T. H. Meng, "Minimizing the peak-to-average power ratio of OFDM signals via convex optimization," *IEEE Global Telecommunications Conference*, San Francisco, vol. 4, pp. 2385-2389, Dec. 1-5, 2003.

[15] A. Aggarwal and T. H. Meng, "A convex interior-point method for optimal OFDM PAR reduction," *IEEE International Conference Communications*, Seoul, vol. 3, pp. 1985-1990, May 16-20, 2005.

[16] R. O'Neill and L. B. Lopes, "Envelope variations and spectral splatter in clipped multicarrier signals," *6th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, vol. 1, Toronto, Sep. 27-29, 1995.

[17] X. Li and L. J. Cimini, "Effect of clipping and filtering on the performance of OFDM," *IEEE Communication Letters*, vol. 2, no. 5, pp. 131-133, May 1998.

[18] J. Armstrong, "Peak-to-average power reduction for OFDM by repeated clipping and frequency domain filtering," *Electronics Letters*, vol. 38, no. 8, pp. 246-247, Feb. 28, 2002.

[19] H. Jeo, J. No and D. Shin, "A new PAPR reduction scheme using efficient peak cancellation for OFDM systems," *IEEE Transactions on Broadcasting*, vol. 58, pp. 619-628, Dec. 2012.

[20] H. G. Myung, J. Lim, and D. J. Goodman, "Peak-to-average power ratio of single carrier FDMA signals with pulse shaping," *IEEE 17th International Symposium on Personal, Indoor and Mobile Radio Communications*, Helsinki, pp. 1-5, Sep. 11-14, 2006.

[21] H. G. Myung, "Introduction to single carrier FDMA," *15th European Signal Processing Conference*, Poznan, pp. 2144-2148, Sep. 3-7, 2007.

[22] C. A. Azurdia-Meza, K. J. Lee, and K. S. Lee, "PAPR reduction in SC-FDMA by pulse shaping using parametric linear combination pulses," *IEEE Communication Letterrs*, vol. 16, no. 2, pp. 2008-2011, Dec. 2012.

[23] H. G. Myung, J. Lim, and D. J. Goodman, "Single carrier FDMA for uplink wireless transmission," *IEEE Vehicular Technology Magazine*, vol. 1, pp. 30-38, Sep. 2006.

[24] C. A. Azurdia-Meza, "PAPR reduction in SC-IFDMA uplink system using parametric pulses," *IEEE Latin-America Conference on Communications (LATINCOM)*, Santiago, pp. 1-5, Nov. 24-26, 2013.

[25] G. Huang, A. Nix, and S. Armour, "Impact of radio resource allocation and pulse shaping on PAPR of SC-FDMA signals," *IEEE 18th International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, Athens, vol. 1, pp. 1-5, Sep. 3-7, 2007.

[26] B. Farhang-Boroujeny, "A square-root Nyquist (M) filter design for digital communication systems," *IEEE Transactions Signal Processing*, vol. 56, no. 5, pp. 2127-2132, May 2008.

[27] B. Chatelain and F. Gagnon, "Peak-to-average power ratio and intersymbol interference reduction by Nyquist optimization," *IEEE 60th Vehicular Technology Conference*, vol. 2, pp. 954-958, Sep. 26-29, 2004.

[28] H. Nyquist, "Abridgment of certain topics in telegraph transmission theory," *Journal of the A.I.I.E*, vol. 47, no. 3, pp. 214-217, Mar. 1928.

[29] N. C. Beaulieu and M. O. Damen, "Parametric construction of Nyquist-I pulses," *IEEE Transactions on Communications*, vol. 52, no. 12, pp. 2134-2142, Dec. 20, 2004.

[30] L. E. Franks, "Further results on Nyquist's problem in pulse transmission," *IEEE Transactions Communication Technology*, vol. COM-16, pp. 337-340, Apr. 1968.

[31] S. Mohapatra, "Peak to average power reduction (PAPR) by pulse shaping using a modified raised cosine filters," *Annual IEEE India Conference (INDICON)*, Gujarat, pp. 1-4, Dec. 18-20, 2009.

[32] C. Tellambura, "Computation of the continuous-time PAR of an OFDM signal with BPSK subcarriers," *IEEE Communication Letters*, vol. 5, pp. 185-187, May 2001.

[33] S. H. Han and A. H. Lee, "An overview of peak-to-average power ratio reduction techniques for multicarrier transmission," *IEEE Wireless Communications*, vol. 12, no. 2, pp. 56–65, April 2005.

[34] P. Sandeep, S. Chandan and A. K. Chaturvedi, "ISI-free pulses with reduced sensitivity to timing errors," *IEEE Communications Letters*, vol. 9, no. 4, Apr. 2005.

ACCELERATING THE INTRODUCTION OF SPECTRUM SHARING USING MARKET-BASED MECHANISMS

Fernando Beltrán

University of Auckland Business School

ABSTRACT

Spectrum management needs to be effective, in that spectrum must be allocated to the right uses, and efficient, in that spectrum must be assigned to those that value it the most. Technological advances and demands for further spectrum availability from mobile broadband operators (among others) require spectrum management to timely and firmly incorporate schemes to increase the technical efficiency of spectrum utilisation. One such scheme is spectrum sharing which has the potential to result in higher spectrum utilisation and greater spectrum value. In such context allocation and assignment, two critical functions to manage the spectrum, are also discussed. It is argued that in the course of deciding about allocation and assignment of spectrum, a spectrum authority can and should include market-based mechanisms that incentivise incumbents to share spectrum needed by entrants.

Keywords— Spectrum management, spectrum sharing, ICT, General Purpose Technologies, licensed shared access, unlicensed spectrum.

1. INTRODUCTION

Radio spectrum has played, is playing and will play a fundamental role in the development of communications networks and services. The unparalleled rise of the cell phone and the quiet revolution in data communications brought about by Wi-Fi and other wireless data technologies demonstrate that effective management of the spectrum is the foundation to robust wireless markets and innovative wireless services.

This paper examines spectrum management as the vehicle deployed by governments to achieve key objectives such as maximising the value of spectrum, its efficient utilisation and its benefits to society. If spectrum management creates conditions for achieving such objectives, it will grant spectrum the affordances that make Information and Communications Technologies (ICT) the kind of disruptor that exhibits considerable technological progress, pervasive use in a wide range of economic sectors, a booster for complementary innovations and a generator of important spillover effects.

Essentially spectrum management is the government function that organises and regulates the utilisation, allocation and assignment of blocks of frequencies so that interference between uses in contiguous bands is minimised. In order to address scarcity and underutilisation of spectrum, spectrum sharing is taking high priority in the agenda of spectrum authorities in many countries. Such institutions have initiated reviews of their national guidelines for spectrum management in order to incorporate spectrum sharing to their regular processes of spectrum allocation and assignment. A review of spectrum sharing policies is included in a later section.

Spectrum management has evolved from its early days when it was focused on interference avoidance [8] to a more modern view whereby, additionally, the spectrum authority seeks to maximise its value [7]. Spectrum value is defined by establishing who derives value from its usage and what the right measurement of value must be. A broad conception of value as found in Barwise et al [2] considers three components of value: private use value, private external use value and social value.

Such approach can enhance a Spectrum Authority's (SA) arsenal of policy tools and regulations when it is applied to spectrum sharing as a management scheme aimed to increase the effectiveness of spectrum allocations and the efficiency of spectrum assignments. In particular, the decision-making process that leads to the assignment of frequency bands to competing parties is based on administering an auction that sells a number of blocks in a number of geographical areas. Although traditionally auctions have assigned spectrum licenses on an exclusive basis, using them to assign shared rights is not only conceivable but possibly an efficient pathway to introduce market mechanisms to decide who gets to share the spectrum.

This paper will first introduce, in Section 2, the concept of General Purpose Technologies -which helps explain the importance of certain technologies to the economy at large- to support its argument that Information and Communications Technologies are of such kind. Spectrum management is discussed in Section 3, while particular aspects and variants of spectrum sharing are presented in Section 4. In Section 5 the paper argues that market-based mechanisms can be used when a spectrum authority decides

to broaden the role of spectrum sharing in future processes of spectrum allocation and assignment. In section 6 conclusions are drawn.

2. ICT AS A GPT

It is not an overstatement that ICT has been a key contributor to economic growth. Over the second half of the 20th century wireless communications have fueled the development of ICT with an unprecedented growth in both reach and scope. Ever since ALOHA started to transmit packets of data back in the late 1960s the smart manipulation of data signals that use the spectrum to propagate and carry information has been a continuous source of innovation in wireless communications. Looking forward the smart and efficient use of the radio spectrum will keep adding to innovation in ICT.

One theoretical element that helps comprehend ICT's impact on economic growth is the concept of General Purpose Technology (GPT). GPT has been defined as "a single generic technology, recognizable as such over its whole lifetime, that initially has much scope for improvement and eventually comes to be widely used, to have many uses, and to have many spillover effects" [12]. A GPT is usually a crude technology that evolves to encompass a broad range of uses, usually applied in the production of a wide range of outputs [12]. In a sense a GPT is an opportunity enabler that opens the door to other technologies and uses, its efficiency improving as it diffuses through an economy. GPTs may transform the economic, social and political structures that embrace them; such GPTs are known as "transforming GPTs". As an illustration, although not widely accepted, Lipsey et al. [12] list 24 transforming GPTs throughout modern human history including domestication of animals, wheel, iron, printing, internal combustion engine, electricity, mass production, computer, Internet and biotechnology.

GPTs are innovations that are slow to develop and diffuse yet gradually impact most activities in the economy. GPTs are defined within historical time, i.e., they start at a point in time, they expand to eventually mature and give way to other technological changes [12]. ICT is generally recognized as a GPT. This means that ICT is seen as sharing GPTs' unique characteristics, characteristics that differentiate them from conventional economic goods and services. It is their historical dimension that makes GPTs contribute to business cycles. For instance, in its first phase, ICT diverts resources from other economic activities with a corresponding slowdown of the economy. Some argue that this is partially what happened with ICT and the productivity slowdown from the early seventies to the nineties [1].

Characterising ICT as a GPT, Rincon et al [18] state that ICT exhibits considerable technological progress and its use is pervasive in a wide range of economic sectors; it is also considered a booster for complementary innovations and a generator of important spillover effects. In the wireless communication sector those characteristics are propelled by access and utilisation of the radio spectrum. The spectrum's transformation from a carrier of analog signals to a carrier of digital information has allowed many sectors to take advantage of its pervasive use. The utilisation of spectrum frequencies continues to fuel a high rate of technological progress both for manufacturers of wireless devices and equipment as well as for developers of software solutions for wireless applications. In their study Rincon et al [18] found that spillover effects, although negative in the short-term, turn to positive after about five years of initial investments. They also find that spillovers across industries are positive and significant.

Seen through the lens of the GPT concept, the effects of ICT have not only been beneficial to the technology and telecommunications sectors per se but to the wider economy across and society. Deployment and innovation in ICT has become intimately linked to the smart utilisation of the radio spectrum. However, since spectrum is a public resource traditionally overseen by government, spectrum management has been rather slow and struggles to keep up with innovations.

3. SPECTRUM MANAGEMENT

Since the times when radio spectrum was first used to guide ships on their Northern Atlantic routes and later commercially by early radio stations, the need for organising it in channels without signal interference was quickly acknowledged [8]. Minimising interference and assigning bands to a handful of uses were therefore the focus of early spectrum management. Fast forward to the second decade of the 21st century and the scope, complexity and diversity of tasks associated with spectrum management are daunting. Its modern conception as stated by Cave et al is '*to maximise the value that society gains from the radio spectrum by allowing as many efficient users as possible while ensuring that the interference between different users remains manageable*' [7]. In other words in addition to interference manageability, Spectrum Authorities (SA) seek to allow more users to use spectrum and more value to be derived from it [3], [4].

Currently the most noticeable trend in spectrum management is a shift away from command-and-control to a market-based approach where users and applicants to licenses act within an incentive-based institutional framework expected to lead to efficient spectrum use. When foreseeing the potential commercial use of a band, typically SAs have relied, first, on deciding the type of use the band will be given to, and then, providing a license to one or more operators for its exploitation. The former is known as **spectrum allocation** and the latter is known as **spectrum assignment** [7].

Every SA seeks to keep excessive interference from occurring and in so doing they keep tight control on allocating spectrum to uses in such a way that similar services tend to cluster in similar bands. The SA is also charged with assigning the spectrum to diverse users. When

spectrum supply exceeded demand, assignment would occur in a very straightforward basis; as time went by and technological advances started to demand more and more spectrum, lotteries and other administrative processes were used to assign the spectrum. Soon the flaws of lotteries were revealed and a long held theoretical proposal became reality: auctions became the mechanism of choice to assign radio frequencies to users. Still, assignment entailed licensing, so licences would protect the holder from undue interference from other spectrum users. Licences were issued on relatively long periods of time, which meant risks associated with technology and policy changes were eliminate in favour of the licensee.

A range of services and applications have benefited from decisions to exempt users of certain bands from licensing. Cordless phones, remote controls and wireless Local Area Networks operate on that basis. The decision to allow for such commons, especially for Wi-Fi technology, has opened immense opportunities for innovation in wireless communications. Put simply, it reveals the importance of having tried a non-conventional way of spectrum utilisation.

However, SAs face quite a number of challenges. Spectrum management needs to embrace new approaches and, in some cases, a full overturn of traditionally held views. It is the case of spectrum value maximisation. For example, in the UK, the Department for Culture, Media and Sport, DCMS, [2] identifies three components of value: private user value, private external value and social value. Private user value, also known as the economic value of spectrum is defined as the present value of the discounted future profits earned by way of using the spectrum [3], [4]. Value is affected by revenues, costs and uncertainty. Private external user value refers to the externalities that arise from the use of spectrum by other users. Social value of spectrum gathers all expressions of value that are not directly attributable to economic activities that may profit from spectrum. Most social value resides in the spectrum allocated to defense, security and public affairs.

4. SPECTRUM SHARING

The preceding discussion pointed at the need for more spectrum management flexibility since the "command and control" or long-term exclusive use approach has proven to be quite a static and rigid management approach. It has been argued that such models have contributed to spectrum scarcity [17]. From a technical perspective, a more dynamic spectrum management is possible if and when new techniques that enable the redefinition of radio frequency parameters, such as the frequency, modulation or output power are allowed to operate.

The combination of new technology, improved radio transmission techniques, and flexible and innovative rules on the use of spectrum facilitate an increasingly attractive aspect of spectrum management which is gathering renewed interest and, in a way, being demanded by enthusiasts and experts: the shared use of the spectrum.

Spectrum sharing is a spectrum utilisation scheme that allows two or more parties to utilize the same range of frequencies while no exclusivity is granted to any of them. In [10] techniques that facilitate spectrum sharing are primarily divided into coordinated, which require coexisting radio-frequency (RF) systems exchange information to share the same frequency band, and uncoordinated, by which RF systems adjust their operation to coexist with other RF systems with little information to share. Coordinated techniques are based on either properties of the communications infrastructure – as in FDMA, TDMA or CDMA - or use channel-based control methods, such as CSMA/CA. On the other hand, uncoordinated techniques include: dynamic channel selection, adaptive frequency hop, listen-before-talk, distributed control power, and cognitive radio [10].

Spectrum sharing started with decisions that split the use of bands, with strong provisions against interference, and has evolved into decisions that have designated some spectrum bands as unlicensed and therefore free to be used by any device within technical parameters dictated by standards. More recently spectrum sharing has found its place into regulatory frameworks promulgated by some SAs. If policy makers and SAs are receptive to spectrum sharing, it may develop as a key spectrum management tool to use, allocate and assign spectrum to achieve the main management objectives discussed above of value maximisation and interference minimisation.

Milgrom et al. [14] argue that a combination of licensed and unlicensed approaches to spectrum assignment may unfold in innovative services which would in turn lead to increased social benefits. Licenses provide rights and obligations whereas unlicensed used of the spectrum, like a no-frills use, is subject to the hassles and inconveniences of the commons.

Spectrum sharing is not concerned with licensing per se; rather it is a spectrum utilisation scheme that erodes exclusivity in spectrum access and utilisation. In recent years several technologies such as Software Defined Radio and Cognitive Radio (CR) have been developed to make shared use of the spectrum possible. Such radios are capable of providing dynamic access to the spectrum, whereby radio frequency parameters are adjusted dynamically to optimize spectrum usage. A conspicuous example is IEEE 802.22, also known as Wireless Regional Area Network or WRAN, a technical standard that includes CR techniques able to use spectrum allocated to television broadcasters under direct coordination of a central database, which keeps up-to-date information about current band utilisation. WRAN is meant to be deployed in rural, low-density geographical areas where broadband access is non-existent. The standard is first of its kind as it is meant for the opportunistic use of frequencies associated with TV bands – known as white spaces- while allowing no interference.

Either through new technologies or through purely administrative allowances that make it possible for several users to share a band – shifting the burden of agreeing to interference-free operation onto those users - spectrum

sharing arrangements challenge the conventional management approach to commercial use, especially for mobile telecommunications services, that has for long conceived spectrum as a resource that must be granted on an exclusive basis [17].

Spectrum sharing can be implemented in one of a range of variants. For instance, license-exempt bands allow the use of a band by any device that complies with a pre-specified technical standard; IEEE 802.11 or Wi-Fi is the best example of it; or bands shared by licensed and license-exempt applications and, licensed and light-licensed commons [19].

In the UK Ofcom has recently released its Spectrum Sharing Framework [15], which provides mobile and wireless broadband operators with legal room to initiate a request to gain access to share specific bands. After consultation with the market, Ofcom summarised the framework in three main aspects that: 1. state the characteristics of use for prospective users who seek access to shared spectrum; 2. advise on the nature and strength of barriers that may limit the future of spectrum sharing; and, 3. discuss the market and technology regulatory tools and enablers of spectrum sharing. Acknowledging that sharing may be detrimental to a licensee's interests, Ofcom will need to decide when spectrum sharing is economically and technically feasible and how it represents a beneficial alternative to the status quo; otherwise it must maintain the current allocation untouched.

In 2009 New Zealand introduced the Managed Spectrum Park (MSP), a special type of licensed commons that operates in the 2575-2620 MHz band. With such scheme the government has sought to encourage "a flexible, cooperative, low cost and self-managed approach to allocation and use" [13] of the spectrum. A MSP allows access to a number of users – usually operators of communication services such as wireless broadband- to the common band on a shared basis and is intended for local and regional services; applicants to a MSP only seek to cover small geographical areas and do not need or want a nationwide license. Sharing may take several forms: it may be that two or three operators split the available bandwidth in an arrangement whose technical aspects need to be sorted out by private agreements, or it may consist of a geographical split within the licence's region.

The European Commission has established two models for sharing frequencies [9]: CUS or Collective Use of Spectrum, and, LSA or Licensed Shared Access. CUS is a license-exempt mode that allows more than one user to use a spectrum simultaneously and with no requirement for a license; variants of the commons fit within the CUS approach. On the other hand, LSA is a scheme which combines traditional command-and-control management with an explicit allowance to share spectrum; in a LSA a limited number of parties are licensed to totally or partially use the band under sharing rules, which have been approved by the SA and then included as terms in the license [19]. Figure 1 uses two "dimensions" to classify several model types of spectrum utilisation based on whether a license is needed, on one dimension, and exclusivity in the use of spectrum, on the other.

Authorised Shared Access (ASA), a special type of LSA promoted in the European Union, is as a mechanism by which a new licensee is granted temporary access to the spectrum already assigned to an incumbent under the prescription that the incumbent does not use it [19]. ASA will allow an access seekers to deploy cognitive radio techniques that will help it learn about on-the-spot channel availability. Such scheme requires bilateral negotiations between the new licensee and the incumbent. ASA also allows multiple new licensees access to one or more incumbents' licensed spectrum.

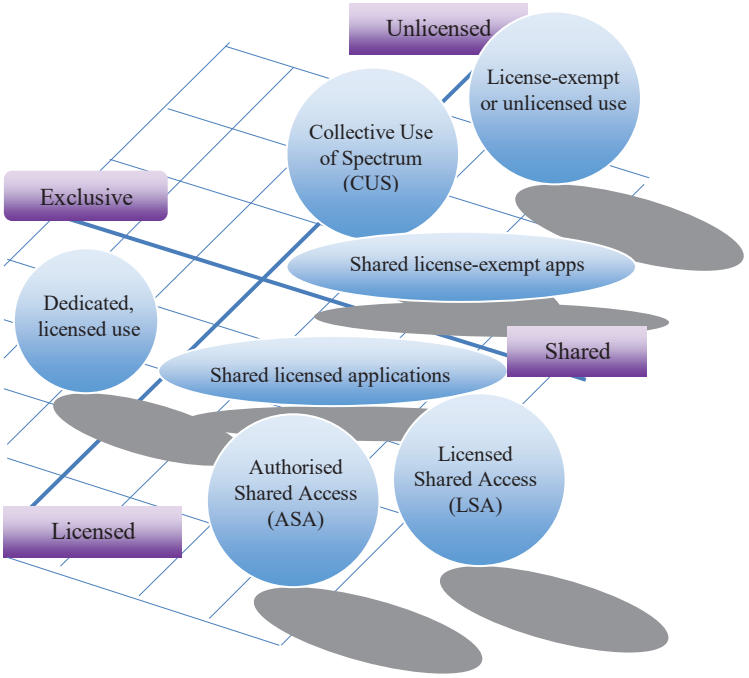


Figure 1: The spectrum sharing landscape

The illustrations provided above indicate that SAs are attracting attention in different regions. Incorporating of spectrum sharing is a process that threatens the conventionally accepted exclusivity of spectrum rights and acknowledges that technology progress and political willingness can come together to favour the introduction of Dynamic Spectrum Management. Such transition surely is a slow process and will require the introduction of legislative and regulatory changes.

5. EFFECTIVENESS OF SPECTRUM ALLOCATION AND EFFICIENCY OF SPECTRUM ASSIGNMENT

This section discusses effectiveness and efficiency in the context of spectrum allocation and spectrum assignment,

respectively. It then turns to argue that spectrum sharing provides a renewed management tool to increase the effectiveness of allocations and the effectiveness of assignments, which should translate into additional benefits to society at large.

The typical problem a SA faces is the following: what is the best use that can be given to a given frequency band and who should be entitled – or licensed – to use it?

As discussed above a SA allocates a spectrum band to a service or services that can occupy the band. International agreements and harmonisation in the possible uses of the spectrum have led SAs to designate radio spectrum bands for their utilisation by prescribed services, which in many cases must follow technical standards of transmission and interference management. Assignment of spectrum to users follows an administrative process that grants them rights over a number of bands Most SAs use auctions for assigning spectrum in bands allocated to commercial communications and broadcasting operations.

Building upon Larbi-Apau and Moseley [11] who state that effectiveness means doing the right thing while efficiency is about doing the things right, it is here suggested that if a SA puts spectrum to its best use, it will maximize the effectiveness of the allocation, and if a SA puts the spectrum in the hands of those who value it the most, it will maximize the efficiency of the assignment. It is clear, from the evolution of some of the so-called ISM bands –such as 2.4 GHz and 5.8 GHz, which were originally conceived for applications non-related to telecommunications, to become the support of Wi-Fi systems, that allowing non-licensed, open access use of the spectrum can create conditions that demonstrated the effectiveness of spectrum management.

In considering spectrum sharing a SA would need to ask itself how the introduction of sharing will affect its main spectrum management functions. In particular it is adequate to wonder what the impact of spectrum sharing would be on allocation and assignment.

In some cases SAs are pressed from different directions to allow new users, particularly wireless broadband operators and other providers of newly developed services to access spectrum bands which are being cleared up from their previous licensees. The US 700 MHz band is one case at hand in which several competing parties demanded that the Federal Communications Commission should designate the digital dividend – those bands left empty by the introduction of digital television – to their particular uses of interest. On one side, mobile telecommunications operators demanded those bands be allocated to 4G services and an auction administered to assign extensive geographical licenses. On the other side, new comers – among them information and contents operators as well as some equipment manufactures - demanded those bands should be designated unlicensed and opened for common exploitation by Wi-Fi services.

One interesting proposal was to design an auction that would solve both problems jointly: allocation and assignment. The auction design, proposed in Bykowsky et al, [6] would allow both types of potential buyers to bid in an auction whose outcomes would determine who would use the spectrum, and in doing so also deciding the service, as well as the amounts to be paid for the licenses. By introducing a slight change to such auction design, which would restrict the use of those bands to the fringe of participants interested in the unlicensed option, a SA would be shifting the burden of spectrum allocation decision to a market-based mechanism that simultaneously over the course of a clock ascending auction would decide whether a band is to be shared or not, and it is, the price to be paid for a shared license. Potential sharers need to solve a collective action problem; in such situations, typically efficiency of the allocation may be compromised as bidders find incentives to free-ride.

Opportunistic use of a frequency band, as enabled by WRAN technology, requires a centralized database to allow CR-based devices to transmit on TV white spaces. Establishing and managing – or outsourcing - such a system requires that the SA covers the costs of equipment and administration. ASAs such as this would probably have to be funded, at least partially, through license fees but most likely will need to be subsidized as the target population is usually sparse and remotely located from urban centres.

A common concern raised by potential new users to SAs is about the efficiency with which government agencies use parts of the radio spectrum, indicating too that they would need shared access to such bands. Upon releasing its spectrum sharing framework [15], Ofcom outlined the request process as one by which a mobile operator or interested party initiate a request to gain access to a specific band that has been either licensed to an incumbent or held by users, such as government agencies, traditional holders of rights. The spectrum access seeker needs to have exhausted a number of options that must precede their sharing aspirations. First, it must have not found any suitable option among the currently available licenses (including license-exempt bands), or found no trading or leasing opportunities. Then, and only then, Ofcom will consider looking into available information to determine whether the request is worth being further investigated [15].

Building provisions to endow spectrum sharing decisions with market-based mechanisms was attempted by Ofcom with its 800 MHz and 2.6 MHz spectrum auction in 2013. A previous 2012 Ofcom's previous consultation on the award of those bands gave ample consideration to auction rules that would lead bidders to reveal preferences for wining bands contiguous to bands won by other bidders with whom potential sharing agreements could be reached [16]. The proposal first explored the pros and cons of allowing bidders to express their preferences consistent with their goal to get blocks contiguous to the blocks won by other bidders with whom they would be sharing the spectrum. Bidders would see benefits from pooling together resources with higher speed and improved quality of service possibly achieved. Bidders, on the other hand, would be exposed to the risk involved in not being able to win the necessary blocks, hence the likelihood of an inefficient auction outcome. This

is another instance of collective action that requires coordination between bidders over the course of the auction, a situation hard to deal with in the context of auction rules that protect anonymity of bidding. In its two-stage auction design Ofcom had proposed three ways to deal with coordination: a) to allow negotiations between winners of the first auction stage, b) to allow joint bidding on the second stage, and, c) to allow bids in the second stage contingent on knowing the identity of potential neighbours (neighbouring bands). We identified significant downsides with each option such that we did not favour any of them. Eventually for its 2013 auction, due to lack of support from respondents and no alternative provisions suggested, Ofcom decided not to facilitate joint bidding or spectrum sharing.

6. CONCLUSIONS

The modern role of spectrum managers has drastically become even more challenging as demand for spectrum has increased remarkably over recent years. Bandwidth-hungry applications and devices seek to get connected over the airwaves is dismal numbers but the availability of spectrum is not keeping up with demand. This paper argues a renewed role for spectrum sharing needs to be embraced by spectrum authorities as other managerial decisions such as reallocation, clearing of non-highly used bands and refarming are not enough. In doing so spectrum management is continuously expected to be serve as the vehicle deployed to maximise the value of spectrum, its efficient utilisation and its benefits to society.

Spectrum sharing can enhance the spectrum authority's capabilities with a management scheme aimed to increase the effectiveness of allocations and the effectiveness of assignments. Market-based mechanisms that include auction for the assignment of rights to share the spectrum are not only conceivable but possibly efficient ways to decide about the best use and user of the spectrum [5].

If spectrum management creates conditions for efficient allocation and assignment through spectrum sharing, it will be fueling the connection between policy and markets that make ICT a pervasive factor in a wide range of economic sectors, a booster for complementary innovations and a generator of important spillover effects.

REFERENCES

[1] P. Aghion, P. Howitt, and G.L. Violante, "General Purpose Technologies and Within-group Inequality", *Journal of Economic Growth*, (2012), 7: 315.

[2] P. Barwise, M. Cave, P. Culham, T. Lavender, N. Pratt, and D. Tambini, "Incorporating Social Value into Spectrum Allocation Decisions", *A report for UK Department for Culture, Media and Sport*, (2015).

[3] C. Bazelon, and G. McHenry, "Spectrum Sharing: Taxonomy and Economics", The Brattle Group, (2014).

[4] C. Bazelon, and G. McHenry, "Spectrum Value", *Telecommunications Policy*, (2013), 37: 737-747.

[5] F. Beltrán, S.K. Ray, and J. and Gutiérrez, "Understanding the current operation and future roles of wireless networks: Co-existence, competition and co-operation in the unlicensed spectrum bands", accepted for publication in *Journal of Selected Areas of Communications*, (2016).

[6] M.M. Bykowsky, W.W. Sharkey, and M., Olson, "A Market-Based Approach to Establishing Licensing Rules: Licensed versus Unlicensed Use of Spectrum", *US Federal Communications Commission, OSP Paper 43*, (2008). https://www.fcc.gov/reports-research/working-papers/market-based-approach-establishing-licensing-rules-licensed-versus

[7] M. Cave, C. Doyle, and W. Webb, "Essentials of Modern Spectrum Management", Cambridge University Press (2007).

[8] R.H. Coase, "The Federal Communications", *Journal of Law and Economics*, (1959), 2:1-40.

[9] European Commission, "Promoting the shared use of Europe's radio spectrum". https://ec.europa.eu/digital-single-market/en/promoting-shared-use-europes-radio-spectrum (2016). (last access, April 28, 2016).

[10] Y. Han, E. Ekicia, H. Kremo, and O. Altintasc, "Spectrum sharing methods for the coexistence of multiple RF systems: A survey", *Ad-Hoc Networks*, In-print.

[11] J. A. Larbi-Apau, and J. L. Moseley, "Performance-based measurement: Action for organizations and HPT accountability", *Perf. Improv.*, (2010), 49: 7–17.

[12] R. Lipsey, K.I. Carlaw, and C.T. Bekhar, Economic Transformations: General Purpose Technologies and Long Term Economic Growth, Oxford University Press, (2005).

[13] MBIE Ministry of Business, Innovation and Employment Managed Spectrum Parks Allocation Rules. New Zealand Government, (2010) www.rsm.govt.nz

[14] P., Milgrom, J. Levin, and A. Eilat, "The case for unlicensed spectrum". Stanford Institute for Economic Policy Research Discussion Paper No. 10-036. (2011).

[15] Ofcom, "A framework for spectrum sharing", available at http://stakeholders.ofcom.org.uk/binaries/consultations/spectrum-sharing-framework/statement/statement.pdf (2016). Last access April 24, 2016.

[16] Ofcom, "Second Consultation on assessment of future mobile competition and proposals for the award of 800 MHz and 2.6 GHz spectrum and related issues", http://stakeholders.ofcom.org.uk/binaries/consultations/award-800mhz/summary/combined-award-2.pdf (2012). Last access May 30, 2016.

[17] PCAST, Report to the President: Realizing the Full Potential of Government-held spectrum to spur economic growth, (2012).

[18] A. Rincon, M. Vecchi, and F. Venturini, "ICT as a General Purpose Technology: Spillovers, Absorptive Capacity and Productivity Performance", UK NIESR Discussion Paper 416, (2013).

[19] SCF Associates, "Perspectives on the value of shared access", Final report for the European Commission, (2012).

SESSION 4

NETWORK EVOLUTION

5G IN RURAL AND LOW-INCOME AREAS: ARE WE READY?

Luca Chiaraviglio,[1,2] *Nicola Blefari-Melazzi,*[1,2] *William Liu,*[3] *Jairo A. Gutierrez,*[3] *Jaap Van De Beek,*[4]
Robert Birke,[5] *Lydia Chen,*[5] *Filip Idzikowski,*[6] *Daniel Kilper,*[7] *Paolo Monti,*[8] *Jinsong Wu,*[9]
1) CNIT, Italy, 2) University of Rome Tor Vergata, Italy, 3) AUT, New Zealand, 4) Lulea University of Technology, Sweden,
5) IBM Research, Switzerland, 6) Faculty of Electronics and Telecommunications, Poznan University of Technology, Poland
7) The University of Arizona, USA, 8) KTH Royal Institute of Technology, Sweden, 9) University of Chile, Chile

ABSTRACT

Current trends in telecommunication networks foresee the adoption of the fifth generation (5G) of wireless networks in the near future. However, a large number of people are living without coverage and connectivity. To face this issue, we consider the possibility of deploying 5G networks in rural and low-income zones. After detailing the current state-of-the-art, we consider the main challenges that need to be faced. Moreover, we define the main pillars to follow in order to deploy 5G networks in such zones, as well as a proposal of a future network architecture.

Index Terms— 5G networks, rural and low-income zones, future Internet, global connectivity

1. INTRODUCTION

The first experiment on the Arpanet (the father of the current Internet) took place on 29th October 1969, with a simple login message exchanged in a network composed of four nodes. Since then, telecommunication networks have incredibly grown, with distributed architectures connecting billions of users. Today, the Internet is considered as a commodity together with energy, water, and food. According to the recent study presented in [1], the Internet is the world's most powerful engine for social and economic growth, and it needs to be open, secure, trustworthy, and accessible to all. In this scenario, the International Telecommunication Union (ITU) has reported that 69% of the world population is covered by the third generation (3G) network [2], which allows users to connect with the Internet. Moreover, the penetration rate of the Internet in North America is above 80% [3]. However, these numbers are hiding the dark side of telecommunication networks, which is the lack of connectivity and/or coverage experienced by a large number of people, in the non-coverage areas, especially rural and low-income ones. Such zones include both low-density regions, but also towns/cities where the Gross Domestic Product (GDP) is extremely low. To this end, at least two billion people are currently experiencing a complete lack of wireless cellular coverage [4], thus preventing them to connect with the rest of the world.

Given this picture, a natural question is: why are there such differences in connectivity and coverage rates across the world? One initial observation is that the telecommunication networks are widely deployed in urban zones rather than in rural and low-income ones. Specifically, users located in urban areas have the possibility to connect to the Internet by means of WiFi, radio access, fixed access and satellite connections. In this context, telecom manufacturers and researchers are focused on the development of the forthcoming 5G technologies, which will be available by 2020. 5G standards are currently investigated by several organizations around the world, including partnerships (such as 5G Public Private Partnership in Europe, IMT-2020 5G Promotion Group in China, The Fifth Generation Mobile Communications Promotion Forum in Japan, 5G Forum in Korea, and 5G Americas) and international events. These efforts aim to build 5G networks that will dramatically improve the user experience, thanks to a sharp increase in the offered data rates, coupled also with extremely low latency times. In this way, services like very high definition video, tactile Internet, virtual reality and Internet of Things will be made available.

5G has several advantages for the spreading of the Internet connectivity (see examples in the white paper of 5G-PPP [5]). Among them, the network introduces a high level of flexibility, which was otherwise not possible with previous technologies. Thanks to this advantage, it is possible to deploy services and network resources where and when they are really needed. In addition, 5G foresees the exploitation of commodity hardware, which opens the way to the development of software solutions implementing networking functions, and potentially decreasing the costs of installing and maintaining devices. Finally, 5G has introduced the concept of "converged solution", where the networks and the services cooperate to deliver high bandwidth and extremely low delay to users.

Even though 5G introduces several positive aspects, the relevant technologies are "urban" in their nature. Specifically, the high performance requirements are made possible by an extremely rich and complex architecture, including: heterogeneous networks of macro and small cells, fronthaul and backhaul transport networks, small computing nodes deployed close to the users, and large data centres. The current models of telecommunication networks, which are business and profit oriented, suggest that 5G networks will be mainly deployed in extremely dense urban zones, where the number of subscribers is sufficiently high to compensate the installation and management costs of the 5G network. On the contrary, rural and low-income zones are less desirable and attractive for operators, since the extremely low

density (and/or low-income) population does not justify the deployment of 5G networks.

In this context, the European Union has launched different calls for research proposals in the context of 5G [6]. However, a specific call of 5G for rural and low-income zones is not included. Rural areas are only indirectly taken into account by the proposals on converged architectures. Such architectures aims at defining a universal network model that can be applied to cities, towns, and small villages. Nevertheless, the associated costs of these networks are still an open issue. In this scenario, the lack of connectivity and coverage for rural and low-income zones is in contrast with the Internet purposes.

In this paper, we take a new view of 5G networks, by looking at them through the lens of rural and low-income subscribers. In this context, several questions arise: What is the current state-of-the-art in the field of 5G research in such zones? What are the main challenges that need to be faced? Is it possible to define a holistic 5G architecture explicitly designed for rural and low-income zones? The goal of this paper is to shed light on these issues, and to define future research directions. Specifically, we believe that a 5G network in rural and low-income zones should be built around the following pillars: i) possibility in deploying cells on board of small Unmanned Aerial Vehicles (UAVs), ii) exploitation of renewable energy sources, iii) reusability of network components and functions, iv) deployment of commodity hardware and v) reduction of CAPital EXpenditures (CAPEX) and OPerating EXpenditures (OPEX).[1]

The rest of the paper is organized as follow. Section 2 reviews the state-of-the-art. The main challenges are reported in Section 3. Section 4 details our vision. Finally, Section 5 concludes our work.

2. STATE-OF-THE-ART

We consider the related work in the following domains: wireless access networks, transport networks, and data centres.

2.1. Wireless Access Networks

Future 5G networks will be characterised by the extensive deployment of small cells [7]. However, this infrastructure requires huge installation and management costs, due to the fact that a large number of sites will be required to be deployed, coupled also with the need of connecting each site to the rest of the network. Currently, one of the great barriers to the development of cellular networks in rural zones is the lack of revenues per square mile [8]. However, even an increase in the competition in the spectrum assignment will not affect the revenue per square mile experienced by operators.

The costs needed to run a wireless network can be divided into two main branches: CAPEX, which are related to the acquisition and the installation of network equipment, and OPEX, which instead incur during the management of the network. Focusing on the former, solutions aiming at the deployment of cellular networks with low CAPEX costs are of mandatory importance. These costs can be reduced by decreasing the number of deployed cells per area size. Moreover, authors of [9] clearly show that the exploitation of massive Multiple Input Multiple Output (MIMO) beamforming is able to reduce the network area power consumption by up to 50% compared to standard cellular networks, while guaranteeing a 10 Mbps cell edge user throughput power. Alternatively, one radical improvement for the reduction of CAPEX costs is the exploitation of new cellular architectures. The Loon project launched by Google [10] aims at providing cellular connectivity by means of balloons naturally moved by the winds of the stratosphere. The same goal is also pursued by the Internet project of Facebook [11], with the main difference that the cells are deployed on board of UAVs flying at high altitude and powered by solar panels.

At the same time, solutions aiming at the reduction of OPEX costs are also necessary. One way to reduce these costs is the exploitation of renewable energy sources. Authors of [12] investigated the adoption of relay nodes powered by renewable sources, showing that the radio resources (in terms of Resource Blocks and transmission power) can be efficiently allocated in order to maximize the user data rates. Moreover, renewable sources can be also exploited to power macro cells. In this context, solar powered cells are being deployed to provide connectivity, mainly in some zones of Africa and Asia [13]. Even though the size of the renewable energy power systems is still an issue [14], the cost of a solar powered cell is comparable with a cell connected to the grid, thus much less than exploiting power coming from expensive diesel generators. In addition, the OPEX costs can be decreased by reducing the power consumption of cells [15]. This is made possible by the exploitation of different power states (such as full power and low power) which are applied over time, by taking advantage of the traffic variability. Finally, another way to reduce the OPEX is the adoption of virtual network elements, for example by means of Cloud Radio Access Networks (C-RAN) [16].

2.2. Transport Network

With the latest 5G requirements of individual user peak access rates between 1 and 50 Gb/s [5], fibre connections become essential and therefore fibre networks will need to reach rural and under-developed areas, where it is important to give priority to other characteristics, i.e., low cost/power; resilience to power disruptions, and low maintenance requirements [17].

When looking at today's fronthaul-backhaul infrastructures, it can be noticed that radio systems employ a "remote radio head" strategy in which the Radio Frequency (RF) signal is sampled and immediately converted to an optical signal for transmission to the Base Band Unit (BBU) processing the base band. These units can be placed at the bottom of the tower or centralized in BBU hotels that can be quite far away

[1] Apart from the connectivity 5G networks may be exploited also to build the Internet of Things paradigm.

from the antenna. In this case however, the radio processing delay requirements become the limitation. In LTE the Up-Link Hybrid Automatic Repeat reQuest (UL-HARQ) has a 4 ms response time requirement that, given the processing time requirements, translates into a maximum fibre distance of 20 km. A small amount of additional headroom can significantly increase the optical reach. This would open the possibility to distribute the low-cost, low-power, high-reliability radio heads across an area extending 20 km from a centralized processing location, where higher reliability, security, and greater efficiency may be possible [18].

When considering core transport networks (i.e., the ones used to move data from/to access networks), renewable energy such as solar will likely be important for many rural areas with limited power availability. Their drawbacks is mainly in their lack of reliability. In this context, several studies have looked at the operation of optical systems in the presence of a variable and renewable energy sources (see e.g., [19]).

Another way to reduce the cost and improve the resource efficiency in a transport network is to introduce advanced network functionalities (i.e., dynamic resource sharing and Network Function Virtualization (NFV)) [20]. This allows for example the allocation of resources on-demand to support specific transport needs that may vary over time, without the need of manually setting up the devices. Dynamic resource sharing is based on the intuition that the same transport resource can be dynamically shared over time for different transport purposes. NFV provides flexibility by dynamically placing network functions in different locations depending on the specific need of a service, e.g., close to the users to exploit traffic locality. Examples of network functions that can be virtualized include for instance Evolved Packet Core (EPC) functionalities for local breakout, virtualization of packet aggregation capabilities, and virtualization of computing and storing functionalities, e.g., network caching.

These advanced functionalities required the presence of distributed (possibly small) Data Centre (DC). Therefore, the possibility to have low-power low-cost DC solution is crucial. In this respect optics can help again [21]. Small form factor optics for DCs use minimal electronics and trade performance for low cost. Long reach transceivers are available up to 100 Gb/s and 80 km reach. Continued progress in reducing the cost and form factor of these optics will be important for applications in rural and low-income areas. Integrated photonics is a promising approach to doing this [22]. In particular, silicon photonics integrates multiple photonic devices on silicon chips, which are often Complementary Metal Oxide Semiconductor (CMOS) compatible, with the potential for optical devices to share the same cost benefits of high-volume micro-electronics. Recently, single-chip Wavelength Division Multiplexing (WDM) chips were manufactured with 500 Gb/s aggregate data rate and as many as 1700 devices [23]. This Indium Phosphide (InP) based chip was designed for long-haul links. Intense research and development are currently underway for silicon photonic transceivers that can be manufactured in high volume and low cost [24].

2.3. Data Centres

Data centres are proliferating worldwide, but mostly in urban areas and developed countries, thus increasing the global digital divide.[2] The reason is that the conditions in rural and developing countries often exacerbate the challenges for operating data centres in addition to prohibitive acquisition costs.

A part from the previously discussed connectivity problems, energy provisioning raises several challenges. Specifically, in developing countries electrification is often not complete and especially rural areas lack access to the electric grid. To cope with energy variability and security, research on demand-response and emergency demand-response [25] tries to optimize the data centre operation based on the energy costs and availability. For this, they exploited the large body of research which went into power consumption flexibility at data centres via capacity right-sizing [26], load shifting over time [27] and across geographies [28].

To overcome the lack of grid access and to green the impact of data centres as well as to drive down the CAPEX costs, data centres are scaled down in size [29], powered via renewable energies [30] and rely on micro-servers with better work done per joule and work done per dollar ratios [31].

To reduce the energy consumption and OPEX costs, research tries to optimize the three major pillars in data centre: IT, cooling and power (see e.g., [33]). The upcoming 5G also promotes the use of softwarization and virtualization as means to drive up efficiency [34], especially via cloud-based radio access networks [35].

3. CHALLENGES

We first review the challenges related to the exploitation of 5G technologies iin rural and low-income zones the first step, and then we consider the socio-economic aspects that need also to be taken into account.

5G Technology Challenges The application of 5G in rural and low-income areas is challenged by the peculiar features of such scenarios. Tab. 1 reports a comparison between a classical 5G urban scenario [5] against 5G rural and low-income ones[3] (whose requirements may be inferred from [36]). In contrast to 5G urban's most advanced Internet services such as High-Definition (HD) streaming, tactile Internet and Internet of Things, many rural and low-income regions are still disconnected from the rest of the world and the lack of Internet coverage is the most critical problem to be tackled. The state of several essential services such as e-Learning and e-Health have remained substandard. Therefore, the infrastructure is required to support the

[2] Data centres requirements may be different than the ones of telecommunication networks, which are mainly based on central offices.

[3] Low-income areas include both low density regions in terms of populations as well as town and cities.

Table 1. Comparison of a classical 5G Urban Scenario with Rural and Low-income Ones

	5G Urban Scenario	5G Rural Scenario	5G Low-income Scenario
Service Type	HD Video, HD Streaming, Tactile Internet, IoT	HD Video, Emergency Service, e-Health, e-Learning	Basic Connectivity, Emergency Service, Delay Tolerant, e-Health, e-Learning
Network Constraints	Maximize Bandwidth, Minimize Delay, Coverage	Coverage, Guaranteed Bandwidth	Coverage
Energy Sources	Power Grid	Power Grid, Renewable Sources	Unreliable Power Grid and/or Renewable Sources
Network Cost from the User Side	Pay per bandwidth	Same as standard urban users	Low
Business Model	Return on Investment	Subsidized by the government	Subsidized by the government
Required Network Flexibility	High	High	High
User Mobility	Pedestrian, Vehicular, High Speed Vehicular	Pedestrian, Vehicular	Pedestrian, Low Speed Vehicular

appropriate set of applications to guarantee these essential services according to their different network constraints such as bandwidth requirements and link reliability. While for urban zones it is important to maximize the bandwidth and minimize the delay, in rural areas it is essential to guarantee a given minimum amount of bandwidth to users (e.g., the one required to deliver video services - but not high definition). Additionally, in rural and low-income zones it is even more important to guarantee basic coverage rather than high bandwidth services. In addition, another aspect that can not be neglected is that in rural and low-income zone the power grid may be not always available and/or unreliable, thus suggesting that renewable sources (such as the sun and the wind) should be exploited. Moreover, the cost of the network from the user side should be kept as low as possible, keeping also in mind that in low-income areas the users should pay much less for an Internet connection compared to the urban regions. This also inevitably influences the associated business models in such zones, which can not be based on the classical Return On Investment (ROI), but rather on the fact that the Internet is a primary need, that should be provided, e.g., by the government rather than private operators.[4] Moreover, the network has to be flexible in all scenarios. For example, in rural and low-income zone the network has, e.g., to deal with the scarcity of electricity, as well as to wisely manage the network resources in order to guarantee coverage. Finally, the user mobility has to be always taken into account. Specifically, rural and low-income zones are characterised by relatively lower mobility compared to urban ones.[5] This feature may also have an influence on the design of the network in such zones.

Socio-Economic Challenges In addition to these aspects, we would like to stress that networking alone is not sufficient to enable effective usage of its applications and services for a sustainable development. Other socio-economic challenges are also to be taken into account. First, affordability is to be considered in the networking development if the citizens want to use it effectively to achieve better living conditions. The cost of broadband connectivity is still higher than the average income in developing countries. Therefore, it can be a financial barrier for the poor citizens and communities from benefiting from networking. Second, it is also crucial to consider how relevant are applications and services provided. Their users, especially the low-income people, need applications that are necessary to their primary development needs such as the basic instructions they can understand, accessible through devices and services that they can afford and use conveniently. Third, human capacity is as critical as applications. Users need knowledge and skills to fully benefit from networking-enabled services, including ICT-based skills in the areas such as computer networking, web and basic applications development and elementary network security are essential in all societies. Governments and policy-makers need to understand the technical aspects of networking and their services, as well as the interplay between technological and public policy domains. Finally, the impact of the networking on environment is also critical. The networking can drive energy efficiency, smart systems and services to enable more productivities. However, networking is to be also a growing source of material consumption and greenhouse gas (GHG) emissions. This negative impact will increase as networking become more and more widespread.

4. OUR VISION

Given the aforementioned challenges, we first define the main pillars which, we believe, are essential for the design

[4]Private operators will not invest the cost of an infrastructure in such zones, due to the low return on investment. Therefore, the government should be either invest on the deployment of a minimum infrastructure or even becoming a network provider.

[5]In urban zones users may travel on high speed trains across the city. In rural zones users are more fixed, e.g., they are located close to their houses. In low-income zones users may be traveling by foot or by means of low-speed vehicles.

and management of 5G networks in rural and low-income zones. Our architecture is driven by the fact that the Internet connectivity is a primary need, that may be provided by the government or government-based entities rather than private ones. This alleviates the need of always guaranteeing a ROI. However, this process should be coupled with the reduction of the costs needed to deploy and manage the network. In the following, we sketch the proposed architecture.

4.1. Main Pillars

Converged Solution We believe that the services and the networks should be managed in a converged way, following a trend that is currently emerging in 5G architectures [37]. In our scenario, the last mile of the network should be orchestrated in conjunction with the metro and core one, without a complete separation between them, in contrast to current networks. Additionally, the services are not running on "top" of the network, but they are lying at the same level of the network components. More in depth, the network provider is also acting as service provider, and different computing components (like storage, local DC, caches) are installed on most network devices both in metro/core and access segments. As a results, there is not a strict separation between the different parts of the network, thus enabling a global optimization of the services and the network. This design choice is also justified by the fact that rural and low-income zones will be likely managed by government-based entities, which will then have the highest flexibility to deploy the services and the different components across the full network.

Reusability of Network Components We believe that another of the main pillars will be the possibility to completely virtualize most of the network components by means of virtual functions that are shared among devices and managed by a centralized entity. This trend is in line with current efforts in the broad area of softwarization, which aims to bring the network and computing functionalities from the hardware space to the software environment. Even though the performance of virtualized network elements is still lower than fully hardware-based components, this constraint is less stringent in rural and low-income zones compared to urban ones, thanks to the fact that the strict requirements in terms of high bandwidth and extremely low delays can be relaxed in such contexts. Therefore, by letting each element to be composed of a set of virtual elements, it is possible to reuse the same network components across different physical devices. This feature triggers the possibility to implement smart resource allocation policies in order to move the network and the computing capacity where it is really necessary or where the sources of energy are currently available. Moreover, with this capability the functionalities of network devices can be directly modified by software upgrades, without the need of changing the underlying hardware, which otherwise may be an expensive operation. Finally, thanks to the fact that the devices are split in software components, it is possible to integrate several functions in a single device, e.g., networking and computing functions.

Exploitation of Commodity Hardware In traditional networks, where most of functionalities are hardware-coded, there is a clear lack of flexibility. Each device (either a Base Station, a router, or a computing node) is composed of proprietary hardware, which can be hardly managed in a converged solution where all the network components and the services need to be controlled in a flexible way. In our vision, the network components are softwarized, and therefore it is possible to adopt general purpose hardware for most of devices. Apart from the flexibility aspects, the exploitation of commodity hardware has additionally two main advantages: i) the costs of deploying the network can be cheaper compared to high performance hardware-coded devices, and ii) since the hardware is the same for most of devices also the operating expenses can be reduced (e.g., in terms of failure costs). Clearly, the use of softwarized functions and commodity devices have an impact on the performance (e.g., bandwidth and delay), which should be always taken into account.

Solar-Powered Energy-Efficient Devices Since the power grid is assumed to be not so widespread in rural zones and not present or unreliable in low-income zones, a clear pulse will be the large exploitation of renewable energies to power the physical devices. Among the available renewable sources, one of the most promising is the sun. Moreover, since most low-income zones are located in the regions of the earth receiving the largest irradiation from this energy source, we expect that a large exploitation of solar panels will be a viable approach. However, even though the solar power can be predicted over time (thanks to weather forecasts), it is obviously not always available (e.g., during night or bad weather conditions). In addition, the size of the solar panels is still an issue for the installation costs. Therefore, the elements of a network have to carefully manage their energy consumption, by implementing, e.g., smart energy saving policies to reduce the amount of power requested. These policies will integrate both the knowledge of the power available and the variation of traffic experienced in the network over time.

Unmanned Aerial Vehicles and Advanced Radio Techniques In order to limit the costs for installing and managing the devices, the number of nodes devoted to the last segment of the network should be reduced as much as possible. Focusing on wireless elements, we foresee the exploitation of different technologies. First of all, thanks to the recent advances in Unmanned Aerial Vehicles (UAVs), the radio elements may be mounted on top of them. For example, it would be possible to move the radio nodes in order to follow the (few) users in a rural network, thus preventing them from the lack of connectivity. In addition, radio elements can be deployed at higher altitudes, in order to exploit two different features: i) the winds that can move such devices in the atmosphere, and ii) the fact that most users will experience Line Of Sight (LOS) conditions, which will decrease the propagation loss compared to classical Non Line of Sight (NLOS) conditions. In case of rural zones in which the number of users is not so low compared to the previous case, the emerging technology of massive antenna arrays allows

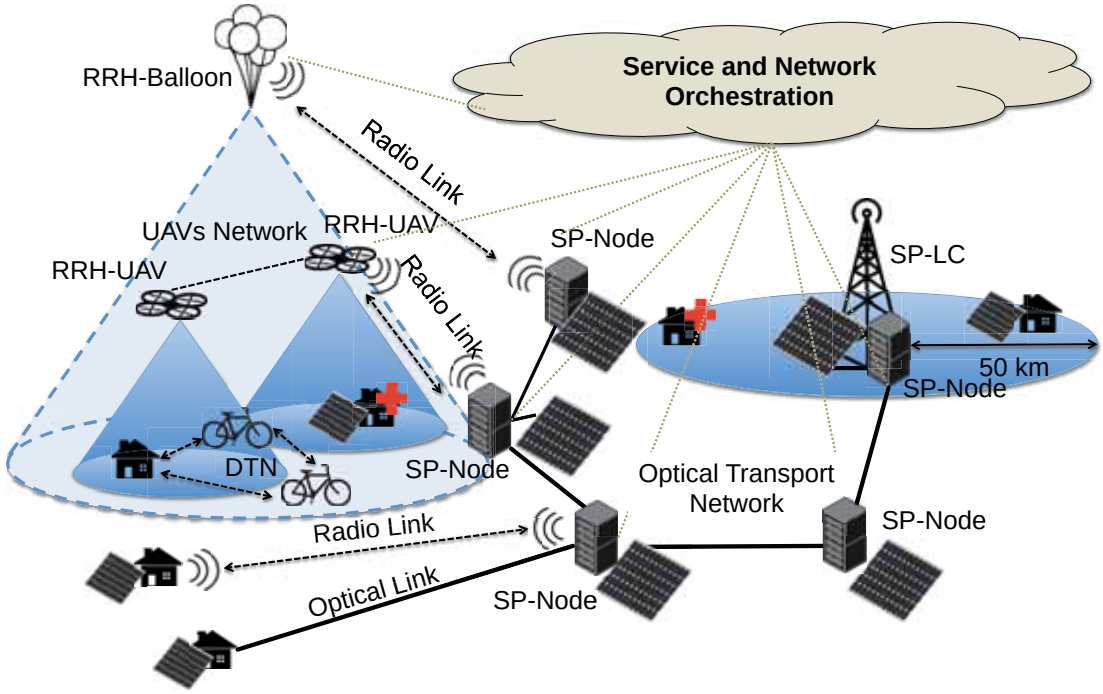


Fig. 1. Vision of a 5G network for rural and low-income zones. (SP = solar powered, LC = Large Cell, RRH = Remote Radio Head, UAV = Unmanned Aerial Vehicle, DTN = Delay Tolerant Network, NODE = Flexible component that can act as micro server, BBU, SDN switch and optical router).

the deployment of radio elements covering ultra-large cell sizes (over 50 km). It is widely acknowledged in academia and industry that deployment of very large antenna arrays at the base station side will form a key component of the 5G radio standards that very recently have started to be addressed by 3GPP. Technologies in academia referred to as massive MIMO are in the heart of the urban, capacity-driven 5G use cases and scenarios. Moreover, the energy-efficiency of these very large antenna arrays is also promising for the noise-limited, low-load rural and remote coverage scenarios.[6] The large arrays can accomplish a beamforming gain that will beneficially affect the link budgets of large macro-cells. The pencil-sharp radiobeams produced by these base stations allow to increase the cell radius without compromising implementation complexity, cost-efficiency or energy consumption. Finally, we foresee the exploitation of new ultra-lean radio-protocols, which will be developed with the explicit goal in reducing the transmission overhead of the radio nodes and improving their energy efficiency [38].

4.2. Overall Architecture

Our vision aims to develop a comprehensive low-cost connectivity architecture that can efficiently support a wide-range of services and applications. Fig. 1 reports the proposed architecture. Focusing first on the access part of the network, we foresee the exploitation of different technologies. Specifically, for rural zones with a limited number of users, coverage and capacity may be provided by RRH mounted on top of UAV (RRH-UAV). These devices can provide flexible coverage over a territory, by considering only the zones where the users are located. Moreover, RRH-UAVs can be exploited when the capacity of the network is needed (e.g., during the day). Each RRH-UAV establish communication with the other RRH-UAVs flying in the same zone. The goal is then to compute the UAV trajectories in order to optimize the coverage of the zone where the users are located. In addition, the RRH-UAV will establish a radio link with a BBU mounted in selected Solar-Powered (SP) Nodes [7]. The challenge here will be to develop smart solutions to reduce the amount of information exchanged between the RRH-UAVs and the SP-Node with the BBU. Notice also that he UAVs can be recharged by power stations fed by solar panels. In addition to this, RRHs are also mounted on top of balloons. These elements are continuously flying in the atmosphere in order to provide basic coverage and emergency services. Finally, each community connected to such an infrastructure may develop Delay Tolerant Networks (DTNs) to further spread the information by means of low-velocity vehicles (i.e., mainly bicycles). As second alternative to provide wireless access connectivity, we foresee the exploitation of Large Cells (LCs), with coverage radius in the order of 50 km. Such cells can be spread in low-income areas where the users requirements are low in terms of bandwidth and delays. LCs are also powered by Solar Panels (SPs), since the power

[6] We refer the reader to the Facebook project ARIES: `https://code.facebook.com/posts/1072680049445290/`, `lastaccessedon1stJuly2016`.

[7] This solution has to be evaluated w.r.t. the LTE UL-HARQ latency constraint.

grid is assumed to be not present or unreliable. The main issue in this scenario will be to deploy efficient solutions to limit the amount of required power in the uplink. Finally, we foresee the exploitation of direct optical connections and dedicated radio links to selected locations (e.g., places where the bandwidth requirements are higher).

Looking then at the other network solutions, we foresee the exploitation of flexible, efficient, low-cost, low-power nodes (SP-Nodes in the figure). Such devices, which will be solar powered, will virtualize different functionalities, including radio, computing, and transport ones. Each functionality can be activated/deactivated depending on where the node is located (in the transport part of the network or in the access one) and when it is needed. Moreover, the interconnection between the SP-Nodes will be realized by means of low-cost and low-power optical connections, which will implement the state-of-the-art features to increase the efficiency of the entire architecture.

Finally, the architecture will be controlled by a centralized orchestrator, which will jointly manage the network and the computing resources. For example, during period of high traffic the computing resources will be moved in the SP-Nodes close to users, while the opposite will be realized during low traffic periods. At the same time, coordination of UAVs, as well as decrease/increase of LC coverage will be done in accordance to the users variation over time. Finally, the orchestrator will perform this allocation of resources also taking into account the variation of power available from the SP.

5. CONCLUSIONS AND FUTURE WORK

We have focused on the problem of providing 5G services in rural and low-income areas. After a deep dive into the current state-of-the-art, we considered the main challenges that need to be faced for a full exploitation of 5G in such areas. In order to achieve this goal, we discussed a number of architectural features, including: the adoption of a converged solution, the reusability of network components, the exploitation of commodity hardware, the deployment of solar powered energy-efficient devices, UAVs and advanced radio techniques. Additionally, we proposed a reference architecture.

As future work, we plan a number of research activities. First of all, a detailed characterization of requirements of users is mandatory in both rural and low-income zone. This task would then drive the definition of more detailed technology solutions targeted to each specific context. In addition, we plan to study the radio issues that may emerge. For example, a large cell may provide coverage over a vast area. However, the uplink budget may be constrained by the user power and therefore there is an issue when the terminal is far from the macro cell, since a high amount of uplink power would be required. Moreover, different functional splits aiming to reduce the amount of data transferred between the RRH and the BBU will be also investigated. In addition, we plan to tackle the problem of computing the optimal UAVs trajectory to serve a set of users in a rural area. Finally, we plan also to analyse the proposed solutions in terms of CAPEX and OPEX, as well as properly dimensioning the solar panels needed to run the networks and services.

6. ACKNOWLEDGMENTS

This work has received funding from the project EU Horizon 2020 Superfluidity (grant agreement No. 671566) and the Polish National Science Center (decision DEC-2014/12/S/ST7/00415). We would like to thank the anonymous reviewers for their suggestions.

7. REFERENCES

[1] "The future of the Internet is at risk say global web experts," http://www.ecs.soton.ac.uk/news/4890.

[2] "International Telecommunication Union (ITU), ICT Facts and Figures the world in 2015," http://www.itu.int/en/ITU-D/Statistics/Pages/facts/.

[3] "World internet penetration rates," http://www.internetworldstats.com/stats.htm.

[4] Mats Eriksson and Jaap Van de Beek, "Is Anyone Out There? 5G, Rural Coverage, and the Next One Billion," in *Comsoc Technology News*, 2015.

[5] "5G-PPP White Papers," https://5g-ppp.eu/white-papers/.

[6] Arturo Atzcorra, "Trend in EU Research for 5G Networks," https://ans.disi.unitn.it/inw2015/presentations/s4-1-azcorra-5g.pdf.

[7] Cheng-Xiang Wang, Fourat Haider, Xiqi Gao, Xiao-Hu You, Yang Yang, Dongfeng Yuan, Hadi M Aggoune, Harald Haas, Simon Fletcher, and Erol Hepsaydir, "Cellular architecture and key technologies for 5G wireless communication networks," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 122–130, 2014.

[8] Anna-Maria Kovacs, "Regulation in financial translation: Will the incentive auction increase mobile-broadband competition in rural america?," Tech. Rep., Georgetown University McDonough School of Business, May 2014, http://www.gcbpp.org/files/EPV/Kovacs_Spectrum_Auctions_Rural_5.1.2014.pdf?utm_source=Anna-Maria+EPV&utm_campaign=EVP+5%2F1&utm_medium=email.

[9] Paal Frenger, Magnus Olsson, and Erik Eriksson, "Radio network energy performance of massive mimo beamforming systems," in *Proceedings of the IEEE 25th Annual International Symposium on Personal, Indoor, and Mobile Radio Communication (PIMRC)*. Sept. 2014, pp. 1289–1293, IEEE.

[10] "Google Loon Project," https://www.google.com/loon/.

[11] "Facebook Internet Project," https://info.internet.org/.

[12] Sobia Jangsher, Haojie Zhou, Victor O.K. Li, and Ka-Cheong Leung, "Joint allocation of resource blocks, power, and energy-harvesting relays in cellular networks," Tech. Rep. 3, Mar. 2015.

[13] Vinay Chamola and Biplab Sikdar, "Solar powered cellular base stations: current scenario, issues and proposed solutions," *IEEE Communications Magazine*, vol. 54, no. 5, pp. 108–114, 2016.

[14] Marco Ajmone Marsan, Giuseppina Bucalo, Alfonso Di Caro, Michela Meo, and Yi Zhang, "Towards zero grid electricity networking: Powering bss with renewable energy sources," in *2013 IEEE international conference on communications workshops (ICC)*. IEEE, 2013, pp. 596–601.

[15] Ziaul Hasan, Hamidreza Boostanimehr, and Vijay K Bhargava, "Green cellular networks: A survey, some research issues and challenges," *IEEE Communications surveys & tutorials*, vol. 13, no. 4, pp. 524–540, 2011.

[16] Peter Rost, Carlos J Bernardos, Antonio De Domenico, Marco Di Girolamo, Massinissa Lalam, Andreas Maeder, Dario Sabella, and Dirk Wübben, "Cloud technologies for flexible 5g radio access networks," *IEEE Communications Magazine*, vol. 52, no. 5, pp. 68–76, 2014.

[17] Daniel C. Kilper and Houman Rastegarfar, "Energy challenges in optical access and aggregation networks," *Philosophical Transactions of The Royal Society A*, vol. 374, no. 2062, Mar. 2016.

[18] Seamas McGettrick, Frank Slyne, Nattapong Kitsuwan, David B. Payne, and Marco Ruffini, "Experimental end-to-end demonstration of shared N:1 dual homed protection in long reach PON and SDN-controlled core," in *Proc. OFC, Los Angeles, USA*, Mar. 2015.

[19] X. Dong, T. El-Gorashi, and J. M. H. Elmirghani, "IP over WDM networks employing renewable energy sources," *IEEE/OSA JLT*, vol. 29, no. 1, pp. 3–14, Jan. 2011.

[20] P. Ohlen, B. Skubic, A. Rostami, M. Fiorani, P. Monti, Z. Ghebretensae, J. Martensson, K. Wang, and L. Wosinska, "Data plane and control architectures for 5g transport networks," *Journal of Lightwave Technology*, vol. 34, no. 6, pp. 1501–1508, March 2016.

[21] Jonas Weiss, Roger Dangel, Jens Hofrichter, Folkert Horst, Daniel Jubin, Norbert Meier, Antonio La Porta, and Bert Jan Offrein, "Optical interconnects for disaggregated resources in future datacenters," in *Proc. ECOC, Cannes, France*, Sep. 2014.

[22] Martijn J. R. Heck, Jared F. Bauters, Michael L. Davenport, Jonathan K. Doylend, Siddharth Jain, Géza Kurczveil, Sudharsanan Srinivasan, Yongbo Tang, and John E. Bowers, "Hybrid silicon photonic integrated circuit technology," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 19, no. 4, pp. 6100117–6100117, July/Aug. 2013.

[23] J. Summers, T. Vallaitis, P. Evans, M. Ziari, P. Studenkov, M. Fisher, J. Sena, A. James, S. Corzine, D. Pavinski, J. Ou-Yang, M. Missey, D. Gold, W. Williams, M. Lai, D. Welch, and F. Kish, "Monolithic inp-based coherent transmitter photonic integrated circuit with 2.25 tbit/s capacity," *IET Electronics Letters*, vol. 50, no. 16, pp. 1150–1152, July. 2014.

[24] Andy Eu-Jin Lim, Junfeng Song, Qing Fang, Chao Li, X.iaoguang Tu, Ning Duan, Kok Kiong Chen, Roger Poh-Cher Tern, and Tsung-Yang Liow, "Review of silicon photonics foundry efforts," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 20, no. 4, pp. 405–416, July 2014.

[25] Cheng Wang, Neda Nasiriani, George Kesidis, Bhuvan Urgaonkar, Qian Wang, Lydia Y. Chen, Aayush Gupta, and Robert Birke, "Recouping energy costs from cloud tenants: Tenant demand response aware pricing design," in *ACM e-Energy*, 2015, pp. 141–150.

[26] Chuangang Ren, Di Wang, Bhuvan Urgaonkar, and Anand Sivasubramaniam, "Carbon-aware energy capacity planning for datacenters," in *IEEE MASCOTS*, 2012, pp. 391–400.

[27] Zhenhua Liu, Adam Wierman, Yuan Chen, Benjamin Razon, and Niangjun Chen, "Data center demand response: Avoiding the coincident peak via workload shifting and local generation," *Performance Evaluation*, vol. 70, no. 10, pp. 770–791, 2013.

[28] Zhenhua Liu, Minghong Lin, Adam Wierman, Steven H. Low, and Lachlan L. H. Andrew, "Greening geographical load balancing," *IEEE/ACM Trans. Netw.*, vol. 23, no. 2, pp. 657–671, 2015.

[29] James R. Hamilton, "An architecture for modular data centers," in *CIDR*, 2007, pp. 306–313.

[30] Iñigo Goiri, William A. Katsak, Kien Le, Thu D. Nguyen, and Ricardo Bianchini, "Parasol and greenswitch: managing datacenters powered by renewable energy," in *ASPLOS*, 2013, pp. 51–64.

[31] Ronald P. Luijten, Dae Pham, Rolf Clauberg, Matteo Cossale, Huy N. Nguyen, and Mihir Pandya, "4.4 energy-efficient microserver based on a 12-core 1.8ghz 188k-coremark 28nm bulk CMOS 64b soc for big-data applications with 159gb/s/l memory bandwidth system density," in *IEEE ISSCC*, 2015, pp. 1–3.

[32] Vasileios Lakafosis, Sreenivas Addagatla, Christian Belady, and Suyash Sinha, "Prometheus: A wirelessly interconnected, pico-datacenter framework for the developing world," in *WWIC*, 2012, pp. 26–39.

[33] Yijun Ying, Robert Birke, Cheng Wang, Lydia Y. Chen, and Natarajan Gautam, "Optimizing energy, locality and priority in a mapreduce cluster," in *IEEE ICAC*, 2015, pp. 21–30.

[34] Chih-Lin I, Corbett Rowell, Shuangfeng Han, Zhikun Xu, Gang Li, and Zhengang Pan, "Toward green and soft: a 5G perspective," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 66–73, 2014.

[35] Aleksandra Checko, Henrik L Christiansen, Ying Yan, Lara Scolari, Georgios Kardaras, Michael S Berger, and Lars Dittmann, "Cloud RAN for mobile networks: a technology overview," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 1, pp. 405–426, 2015.

[36] Matteo Fiorani, Björn Skubic, Jonas Mårtensson, Luca Valcarenghi, Piero Castoldi, Lena Wosinska, and Paolo Monti, "On the design of 5g transport networks," *Photonic Network Communications*, vol. 30, no. 3, pp. 403–415, 2015.

[37] G. Bianchi, E. Biton, N. Blefari-Melazzi, I. Borges, L. Chiaraviglio, P. de la Cruz Ramos, P. Eardley, F. Fontes, M. J. McGrath, L. Natarianni, D. Niculescu, C. Parada, M. Popovici, V. Riccobene, S. Salsano, B. Sayadi, J. Thomson, C. Tselios, and G. Tsolis, "Superfluidity: a Flexible Functional Architecture for 5G Networks," *Transactions on Emerging Telecommunication Technologies*, on press 2016.

[38] Sibel Tombaz, Pal Frenger, Fredrik Athley, Eliane Semaan, Claes Tidestav, and Anders Furuskar, "Energy performance of 5G-NX wireless access utilizing massive beamforming and an ultra-lean system design," in *IEEE GLOBECOM*, December 2015, pp. 1–7.

DESIGN OF SCALABLE DIRECTORY SERVICE FOR FUTURE IOT APPLICATIONS

Ved P. Kafle, Yusuke Fukushima, Pedro Martinez-Julia, and Hiroaki Harai

National Institute of Information and Communications Technology, Tokyo, Japan

ABSTRACT

Unprecedentedly a massive number of devices are getting connected in the coming era of the Internet of Things (IoT). For discovery, remote access and management of these IoT devices, an IoT directory service is needed to store and provide their various attributes such as location, generated data types, owner's name, and security keys. In this paper, we present the architectural design of the IoT directory service that is capable to store a huge number of heterogeneous records and provide fast lookup (latency of few milliseconds) and dynamic update (latency of few seconds), while fully complying with owner- or user-centric security and privacy policy. To meet the performance requirements despite fluctuations in the workload and networking environment, we leverage tools of network function and resource virtualization to dynamically allocate and adjust the computational and network resources assigned to the directory service.

***Keywords*—**Internet of Things, IoT, directory service, M2M, standardization

1. INTRODUCTION

The Internet of Things (IoT) is envisioned to resolve several challenges of the modern society by collecting and processing detailed information about events and environments with the help of billions of newly connected devices or things. It makes human life safer, healthier, and more productive and comfortable. At the same time, it provides substantial business opportunities to various vertical industries and social sectors involved in the fields of automobile, energy and utilities, transport, logistics, smart cities, healthcare, fitness, sports, public safety, and so on. Consequently, the International Telecommunication Union - Telecommunication Standardization Sector (ITU-T) has recently established the Study Group 20 for the exclusive study and standardization of IoT technologies and applications including smart cities and communities [1]. A paper presented in the ITU Kaleidoscope conference 2015 provides a good review of ITU-T's IoT related activities [2].

The IoT is considered as the major contributor to 50 billion devices expected to get connected to the next generation (known as 5G) mobile networks in 2020 and beyond [3]. It would also change the communication paradigm from human-centric to machine-centric by introducing a massive number of machine-to-machine (M2M) communication devices that operate without involving humans in the communication loop [4]. However, before billions of new devices or things, besides the conventional computers, smartphones, and tablets get connected properly, there are many technical challenges we need to resolve. One of such challenges is that IoT applications require a scalable, performance-guaranteed directory service that stores and provides information about the IoT devices such as name/ID, location, owner's name, generated data types (e.g. temperature, air pollution and illuminance levels measured by sensors), security keys, and credentials. IoT applications will retrieve the information, also called records, stored in the IoT directory service to discover the devices and securely obtain data and services provided by them. The IoT directory service will also play an essential role in providing IoT device managers with information necessary for discovery, management, and remote configuration of IoT devices.

In this paper, we present the architectural design of an IoT directory service that stores a huge number of heterogeneous records (about one billion) and provides fast lookup (latency of few milliseconds) and dynamic update (latency of few seconds) of records, while fully complying with the device owner- or user-centric privacy policy. To meet the aforementioned performance requirements despite fluctuation in the workload and network environment conditions, our proposal instantiates on-demand caches and replicas, and leverages the network function and resource virtualization tools to dynamically adjust computational and network resources allocated to them.

The current Internet's Domain Name System (DNS) and related schemes are not adequate for storing records related with IoT devices. The DNS was not designed to store heterogeneous types of names and resolve them into various records within bounded-time, not to mention its limited support for updating such records dynamically and the slow propagation of the updated records. Previous related studies, such as [5-7], have proposed few schemes of dynamic name services. However, they also have some limitations. For example, Auspice [5] does not consider the record privacy and access control. Multi-level Distributed Hash Table (MDHT) [6] does not consider frequent updates of the records (as it assumes only 1% records change per day), and CoDoNS [7] limits its proposal to fast lookup and resilience to attacks through proactive caching, but updates are not in its focus. They also do not consider the issues of privacy. Similarly, the directory services specified in Recommendation ITU-T X.500 [8] are capable to store heterogeneous types of names and resolve them into

various types of attributes. However, they are not designed to minimize the response time.

We aim at meeting mainly the following two performance requirements of the IoT directory service: (1) being capable of storing a huge number of records (about 1 billion records), and (2) being capable to provide fast resolution of names into related attributes (latency of few milliseconds). We require to keep the lookup latency as low as possible because, accordingly to the ITU Report on Tactile Internet [9], the IoT applications involving human senses (e.g. sight, audio, or touch) require the system response within the order of magnitude of 1 ms to 1 s. Our initial target is 10 ms to enable those applications achieve their response times without confronting the latency added by current name resolution systems.

This research has two parts: (1) architectural design of record storage, replication, lookup, and update functions, and (2) dynamic allocation of computational and network resources to the directory service so that the performance requirements are always met despite fluctuation in the workload. However, because of space limitation, this paper presents only the first part, i.e. the architectural design, and gives only a very brief outline of the resource allocation procedure.

The remainder of the paper is organized as follows. Section 2 reviews the DNS performance as a related technology and provides a list of the design approaches used in the proposed IoT directory service. Section 3 presents the architecture detail. Section 4 describes the record replication, lookup, update, and resource allocation procedures. Section 5 concludes the paper by outlining the future work items.

2. ANALYSIS OF RELATED WORK AND DESIGN APPROACHES

In this section, we first present a review of DNS lookup performance by carrying out a simple experiment as related work, and then list up the approaches we have used in the design of the proposed IoT directory service.

We stored 10,000 records in *BIND 9* software in a computer and measured the lookup latencies by sending queries at various rates by using *DNSPerf* from the same computer. The average lookup latencies were 0.10 ms, 0.15 ms and 2 ms, respectively, for 1,000 queries/s, 10,000 queries/s and 100,000 queries/s. We also measured the update latency by using *nsupdate* command, which was on average 6 ms. These lookup and update latencies do not include any network latency as both the resolver and server were residing in the same computer. To assess the impact of network latencies on the lookup latencies, we ran a simple experiment using Google *namebench* tool from a computer connected to our lab network. We scanned 11 DNS servers located in the surrounding networks (at 3-4 hops distance) and measured the average value of name resolution latencies of each server by sending 250 queries from a Google Chrome resolver program. We found that the lowest value was of 35 ms from the fastest server, while the highest value was around 280 ms from the slowest server. These values also include the propagation latency, and we can infer that the DNS servers are not currently placed optimally.

From the above observation, it is clear that the network latency significantly dominates the DNS lookup latency. Since the DNS BIND database implementation has small query latency, we may leverage it for storing IoT records (of course, with some modification to store heterogeneous names and expedite update process) or other NoSQL databases. However, to reduce the network latency and maintain high performance despite fluctuation in workload and network conditions, we need to apply new design approaches as listed below.

(1) IoT service-wise directory services: To meet the distinct performance requirements of various services, we consider that each IoT application will have its own directory service. For example, automated/safe driving and smart-grid control applications will have two different directory services.

(2) On-demand, trackable record caches: To achieve low latency lookup by reducing the communication distance between a client (i.e. querying node) and the directory server, caches are created and maintained on-demand in replica servers located closer to the client. The caches are trackable by the authorized directory so that any update of dynamic records in the authorized database can be reflected in all cache replicas instantly by a fast update process.

(3) Dynamic resource provisioning: To meet the IoT directory service performance requirements despite fluctuation in the workload and network conditions, the appropriate amount of computational and network resources allocated to the replica servers is adjusted dynamically. Thus, we consider managed networks that have the capability to provision the resources dynamically.

(4) **Leveraging established component technologies:** To improve its deployability qualities, our approach relies on established tools for database management, software-defined networking (SDN), network function virtualization (NFV), and virtual machine (VM) to allocate and adjust resources for storage, computation, and communication. Although the VM setup and usage may incur additional latency, VM provides an easier and effective tool for segmentation, assignment, control and configuration of resources. We expect that this approach would be helpful for standardization and incremental deployment of IoT application services in collaboration with business partners such as equipment vendors and network service providers.

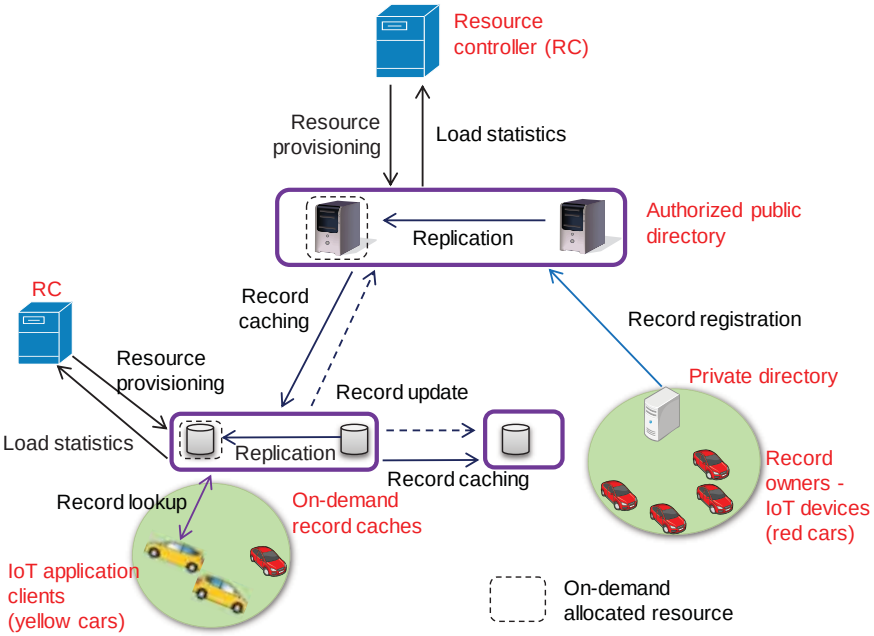


Figure 1. Components and processes of the proposed IoT directory service.

3. IOT DIRECTORY SERVICE ARCHITECTURE

Figure 1 shows the components and processes involved in the proposed IoT directory service. The major components (shown by boxes in the figure) are: (1) Authorized public directory, (2) On-demand record caches, (3) Resource controller, (4) Private directory, (5) Record owners, and (6) Record clients. Similarly, the processes involved (shown by arrows in the figure) are: (a) Record registration, (b) Record caching, (c) Record replication, (d) Record lookup, (e) Record update, (f) Load statistics monitoring, and (g) On-demand resource provisioning. The components and processes are described below.

3.1. Architectural Components

(1) **Authorized public directory:** It contains the required database management functions to store, lookup, update, and delete records, and replication functions to copy the whole database, or some records, to designated replica servers. It is called "authorized public directory" because it stores records of authorized IoT devices or things only and provides the records to authorized clients that are eligible to obtain them (i.e. each record contains its access policy). Each record in the database is composed of various attributes (e.g. ID, location, data types, security key and certificate) linked to the IoT device name. The names may be of heterogeneous types, and can be written in different formats (e.g. text, binary). However, they are unique in the database so that each record is uniquely identified by the name. To reduce the workload and lookup latency of the authorized public directory, the replication function creates a copy of specified records and stores them in the designated replica servers when a command for doing so is received from the resource controller. In addition, if some replica servers are no longer in use due to lack of queries for the copies of records stored in them, these copies are deleted and resources allocated to them are returned to the resource pool.

(2) **On-demand record caches:** They store copies of records distributed from the authorized public directory to the network edges closer to clients or IoT applications. They accept record lookup requests from IoT applications and provide relevant records in the response. They are tracked by the authorized public directory, i.e. the authorized public directory stores the reachability information of all record caches in a list and uses the list to access them when any change in the authorized public directory records needs to be reflected immediately in the trackable caches as well. Similarly, to reduce the record update latency, some cache servers, called anchor cache server, are delegated by the authorized directory to accept and process record update requests. Similar to the authorized public directory, the cache can also make its replicas to reduce its load and thus record access latency. Moreover, any anchor cache server can distribute records to other replica servers created by the resource controller as the record owner moves from one network to the other. This process is regulated by the distribution policies provided by record owners and initiated by the authorized public directory, so all caches are able to provide "authoritative" answers to client queries, which can be totally trusted as long as the records (or their cache copies) have not expired.

(3) **Resource controller:** It allocates, adjusts, and adapts resources of the authorized directory, cache servers as well as of their replica servers on the basis of regular monitoring of their performance and load statistics.

The resources can be hardware (e.g. CPU cycles, memory, disk space, and network bandwidth) as well as software modules (e.g. network protocols, configuration parameters, software commands, adaptation tools). It obtains the status of hardware resources (i.e. CPU load statics, available memory size and storage size) as well as the status of database (e.g. access latency, number of records, access frequency, number of replicas, and trackable caches holding a record) by monitoring them periodically, and then decides if more resources have to be assigned to create new replicas for distributing workloads, or assigned resources to existing servers must be adjusted to meet the changing workload.

(4) **Private directory:** It provides a proxy gateway function to register records of IoT devices in the public directory. It is the authority that determines the privacy level of each record belonging to the IoT devices it manages.

(5) **Record owners:** They are IoT device owners who possess the records and are the final authority to assign the privacy level to each attribute of the records.

(6) **Record clients:** They are the IoT applications that send queries to cache servers to obtain the desired records of IoT devices.

3.2. Processes

Each of the below processes involves both processing inside a component and interaction between two or more components of the IoT directory service.

(a) Record registration: A record is created when an IoT device/object finishes its initial configuration and gets connected to the network. Either the device/object itself or its proxy gateway creates the record. The record includes various attributes: name, ID, location, data types it generates (e.g. temperature, pressure, and video) security keys, certificates, privacy levels, etc. The record is firstly stored in the private directory (e.g. home directory, personal directory) for local use only. The private directory assigns a privacy level to each attribute of the record, e.g. location or address is visible to anybody, but data types are visible only to those who share the same shared key. After setting up the privacy levels, the private directory registers the records in the authorized public directory. The privacy/identity of the record is well-protected by encrypting messages exchanged between the private directory and authorized public directory servers, e.g. using transport-layer security (TLS) [10]. The registration process is not as time-critical as the lookup process.

(b) On-demand record caching: The authorized public directory provides records to on-demand caches after receiving a *cache command* from the resource controller. The resource controller uses several logics (described in the next section) to decide about the potential cache locations and prepares cache servers. Each cache copy has a timeout value, and is deleted on timeout.

(c) Replication: The authorized directory or on-demand cache can replicate the records stored in it to replica servers upon receiving a *replication command* from the resource controller. The replication of records may take place in different granularity (i.e. a single record or a group of records). The authorized directory and cache servers keep a list of all replica servers storing records copied from them. Similarly, replica servers also keep a list of replica origins from where the records have been replicated (also known as anchor points).

(d) Record lookup: When a client needs the record of an IoT device, it sends a *record lookup* request to a nearly cache server securely and receives a response message containing the record. We assume that the client obtains info about the address of the cache server at the time of its attachment to the network (e.g. like DHCP providing a DNS server address in the Internet) and proactively establishes a security association using an existing security mechanism, such as Datagram Transport Layer Security (DTLS) [11], which has been adopted by the IETF to provide a security layer to the Constrained Application Protocol (CoAP), a specialized web transfer protocol having potential to be used in IoT application.

(e) Record update: The record update process is initiated either from the record owner device that changes its parameters stored in the directory or from the network (e.g. proxy gateway) that detects the change in attributes, e.g. a mobile IoT device's address when it is moving from one network to another. The latter approach is better suited for resource-constrained IoT devices because they are not required to involve in the update process, thus helping in the reduction of signaling overhead and power consumption.

(f) Load statistics monitoring: This process runs in all components to monitor two types of loads: database load and system load. It then provides the load statistics to the resource controller. The database load statistics include information about the number of records in database, record lookup and update frequencies, and corresponding latencies. Similarly, the system load statistics includes information about the usages of CPU, memory, bandwidth, number of replica servers, number of cache servers, etc.

(g) **Resource provisioning:** This process runs in the resource controller to allocate and adjust hardware and software resources on-demand. It implements a resource allocation algorithm that takes the system and database load statistics as input parameters and makes the resource allocation, adjustment or adaptation decision.

3.3. Security and Privacy Protection

Privacy and security issues are key aspects of any network architecture. They are more important for IoT applications that deal with delicate and highly private data. Therefore, we have considered both issues from the design phase of the IoT directory service, and it adjusts to the different levels of security and privacy requirements of different IoT

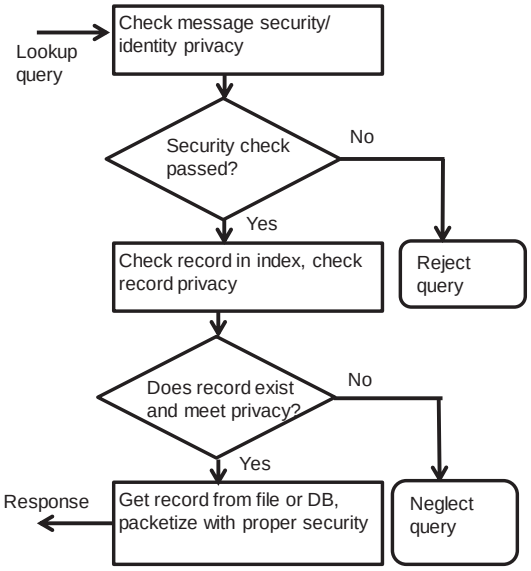


Figure 2. Flowchart of security and privacy scheme.

devices. For instance, some environment sensing IoT devices allow everyone to obtain from the IoT directory their address and types of sensing data they provide, while they hide the owner's name. Similarly, some IoT devices might only allow the authorized clients (e.g. belonging to the same owner or located in the same locality) to know their address and data types.

Figure 2 shows a flowchart of the security and privacy enforcement scheme. The security and privacy checks are carried out in two steps: (1) message security and identity privacy check, and (2) record privacy check. In the first step, it is checked if the lookup query message itself has been secured and privacy protected at the level satisfying the basic requirements, e.g. message integrity protected or name/identity privacy is protected by encryption. In the second step, it is checked if the querying client is authorized to get the requested record by satisfying the privacy policy. Only the lookup queries that pass through both checks are answered with the records. Privacy policies can be applied to full records or part of them.

4. RECORD CACHING, REPLICATION, LOOKUP, UPDATE, AND RESOURCE ADJUSTMENT

In this section, we describe the mechanisms for caching, replication, lookup, and update of the records with the objective to meet the target performance requirements.

4.1. Record Caching Locations

The proposed IoT directory service establishes record caches in locations closer to the potential clients. Such locations are determined by using application-dependent heuristic methods. For instance, record caches are instantiated in the location where the record owner (e.g. a car) has currently moved to and where the possibility that an IoT application (e.g. running on the other cars) from the surrounding making a lookup query for the record is high.

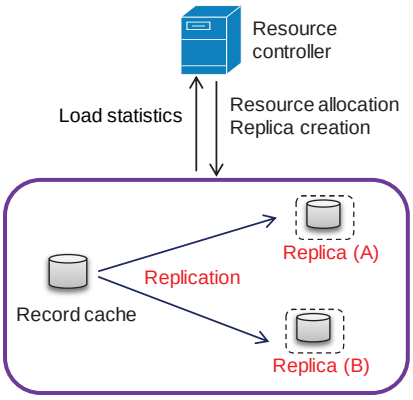


Figure 3. Record replication

4.2. Record Replication

When the volume of records in the authorized directory or a cache server is large, the frequency of lookup queries becomes high, and the monitored metrics indicate that the performance starts degrading and about to miss the target levels, the resource controller prepares a replica server and asks the authorized directory or cache server to move some records to the new replica server. For example, in Figure 3, two replicas A and B are created where records from the cache server are copied.

4.3. Record Lookup

As mentioned earlier, to get a desired record of an IoT device, a client IoT application sends a lookup query to a nearby record cache server, as shown in Figure 4. The cache server verifies if the lookup query message passes through the first step of security and privacy check (as described in Section 3.3). The cache server then checks its record index to determine if the record exists and which replica server holds it. It then forwards the query to the corresponding replica server, which executes the second security and privacy check to determine if the client is authorized to receive the requested record or parts of it. In case the query passes through the record privacy check, the replica server retrieves the record from its database, configures a response message with the record, and transmits it to the client.

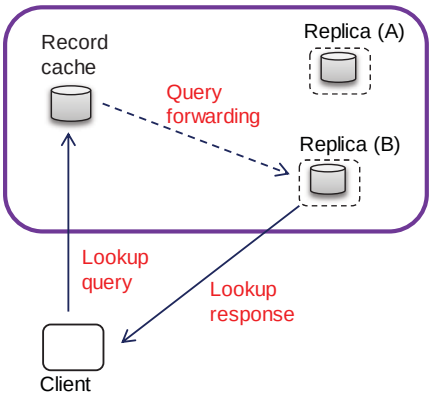


Figure 4. Record lookup

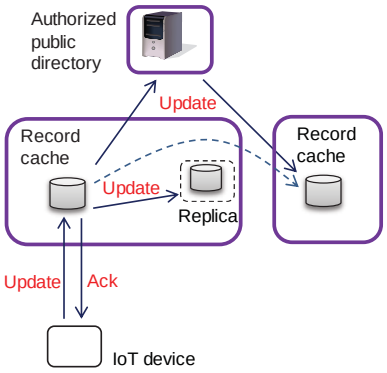


Figure 5. Record update

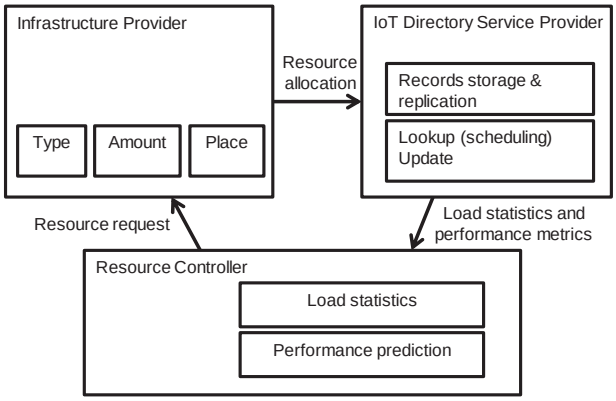


Figure 6. Resource allocation

4.4. Record Update

As mentioned earlier, an update request originates either from the record owner IoT device that changes its parameters due to an event (e.g. changing location due to mobility, adding a sensor to the IoT device to generate a new type of data, or changing certificates and keys for security enhancement) or from the network (e.g. proxy gateway) that monitors the new parameters being assigned to or associated with the IoT device. As shown in Figure 5, the update request is sent to the anchor cache server to update its local cache copy. The cache server verifies the message authenticity, updates the record stored in it (and its replicas) to minimize the time of record inconsistency in the locality, and forwards the update request to the authorized public directory. If the record has been stored in more than one cache servers, the authorized public directory server updates all the caches (known from the trackable cache list).

Alternatively, if the other cache servers have obtained the record copy from the anchor cache server, before forwarding the update request to the authorized directory, the anchor cache server forwards the update request to the other cache servers (as shown by a dotted arrow in the figure). To further minimize the signaling overhead, the update requests flow in bundle (containing update information of several IoT devices, e.g. in the event of group mobility) between the authorized public directory and cache servers.

As update times are crucial, our proposal reduces the number of elements involved to the minimum. At the same time, it forwards any update to the first component that knows where a copy of the record has been cached. This simplifies the process of maintaining cache coherency, and thus improves the overall performance of the system.

4.5. Resource Allocation

Figure 6 shows the resource allocation and adjustment process involving the infrastructure provider, IoT directory service provider, and resource controller. As mentioned earlier, the resource controller regularly obtains load statistics and performance metrics from the IoT directory service system and estimates required amount of resources to achieve the target performance levels. Accordingly, it sends a resource request containing the description of required resources to the infrastructure provider. The infrastructure provider analyzes the resource request and reserves appropriate types and amount of resources in appropriate places. The resources are allocated to the IoT directory service provider. This close loop operation among the resource controller, infrastructure provider and IoT directory service provider guarantees that the performance is always met in spite of the workload and network condition fluctuations.

5. CONCLUSION

We presented the architectural design of a novel IoT directory service, which is capable to store and provide dynamic records of a huge number of IoT devices. We described the components and processes involved in the proposed IoT directory service architecture and outlined the scheme of security and privacy protection and on-demand adjustment of virtualized computational and network resources allocated to the IoT directory service so that its performance is maintained even when the workload fluctuates.

In future work, we will design the resource allocation algorithm for the resource controller, implement the IoT directory service in a testbed network, and study its performance in detail by varying the workload and network capacity. We will also gradually bring this research outcome to ITU for standardization.

REFERENCES

[1] ITU-T SG20: IoT and its applications including smart cities and communities (SC&C), http://itu.int/ITU-T/go/sg20.

[2] V.P. Kafle, Y. Fukushima, and H. Harai, “Connecting the world through trustable Internet of Things,” ITU Kaleidoscope Conference, 2015.

[3] A. Osseiran et al., "Scenarios for 5G mobile and wireless communications: The vision of the METIS project," IEEE Commun. Mag., Vol. 52, No. 5, pp. 26-35, May 2015.
[4] Recommendation ITU-R M.2083-0 (2015) "IMT Vision – Framework and overall objectives of the future development of IMT for 2020 and beyond".
[5] A. Sharma et al., "A global name service for a highly mobile internetwork," ACM SIGCOMM 2014.
[6] M. D'Ambrosio et al., "MDHT: A hierarchical name resolution service for information-centric networks," ACM ICN 2011.
[7] V. Ramasubramanian and E.G. Sirer, "The design and implementation of a next generation name service for the Internet," ACM SIGCOMM 2004.
[8] Recommendation ITU-T X.500 (10/2012), Information technology – Open Systems Interconnection – The Directory: Overview of concepts, models and services.
[9] The Tactile Internet, ITU-T Technology Watch Report, August 2014.
[10] T. Dierks and E. Rescorla, "The transport layer security (TLS) protocol version 1.2," RFC 5246, IETF, August 2008.
[11] E. Rescorla and N. Modadugu, "Datagram transport layer security version 1.2," RFC 6347, IETF, January 2012.

A STACK4THINGS-BASED PLATFORM FOR MOBILE CROWDSENSING SERVICES

S. Distefano

Kazan Federal University
Higher Institute for
Information Technology and Information Systems
Kazan, Russia
Email: sdistefano@kpfu.ru

*A. Puliafito, G. Merlino, F. Longo, D. Bruneo**

Università di Messina
Department of Engineering
Messina, Italy
Email: {apuliafito,gmerlino,
flongo,dbruneo}@unime.it

ABSTRACT

As mobiles grow pervasive in people's lives and expand their reach, Mobile CrowdSensing (MCS) and similar paradigms are going to play an ever more prominent role. There is a pressing need then to ease developers and service providers in embracing the opportunity, and that means offering a platform for such efforts. This in turn means providing a solid foundational architecture with abstractions and sound layering for MCS application designs to be mapped over it. This should base on a flexible infrastructure able to provide resources to MCS applications according to their requirements, hopefully on-demand. A service-oriented/Cloud model can perfectly fill this gap. This paper is a first step in this direction, proposing to adopt Stack4Things (S4T), an OpenStack-based platform for managing sensing and IoT nodes, for runtime customization of resources and their functions to support MCS services and applications. This implies developing and extending the S4T platform further to the specific requirements coming from off-the-shelf, e.g., Android-based, mobiles, as well as describing an example S4T-powered MCS application, Pothole Detection Mapping, to highlight the role of the platform.

***Keywords*—** Mobile crowdsensing, Cloud, IoT, OpenStack, Android.

1. INTRODUCTION AND MOTIVATIONS

Mobile CrowdSensing (MCS) comprises by definition a category of applications where individuals carrying sensor-hosting embedded computers (e.g. smartphones) get collectively engaged in information gathering and sharing efforts to analyze and georeference events which may be interesting for individuals and communities alike. One of the main advantages of MCS is the possibility to conduct sample collection, data mining, etc., without accounting for the corresponding experiments in advance, just leveraging natural daily life patterns arising from human activities as they happen and leave behind breadcrumbs in form of samplings ready to be collected.

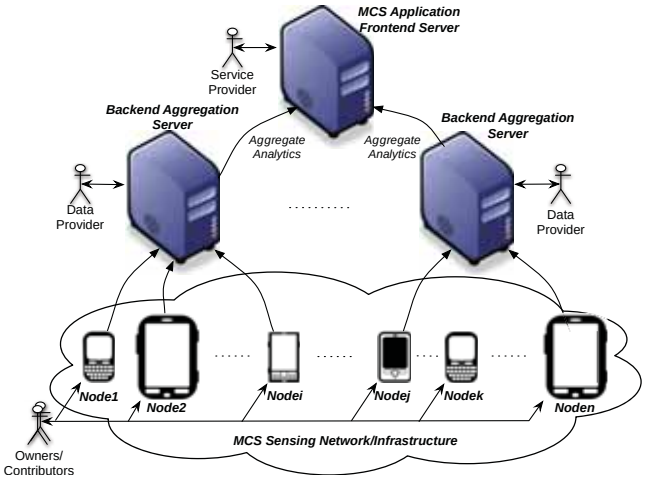


Figure 1. The MCS reference scenario.

Typical MCS applications mainly implement a client-server interaction pattern where a *service provider* offers MCS-based services to *end users*, leveraging *contributors* willingness to provide their physical (sensing) resources. Data are therefore collected and processed by (backend and frontend) servers to carry out aggregate analytics and feed back relevant results to end users. Starting from the lowest level, through heuristics and algorithm design, local analytics may provide a category of functions, among which simple ones are interpolation, extrapolation and outlier filtering, which may enhance a standard MCS application as shown in Figure 1. A few drawbacks of such siloed pattern limiting the potential of the underlying paradigm are: i) unoptimized runtime, as multiple applications would execute on the same nodes without taking into account such configuration, possibly duplicating sensing or processing activities on resource-constrained devices, thus also limiting scalability of the platform; ii) necessity of enrolling, collecting and managing resources, i.e., lack of a coordinated sens-

*This research effort was supported by the EU 7th Framework Programme under Grant Agreement n. 610802 for the "CloudWave" project and by the EU Horizon 2020 Research and Innovation Program under the Grant Agreement n. 644048 for the "BEACON" project, and has been carried out in the framework of the CINI Smart Cities National Lab.

ing resource management. In particular this last point is crucial, as some mechanisms and facilities for dealing with sensing resource management are of strategic importance to enable the actual potential of the MCS paradigm. Furthermore, specific patterns able to support the MCS paradigm by, for example, implementing coordinated, self-managed cooperation mechanisms among nodes could boost MCS to a new-hot trend in the IoT scenario, as a feasible solution for things management and related applications and services development.

This paper moves towards this direction, proposing Stack4Things [1] (S4T), a framework for the management of distributed smart objects in the IoT context through a service oriented, on-demand, cloud provisioning model, for the management of the MCS sensing infrastructure. This goes further beyond current IoT-Cloud trends, mixing the two paradigms for the management of (mainly data) resources in the cyber-physical space, towards a brand new, crowd-driven approach to IoT and related app development. Specifically, starting from the S4T infrastructure-oriented framework, the architecture is adapted and extended in the following to support MCS, mainly providing node enrolment, customization, networking and control facilities in a service-oriented/Cloud fashion. Then an example of an MCS application taking advantage of the functionalities exposed by the platform is discussed.

2. MCS AND IOT PARADIGMS

MCS as a paradigm embraces many approaches to crowd-sourcing sensor data, including both participatory and opportunistic sensing. On one hand participatory sensing [2] may be defined as any crowd-sourced sensing activity where each member of the crowd is actively involved, giving feedback when asked or otherwise tagging measurements on a voluntary basis. This is to be contrasted to an opportunistic perspective [3], where sensing is essentially unmanned: MCS would tap into mobile devices just because people carry those around in their pocket all day long anyway, and may just be involved once with a fire-and-forget experience, leaving then all crowd-sensing activities to unassisted background processes.

As devices are carried around by individuals, owners may eventually be in the (data feeding) loop. Their mobility and situational awareness may be leveraged, in an opportunistic and participatory fashion respectively, to support the collection of finer grained information and semantically tagged data. For MCS applications to succeed, there have to be appropriate incentive mechanisms to recruit, engage and retain human participants. In this sense, a centralized credit system, assigning and managing credits and rewards, is usually adopted as incentive mechanism in a participatory strategy. On the other hand, in opportunistic scenarios, the gamification approach is usually adopted, building up a credit collection race among contributors to incentive their participation.

As a more subtle differentiation of the two approaches, we may consider how they diverge in reference to the actors benefiting from the crowd-sourcing. It would follow that participatory may be considered any MCS, and services that may derive from it, when a community, or the public at large, is the one entity taking mostly (and primarily) advantage, e.g., [4, 5, 6]. Whereas opportunistic would be MCS where any outcome would eminently center around single individuals. The latter is usually linked to measurements of individual phenomena, i.e., where samples are to be traced back to individuals producing them, usually featuring other individuals as a way to augment processing on otherwise purely personal trails. Conversely, community sensing revolves around naturally anonymized aggregation of data.

Table 1. Taxonomy table of MCS applications.

MCS **categorized (by)**	**Approach**	
	Participatory	*Opportunistic*
Owner involvement	Active, human-assisted sensing / tagging	Background, unmanned data collection
User benefit	Public interest	Individual utility
Fruition modality	Pull / non-contextual	Push / contextual
Interaction model	Centralized (client-server)	Distributed (mesh)
Incentive mechanism	Credit systems (bank)	Credit collection race

While still focused on the end-user perspective, another aspect to be considered is how information produced by an MCS system is to be consumed, or made relevant to the situation under which fruition would occur. A typically proactive, participatory pattern for users may consist in merely consulting an MCS-derived knowledge base, thus leveraging information as-is, i.e., non-contextually and in a pull fashion. Conversely an opportunistic fruition mode would be based around push-based notifications, depending on certain inferred metrics on the (dynamic) environment surrounding the user, thus contextual in nature. Context itself may be exploited to dynamically allocate sensing tasks to the best subset of participants [7], or other metrics may be combined and evaluated to rank participants for such kind of allocation, e.g., measuring credible interactions among participants [8]. Moreover, also in terms of interactions, at least first-time enrollment requires input on the side of owner, employing a client-server model. Yet even opportunistic schemes, featuring distributed behavior and cooperative strategies, may be considered, dependent on the underlying topology, as is the case for mesh-like ones in device-dense environments.

A synthesis of the approaches and categories of MCS applications is presented in Table 1, in particular with reference to the multifaceted definition both kind of approaches embed, where each row represents a degree of freedom with respect to this dichotomy. This way, a wide range of possibilities for MCS application paradigms, from pure participatory to wholly opportunistic ones, may be identified, including also hybrid solutions horizontally spanning one or more axes.

MCS can be also considered as an Internet of Things (IoT)-related paradigm. Indeed, the whole IoT research community agrees on the notion that things are to be interconnected over some potentially global network (possibly, but not exclusively the Internet), to be exploited for whichever scenario, and in pa rticular by specific applications and services. One distinguishing feature is the autonomous, chatty nature of

such mesh. As now happens with applications made up of distributed components overlaid on the network, interacting by means of well-established protocols and paradigms, e.g. RESTful services, also things should network themselves and their corresponding agents, with no (direct) operator intervention (think M2M). Under such premise, MCS may just become a pattern under the IoT umbrella term, i.e. a specialization of the platform that an IoT would represent for sensing-related, mobility-enabled, crowd-sourced use cases. This perspective should be particularly appreciated in light of opportunistic developments, as attaching semantic description to things, endowed with distributed, event-triggered logic, may really provide chances for nodes to discover each other services and weave an even more powerful abstraction with respect to the Web of Things.

Although IoT can be mainly associated with the participatory pattern, some work on opportunistic IoT and sensing environment is available in literature. For example, in [9] an opportunistic IoT framework is proposed, mainly extending opportunistic networking towards participatory sensing, enabling opportunistic information sharing among things to also support mobile social networking. Similarly, opportunistic mobile networking is the topic of [10], mainly focusing on low level data forwarding issues through a framework able to support and optimize opportunistic sensing. The best way to exploit the seemingly limitless IoT potential for sensing at a global scale is by coupling opportunistic and participatory application patterns for mobile crowds with the infrastructure underneath. The proposed solution is aimed at horizontally enabling *things* for massive development and deployment of any sensing task and based on a novel application of Cloud computing technologies to IoT, as described in the following.

As we pointed out already, not only should we optimize any MCS despite shaky grounds, but we aim to turn most shortcomings to our advantage. More specifically, uneven global connectivity when coupled with dense, mesh-like topologies lends itself naturally to be addressed by means of routing protocols as those designed for MANETs. This leads to avoiding otherwise likely network congestion and servers' overload while also making room for any kind of crowd-local, opportunistic and cooperative behavior.

In particular when compared to the building blocks of MANETs par excellence such as, e.g., WSNs, nowadays consumer-class mobile devices feature plenty of computing, storage and communication capabilities, making these platforms many orders of magnitude more advanced than mote-like ones, also in terms of sensing, considering that mobiles are usually multi-modal by default, thus enabling a wider range of applications. Moreover, the impact of massive deployments in the field can never be overestimated, as these devices are available by the billions, always-on and mostly online, a consistent slice of people's daily activities and recurring habits. Leveraging such huge populations means building potentially instant-on large-scale applications while at the same time lowering expenses and time-to-market, also avoiding altogether the long setup times and high upfront costs involved in case of ad-hoc sensing infrastructure, when not absolutely required.

Dynamic behavior and chances for data reuse are also differentiating factors in evaluating MCS against WSN-like configurations. When dealing with MCS, the composition of device populations, as well as the kind and quality of data produced, where quality may be expressed both in terms of latency and precision, may change in time due to mobility patterns, power requirements and communication subsystems, including owner-mandated local preferences. In traditional WSNs the populations (and the kind of data produced) are usually known a priori, thus managing quality and designing according to requirements is more straightforward. Moreover, with regard to MCS, support for multiple concurrent applications is feasible, albeit subject to careful planning and design stages. When it comes to sensor networks instead, deployments are typically geared for a specific application, thus repurposing or resource sharing are rarely accounted for.

Still casting challenging constraints into opportunities, evolutions in time of crowds' shapes and composition may lead to fast dissemination of information, aiding in the pursuit of (global) optimization objectives. Moreover we care about actual MCS usefulness, for contributors themselves too, thus any solution can't be considered absolutely all-round if lacking built-in feedback and assisted guidance. Even with regard to data itself, both in terms of raw format and application requirements, sensor readings may not be suitable for direct consumption. In this sense specific local analytics may be performed on-board, to (pre)process raw data, in order to obtain intermediate results ready for transmission and further processing.

Even not taking into account the scalability issues any kind of backend may be subject to when dealing with huge raw data uploads, on-board resources in mobile devices are on the rise and already remarkable in many cases, thus it seems natural to tap into such potential to shift in part the burden of computations toward the edge. Moreover, a renown tradeoff in conventional mote-class WSNs lies in leveraging (local) computation to save on the amount of energy and bandwidth, as needed to transmit preprocessed data in comparison to raw readings. Last but not least, delay-sensitive applications may benefit from such savings, if the overhead of local processing is negligible when compared to transmission-induced latencies.

3. A SERVICE-ORIENTED MCS INFRASTRUCTURE

The Stack4Things framework [1, 11], also referred to as S4T, is our effort to pursue a new approach for the management of smart objects and things in the IoT scenario, following an on-demand, service oriented provisioning model. This shifts the IoT paradigm towards the Cloud one, merging benefits of both: on the one hand providing control and management capabilities to IoT (sensing and actuation) resources, on the other enriching the Cloud paradigm with pervasive I/O capabilities to directly interact with the environment. Altogether

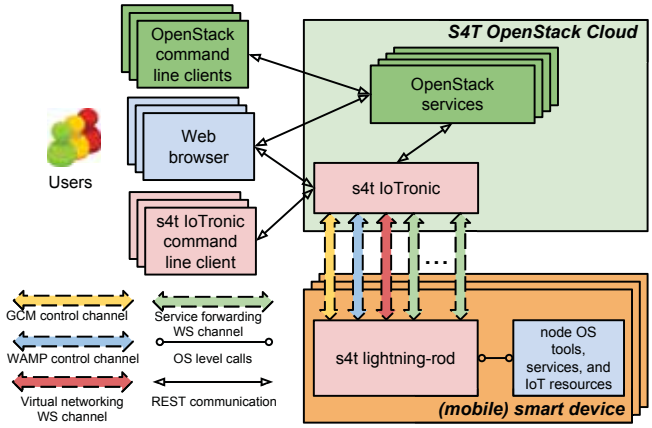


Figure 2. Stack4Things overall architecture for MCS.

these ideas and challenges have been framed into a novel utility paradigm for IoT [12], mainly highlighting, from a functional perspective, its focus, i.e., to establish a sensing-Cloud.

Key aspects of SAaaS paradigms are the i) device-centric philosophy [13] of providing actual (sensing and actuation) devices or things, even if virtualized, to higher levels (users, app developers or providers) and the ii) unified management and control of all the underlying resources. This implies to provide, on the one hand, mechanisms for customize and contextualize the resources, pushing intelligence to the extreme edge by injecting code on them and, on the other, a control surface able to manage these resources and provide them as a "utility", in a Cloud-oriented model. S4T design has thus been tackled by extending certain *OpenStack* subsystems, a well-known and widely adopted Cloud management framework, to smoothly integrate and leverage as much existing functionalities as possible.

Figure 2 shows the S4T overall architecture for MCS, derived from the one defined in [1] by specializing its modules in the MCS domain, in particular considering mobile devices. It highlights interactions and communication facilities between end users, the Cloud and mobile nodes hosting sensing (and, sometimes, even actuation) subsystems.

On the MCS node side, the *S4T lightning-rod* runs under the device-native (typically SDK-enabled) environment available for developers and interacts with a (subset of) OS services and tools available on the device for which authorizations have been granted, as well as with (physical or even virtual) sensing and actuation resources, through UNIX-style filesystem-based abstractions of the underlying interfaces, either GPIO for embedded boards or, typically, (SDK-specific) API-mediated for mobiles. As well, it connects to the Cloud and anchors the node to the (centralised) infrastructure, allowing the end users to manage node-hosted resources even when nodes are behind gateways or other network appliances implementing NAT, firewall rules, or any other restrictive policies. Among mechanisms enabling this flexibility, a key role is played by WebSocket-based tunneling, and either Google Cloud Messaging (GCM)- or WAMP-based messaging between the S4T lightning-rod and the *S4T IoTronic service*, the former being the point of contact with the Cloud infrastructure where the MCS application has to be deployed, the corresponding subsystem of which is the latter. *Web Application Messaging Protocol (WAMP)* [14], as a subprotocol, belongs to the IETF WebSocket (WS) standard, and defines some communication semantics for WS-transported messages, providing publish/subscribe (pub/sub) primitives, and remote procedure calls (RPC) as well, either simple or routed ones.

The S4T IoTronic service is designed as an OpenStack one, exposing the capability to manage one or more smart boards, remotely. This may be done either via a command-line interface, an *OpenStack CLI*, or a specific *S4T IoTronic CLI*, or even a Web browser though either set of REST APIs, as provided by the core OpenStack services and S4T IoTronic.

4. STACK4THINGS MCS PLATFORM

With respect to the node-side runtime, and specifically mobiles, in particular focusing on Android, there are the Sensing APIs and the environment-provided notification subsystem as key elements of the underlying platform at the basis of our design choices. The Sensing APIs are opaque in terms of our MCS platform, as any sensor tuning and reconfiguration requests are expected to be relayed to some environment-native subsystem managing this kind of requests, as the Sensing APIs in case of the Android ecosystem. The Android-native notification subsystem is useful to minimally involve provider-enabled Cloud-based mechanisms for push-based communication to devices, thus avoiding to track network conditions in general.

Google Cloud Messaging (GCM) [15], the most current Google-provided notification service, is leveraged for exchanging (bootstrapping) asynchronous messages in Android mobiles, an enabling step to support runtime customization mechanisms and other Cloud-initiated primitives, including on-demand activation of any WAMP-/WebSocket-based facilities, when not overly restricted in terms of limits on generated traffic. The choice of supporting the instantiation of WAMP-/WS-based channels, actually a custom communication bus, under a provider-supported mobile platform such as Android, stems from the inherent limitations (and costs) linked to the usage of (Android-native) GCM, which does not support significant payloads (e.g., file transfers), but is designed and marketed specifically for push notifications, and in our case employed also for bootstrap signaling.

Figure 3 shows the S4T MCS platform architecture in terms of the node-side components. The Sensing APIs provide the corresponding abstractions at the lowest level, upon which the node-side subsystem, called *S4T lightning-rod*, is built to arbitrate access to sensors/actuators, through a set of libraries for transducer virtualization, the *S4T virt libraries*, which provide interfaces for reading (and writing) sensors (and actuators, respectively), as well as settings parameters for devices' operating modes. Such operations are therefore available at the proper level of abstraction semantics. This means either locking or releasing resources in accordance to

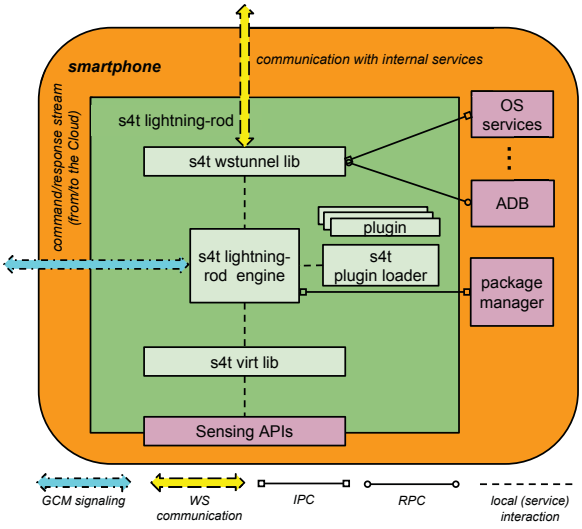


Figure 3. Stack4Things node-side MCS platform architecture.

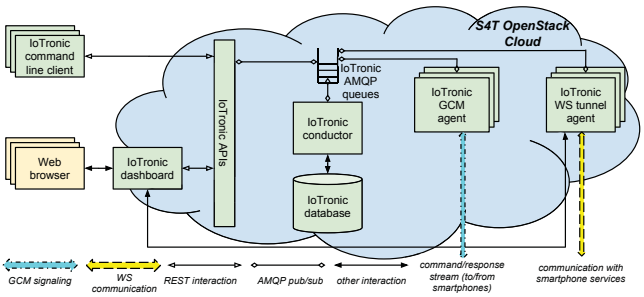


Figure 4. Stack4Things Cloud-side MCS platform architecture.

successful booking requests, and to constraints dependent on the APIs granularity, specific to the transducing resources.

The *S4T lightning-rod engine* sits at the core of the node-side software architecture, and communicates with the Cloud through a GCM-based messaging channel, at the very least, and, when needed, switching to a WebSocket-based full-duplex channel for WAMP messaging (see also Figure 4), in particular to support a wider range of functionalities, such as, e.g., sending data to (and receiving data from) the Cloud, or even executing commands input by the users via the Cloud, respectively. User-provided commands may be related to the interaction with the node-hosted resources (through the S4T virt libraries) and with operating system-level resources and tools (e.g., package manager, background services, filesystems). Bootstrapping communication with the Cloud is thus enabled by primitives for push-based messaging available through the platform-native SDK. Furthermore, WebSocket-based libraries (*S4T wstunnel libraries*) augment the engine to act as a WS-powered (reverse) tunneling server, connecting to a specific WS endpoint in the Cloud. This mechanism [16] enables internal services to be exposed through the tunnel for direct access by external users, whose incoming traffic is forwarded automatically to any internal service of choice running in the background. Outgoing traffic is captured and redirected into the tunnel and eventually reaches the end user that connects to the WS server in the Cloud, a user then ready, at last, to consume the node-hosted service.

The S4T lightning-rod subsystem also provides a *plugin loader*. By this component, custom plugins may be injected into the node environment from the Cloud, and loaded at runtime to expose specific (user-defined) primitives, including system-level interactions, albeit subject to sandboxing, such as, e.g., involving the system-native package manager, or any service automatically started at boot-time. The S4T Cloud-side MCS platform architecture (see Figure 4) features an OpenStack service newly designed for IoT, named *IoTronic*.

IoTronic is characterized by the standard architecture of an OpenStack service. The *IoTronic conductor* includes the core logic of the service, and manages the *IoTronic database*, which stores required information about IoT devices, e.g., unique identifiers of nodes, any ownership or delegation role nodes may obey to with respect to users and tenants, board properties and hardware/software characteristics. The conductor also dispatches RPCs among other components.

The *IoTronic APIs* exposes RESTful interfaces for end users. The OpenStack Horizon dashboard has been extended to expose all the functionalities provided by the IoTronic service and other S4T subsystems. The dashboard exposes also access to node-side services, relaying traffic from the user to the *IoTronic WS tunnel agent*. This agent acts as a controller for the WS endpoint to which nodes connect through wstunnel libraries.

Likewise, the *IoTronic GCM agent* works as a bridge for GCM-based communication between any node and other components. The command stream gets delivered by this agent during bootstrapping phases, as well as under situations of restrictions on traffic. It converts AMQP messages on the wire into GCM-compliant ones, and brokers from GCM to AMQP as well, conversely. In accordance to the current philosophy among OpenStack contributors, any communication occurring among Cloud-side subsystems, in this case IoTronic components, is implemented as pub/sub messaging over the network via AMQP queues. This enables the whole design to be as scalable as needed, as all components may be deployed on different hosts without affecting service behaviour. Also, multiple *tunnel agents* and *GCM/WAMP agents* may be instantiated, in that case each dealing with a subset of enrolled mobiles. Given all of the aforementioned options, high availability and redundancy may be guaranteed as well.

5. AN EXAMPLE

As an example of a S4T-powered MCS application, let us feature Smart City-like traffic service, specifically focusing on a Pothole Detection and Mapping (PDM). aiming at detecting and identifying road potholes through traveler mobiles. The PDM application is based on two components: an Android app, running on S4T-enabled mobiles and a central-

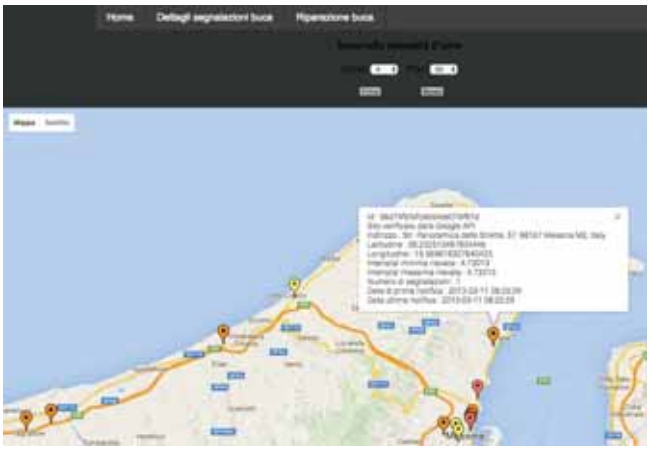
ized back-end collecting, filtering and preprocessing sensed data.

Aggregated data on geolocalized accelerometer samples from mobiles are used to create a first draft of the road map identifying potential potholes on the region of interest. This information is collected by the PDM app running in background on contributor devices to capture sudden changes on acceleration which could be related to a pothole. Such data are therefore locally evaluated and, in the case they match well-known (pothole) patterns, they are notified to the server side app together with their geospatial coordinates. Specifically, the measurements at the basis of the detection mechanisms are related to the accelerometer-provided acceleration vector, its (computed) euclidean norm and in particular the (impulsive) vertical (z-axis) displacement. Based on these input data, a Web-based application has been developed which, by interacting with the server-side database, implements preprocessing operations such as computing the differential values of the acceleration vector along the three axes of the frame of reference. This could be used to obtain further insights on the road segment condition highlighting issues in the roadway, which may be exploited by administrations to setup proper maintenance plan prioritizing on the extent of damage. To reduce the impact on mobile resources, data processing, mining and filtering are mainly moved to the Web server application, while just minimal pre-filtering operations, to prevent false positives, are implemented client side on mobiles.

A testbed composed of voluntarily contributing students and colleagues, going around Messina for almost a month, has been considered for the experiments. The results thus obtained are shown in Figure 5. In particular, Figure 5a depicts a screenshot of the app developed on Android devices, while Figure 5b shows the results collected by the above mentioned population in a month of experimentation, where the detected potholes are showed on a map.

The power of crowdsensing may be further appreciated by avoiding to centralize duties which may as well be performed client-side, actually just binding the processing task to the actor more intimately involved with the corresponding data, i.e., the contributor as data producer but also as volunteering compute node. Indeed this approach has been followed to actually fix imprecise locations of potholes. In particular, well-known limitations with precision in GPS positioning, typically on the order of a few meters at most, are not an issue when leveraged for navigation systems, due to the continuous tracking and the road maps as trails to follow. Yet, in the case of pothole detection, it is not trivial to identify with precision their positions, without resorting to huge amounts of (otherwise unneeded) data to be pushed and processed, to hook points of interest to specific segments, and ultimately infer the most plausible locations. If results are to be retained, albeit coarser-grained, a useful preliminary refinement may just consist in identifying the (shortest) segment among those originating from the (potential) pothole being detected, where those segments intersect neighboring roads perpendicularly. Fortunately there are already cloud-

(a) Android app



(b) Web app

Figure 5. PDM app screenshots

enabled service-oriented interfaces, like the Google Roads APIs, providing exactly this kind of information, i.e., nearest point, with respect to a location, belonging to a road. Considering the scale at which this kind of system could potentially operate, here crowdsensing comes to the rescue not only by leveraging mobiles as sources of interesting data, but also by offloading Cloud-side S4T subsystems, i.e., letting the mobile app itself query the APIs to infer road-compatible locations, before submitting two (related) datapoints, the raw (i.e., sampled) coordinates (for historical and analytics purposes only) together with the refined coordinates, the latter immediately ready to be leveraged for end-user visualisation on the dashboard.

In particular, this is done in our example by injecting and instantiating at runtime the corresponding code, tasked with querying the Roads APIs, as plugin through the aforementioned plugin loader. We can thus notice that our Cloud-enabled management of IoT infrastructure, coupled with a

plugin-based runtime customization system, enables us to leverage the crowd for sensing-related activities, augmenting results by offloading preliminary validation steps of the measurements to the crowd, just leaving long running tasks (e.g., analytics, storage) to the Cloud by default. This option turns out to be especially useful when time/space-local context, only available mobile-side, may be leveraged to preprocess results to be sent, as mentioned above. In particular, the preprocessing may involve other measurements to be sampled opportunistically, which would not overburden the Cloud anyway as this context would be discarded immediately after sending the (refined) sample to the Cloud. In this case, context may be other geocoordinates to be sampled for a brief interval after the main datapoint, in order to assess the confidence interval of the sampled coordinates, before querying the Roads API with a (possibly adjusted) position, to be hooked to the nearest road segment.

6. CONCLUSIONS

In this paper we explored MobileCrowdSensing (MCS) as a paradigm, especially in relation to IoT and edge computing, and adapted the Stack4Things (S4T) framework to account for basic mechanisms needed to enable MCS. This work in particular features a taxonomy and model of MCS, then presents S4T as an Infrastructure-as-a-Service framework for IoT and mobiles, amenable to be employed as a service-oriented platform for MCS applications, with specific focus on Android devices, in particular being able to instantiate and deploy at runtime custom code, useful to offload computation to the edge when deemed useful. These offloading capabilities have been put to use through an MCS application for mobiles, meant for crowdsourcing the mapping of road surface distress conditions, highlighting the potential of the envisioned approach.

REFERENCES

[1] F. Longo, D. Bruneo, S. Distefano, G. Merlino, and A. Puliafito, "Stack4Things: a Sensing-and-Actuation-as-a-Service framework for IoT and Cloud integration," *Annals of Telecommunications*, pp. 1–18, 2016.

[2] N. Lane, E. Miluzzo, H. Lu, D. Peebles, T. Choudhury, and A. Campbell, "A survey of mobile phone sensing," *IEEE Comm. Mag.*, vol. 48, no. 9, pp. 140–150, 2010.

[3] J. Burke, D. Estrin, M. Hansen, A. Parker, N. Ramanathan, S. Reddy, and M. B. Srivastava, "Participatory sensing," in *Workshop on World-Sensor-Web (WSW'06): Mobile Device Centric Sensor Net. and App.*, 2006, pp. 117–134.

[4] M. Haklay and P. Weber, "Openstreetmap: User-generated street maps," *Pervasive Computing, IEEE*, vol. 7, no. 4, pp. 12–18, Oct 2008.

[5] Y. Chon, N. D. Lane, F. Li, H. Cha, and F. Zhao, "Automatically characterizing places with opportunistic crowdsensing using smartphones," in *Proc. of the 2012 ACM Conf. on Ubiquitous Computing*, ser. UbiComp '12. New York, NY, USA: ACM, 2012, pp. 481–490.

[6] S. B. Eisenman, E. Miluzzo, N. D. Lane, R. A. Peterson, G.-S. Ahn, and A. T. Campbell, "Bikenet: A mobile sensing system for cyclist experience mapping," *ACM Trans. Sen. Netw.*, vol. 6, no. 1, pp. 1–39, Jan. 2010.

[7] A. Hassani, P. D. Haghighi, and P. P. Jayaraman, "Context-aware recruitment scheme for opportunistic mobile crowdsensing," in *Parallel and Distributed Systems (ICPADS), 2015 IEEE 21st International Conference on*, Dec 2015, pp. 266–273.

[8] J. An, X. Gui, Z. Wang, J. Yang, and X. He, "A crowdsourcing assignment model based on mobile crowd sensing in the internet of things," *IEEE Internet of Things Journal*, vol. 2, no. 5, pp. 358–369, Oct 2015.

[9] B. Guo, D. Zhang, Z. Wang, Z. Yu, and X. Zhou, "Opportunistic iot: Exploring the harmonious interaction between human and the internet of things," *Journal of Network and Computer Applications*, vol. 36, no. 6, pp. 1531 – 1539, 2013.

[10] D. Zhao, H. Ma, S. Tang, and X.-Y. Li, "COUPON: A Cooperative Framework for Building Sensing Maps in Mobile Opportunistic Networks," *Parallel and Distributed Sys., IEEE Trans. on*, vol. 26, no. 2, pp. 392–402, Feb 2015.

[11] F. Longo, D. Bruneo, S. Distefano, G. Merlino, and A. Puliafito, "Stack4Things: an OpenStack-based framework for IoT," in *Future Internet of Things and Cloud (FiCloud), 2015 International Conference on*, Aug 2015, pp. –.

[12] S. Distefano, G. Merlino, and A. Puliafito, "A utility paradigm for IoT: the Sensing Cloud," *Pervasive and Mobile Computing*, no. 0, pp. –, 2014.

[13] ——, "Device-centric Sensing: an alternative to Data-centric approaches," *IEEE Systems Journal*, 2015.

[14] T. Oberstein and A. Goedde, "The Web Application Messaging Protocol," Working Draft, IETF Secretariat, Internet-Draft draft-oberstet-hybi-tavendo-wamp-02, October 2015. [Online]. Available: http://www.ietf.org/internet-drafts/draft-oberstet-hybi-tavendo-wamp-02.txt

[15] "Google Cloud Messaging for Android." [Online]. Available: http://developer.android.com/google/gcm/index.html

[16] G. Merlino, D. Bruneo, F. Longo, S. Distefano, and A. Puliafito, "Cloud-based Network Virtualization: an IoT use case," in *Ad Hoc Networks*, ser. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, N. Mitton, M. E. Kantarci, A. Gallais, and S. Papavassiliou, Eds. Springer International Publishing, 2015, vol. 155, pp. 199–210.

A POPULARITY BASED CACHING STRATEGY FOR THE FUTURE INTERNET

Suhaidi Hassan, Ikram Ud Din, Adib Habbal, and Nur Haryani Zakaria

InterNetWorks Research Laboratory, School of Computing, Universiti Utara Malaysia

ABSTRACT

Information-Centric Networking (ICN) is an attractive network model receiving increasing consideration by the research community because of its inspiring features. To better manage the Internet usage move from host-centric communication to receiver-driven content retrieval, revolutionary ICN architectures have been proposed. A distinguished characteristic of these innovative architectures is to provide ubiquitous and transparent in-network caching to enhance network resource utilization and accelerate content dissemination. With the exponential increase of Internet traffic, the issue of content storage is a growing concern in ICN. In this paper, we present a caching strategy that considerably increases cache hit rate and reduces stretch ratio, which are the most important metrics in the evaluation of ICN caching. Through extensive simulations, it is shown that our proposed work is a favorable and realistic contribution for the standardization exercise of data caching for achieving accurate and valid network performance in the future Internet.

Keywords— Future Internet, ICN, CPCE, caching, content popularity

1. INTRODUCTION

Toward the start of the Internet, clients were scholastic in nature, for the most part inspired by mail trade and document exchanges [1]. Moreover, sharing of resources was a vital issue that forced significant difficulties with respect to correspondence among end hosts [2]. However, since the last decade the Internet popularity has brought about the activity on the Internet to become significantly reliable [3]. Information sharing and dissemination, for example, scholastic, social, and business, over the Internet is the major cause of the Internet growth. Distribution of named data is a noteworthy application in the existing Internet. Along with online content dissemination, other distribution technologies, for instance, Content Distribution Network (CDN) and Peer-to-Peer (P2P) communication have been well developed and are advancing the communication framework for getting contents by name, regardless of the location of the main server [3,4]. Keeping in mind the end goal to react to expanding activity volume in the existing Internet for applications are utilized that use caching and content distribution and replication in various particular ways [1].

This research is supported by the Ministry of Higher Education Malaysia Research Grant Scheme with reference number FRGS/UUM/13258.

It is therefore important to deduce that the existing Internet architecture can be considered as a restricting factor of the current Internet growth and the design of new applications. To add more, several studies, for example, [5, 6] have shown that the advancement of the Internet design is driven by incremental and responsive increases and therefore change in the architecture is indispensable. In this manner, globally the research community is working towards understanding the architectural challenges to decide the rule that will drive the future Internet design. Numerous research projects (e.g., US NSF GENI [7], EU-FIA [8], and AsiaFI [9]) have been funded in the last few years to define the existing limitations and future needs for the Internet architecture, and therefore Information-Centric Networking (ICN) [1, 10–12]] is one of the considerable results of these research activities. ICN has different modules, such as naming, routing, mobility, security, and caching. However, the majority of our research community is attracted by the caching module because of the limited cache size of network nodes. In ICN caching, the network nodes have the ability to cache a content locally once it is downloaded by the end users. Therefore, if new requests arrive for the same content, they are satisfied locally rather than contacting the original server.

2. PROBLEM DEFINITION

In ICN, the contents are cached locally by network nodes (e.g., routers). These routers may pose strict constraints with respect to cache management. Thus management becomes a critical issue for content caching and caching strategies. When the network becomes stable and the router's cache overflows, a replacement policy, such as Least Recently Used (LRU), Least Frequently Used (LFU), or Random policy is used to evict one of the cached contents to make room for the new arrived one. However, besides the replacement policy, content caching is the main issue, i.e., which content should be cached and at which location it needs to be stored so that to efficiently utilize memory and bandwidth consumption. For that, many strategies have been proposed, for example, Cache Everything Everywhere (CEE) [2] - the default ICN strategy, Cache Less for More [13] – caches contents at a node which has the maximum betweenness centrality, Probabilistic Caching (ProbCache) [14] – stores contents near the users, Cooperative In-network Caching (CIC) [15] - where the content is divided into different chunks and cached at more than one node, Cache Aware Target Identification (CATT) [16, 17]- where the content is cached at a single node on the publisher-

subscriber path, Optimal Cache Placement based on Content Popularity (OCPCP) [18] - where the popularity of incoming content is calculated on the basis of cached contents and the new content is cached based on its popularity value, Network Coding based Cache Management (NCCM) [19] - jointly considers content routing and caching strategy through Linear Network Coding (LNC), WAVE [20] - caches contents based on their access count, Most Popular Caching (MPC) [21] – caches contents based on their popularity values, Dynamic Fine-Grained Popularity-based Caching (D-FGPC) [22] – the modification of MPC, however, unlike MPC in FGPC the threshold value for content caching is changed dynamically based on content popularity values. Moreover, some experimental assessment of cache management in ICN are presented in [23]. Some other research on caching has been proposed in [24–31].

All of the mentioned strategies investigated only one aspect of the caching (i.e., either content placement or replacement), and none of them covers both aspects. Therefore, a flexible caching strategy is needed so that to place the contents at the best possible position and (on the arrival of a new content) replace one of the cached contents. Actually, the contents are cached in the random access memory (RAM) while the available static random access memory (SRAM) or dynamic random access memory (DRAM) is limited in the size. The DRAM, which is a volatile memory and needs to be refreshed regularly [32] is currently available at 10GB maximum [33, 34]. In other words, cache size is the biggest constraint in ICN caching. However, in most of the existing available strategies, the maximum cache space is occupied. In addition, if the memory of a router becomes full and a new content arrives, none of the existing strategies has any policy for that but simply replaces one of the cached contents by the new arrived content. This is achieved using a replacement policy, i.e., either Least Recently Used (LRU) or Least Frequently Used (LFU). As LRU and LFU replace the content based on the access time, the replaced content may be very popular and therefore the subsequent requests for the same content will be forwarded to the server. In this way, extra content retrieval delay may occur and thus maximum bandwidth is utilized.

To overcome such problems, we propose in this paper a new caching strategy, Cache Popular Content Everywhere (CPCE), which caches popular contents on all network nodes on the publisher-subscriber path. The CPCE strategy is explained in the following section.

3. PROPOSED STRATEGY

The proposed CPCE caches contents at all *on-path* routers available all the way from publisher to subscriber, as in CEE. However, the difference between our proposed caching strategy (CPCE) and the CEE is such that CEE caches every incoming content while CPCE caches contents once their popularity reaches a specified threshold value. In other words, a content is cached when its popularity reaches the threshold value in the Request Table (RT) - a table which locally calculates popularity values based on content requests, that

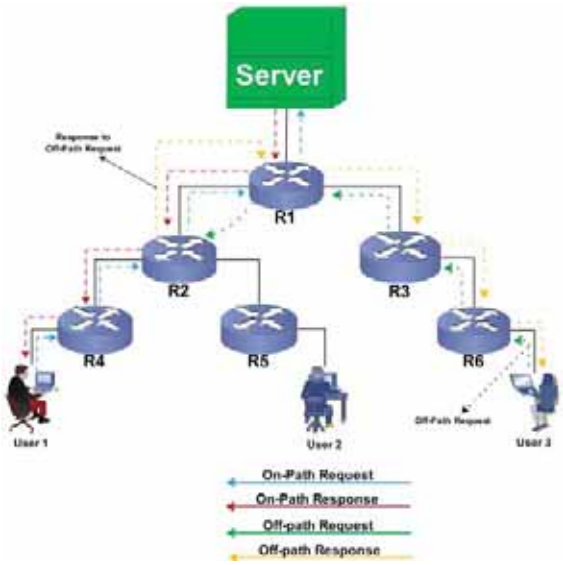


Figure 1. An example topology

is, if($C_p >= V_t$), where C_p is the content popularity and V_t is the threshold value. The V_t is kept 10 in the CPCE to avoid flooding as all incoming contents have the C_p value of 1. Therefore, a content in the CPCE is only considered popular when its C_p reaches 10. Furthermore, if the cache of all *on-path* routers overflows and a new content arrives, the CPCE uses the Least Recently Used (LRU) policy to replace one of the cached contents from each router except that the router which has the maximum outgoing interfaces, denoted by R_{max}, to accommodate the new arrived content. The reason to leave the content at the R_{max} is to avoid maximum bandwidth utilization, as majority of the content requests pass through that router. However, the question arises of how long R_{max} will keep contents as the cache of that may also overflow. Therefore, in case of cache overflows at R_{max}, on the arrival of a new content, LRU policy is used to evict one of the cached contents accessed least recently. The evicted content is cached at the underlying router, denoted by UR_1 (R2 in Figure 1), placed immediately below R_{max}. Furthermore, if the cache of UR_1 is also full then the Random policy is applied here to accommodate the content coming from R_{max}. The purpose of using the Random policy is to avoid searching overhead for content replacement as other replacement policies, e.g., LRU, take some time to find the contents based on their access time. Here the replaced content from UR_1 is moved to the next down cache router, i.e., UR_2, and the same procedure is followed until the content reaches the router placed near the subscriber, i.e., UR_n.

As the LRU does not care of content popularity and it is deployed at R_{max}, it may also evict the most popular content. Now, if a new content request arrives for the evicted popular content from *off-path*, it may also go through R_{max}; however, if hit does not occur at R_{max}, it can be found from RT that the requested content is available at the UR_i.

Hence, even the CPCE is designed for *on-path* caching but *off-path* nodes can also benifit from its versatility.

4. SYSTEM MODEL

Cache management in ICN may acquire the Hypergraph [35] characteristics, where an association is obtained between on-path routers and the original server. Our proposed CPCE strategy follows graph theory, called Hypergraph [35–37]. Let a network with routers (V) and connections (E) be represented as a Hypergraph (H) [35–38], such as: H=(V,E), where V=$\{v_1, v_2, ..., v_n\}$, and E=$\{e_1, e_2, ..., e_m\}$. Therefore, the network relationship can be defined as R=$\{r_1, r_2, ..., r_n\}$, where r_i is the i^{th} router and n is the number of total routers. Similarly, the connections denoted by E is such that E=$\{e_1, e_2, ..., e_m\}$, where e_j is the j^{th} connection and m is the number of total connections.

The GEANT topology is used for the validation of the CPCE strategy using Hypergraph, [39]:

A topology T, containing R routers and E connections as

$$T = \{R, E\} \tag{1}$$

with the objective as GEANT maintains the ICN formulation as

$$T^{'} = \{R, E\} \tag{2}$$

as GEANT consists of 22 nodes (see Figure 2), therefore

$$R = \sum_{i=1}^{22}(r_i), \tag{3}$$

where r represents the number of nodes (routers), and as the number of connections (E) in GEANT topology is 38, therefore

$$E = \sum_{j=1}^{38}(e_j). \tag{4}$$

Due to the Internet heterogeneity, each router has a connection pair e_j, such as: [$e_1 = \{r_0, r_1\}$, $e_2 = \{r_1, r_2\}$, $e_3 = \{r_2, r_3\}$, $e_4 = \{r_3, r_4\}$, $e_5 = \{r_3, r_5\}$, $e_6 = \{r_5, r_6\}$, $e_7 = \{r_6, r_7\}$, $e_8 = \{r_7, r_8\}$, $e_9 = \{r_8, r_9\}$, $e_{10} = \{r_9, r_{10}\}$, $e_{11} = \{r_{10}, r_{11}\}$, $e_{12} = \{r_{11}, r_{12}\}$, $e_{13} = \{r_4, r_{12}\}$, $e_{14} = \{r_{12}, r_9\}$, $e_{15} = \{r_{12}, r_5\}$, $e_{16} = \{r_5, r_{13}\}$, $e_{17} = \{r_5, r_{14}\}$, $e_{18} = \{r_{13}, r_{14}\}$, $e_{19} = \{r_{14}, r_{15}\}$, $e_{20} = \{r_{14}, r_{16}\}$, $e_{21} = \{r_{16}, r_{17}\}$, $e_{22} = \{r_{17}, r_1\}$, $e_{23} = \{r_{17}, r_{15}\}$, $e_{24} = \{r_{17}, r_{18}\}$, $e_{25} = \{r_{18}, r_{14}\}$, $e_{26} = \{r_{18}, r_{19}\}$, $e_{27} = \{r_{18}, r_{20}\}$, $e_{28} = \{r_{19}, r_{20}\}$, $e_{29} = \{r_{20}, r_{21}\}$, $e_{30} = \{r_{21}, r_{14}\}$, $e_{31} = \{r_{21}, r_0\}$, $e_{32} = \{r_{21}, r_4\}$, $e_{33} = \{r_4, r_{18}\}$, $e_{34} = \{r_4, r_7\}$, $e_{35} = \{r_4, r_{15}\}$, $e_{36} = \{r_{15}, r_{12}\}$, $e_{37} = \{r_{15}, r_1\}$, $e_{38} = \{r_1, r_3\}$].

Thus by the definition of Hypergraph, directed and undirected connections are achieved. In Figure 2, r_0 and r_1 have a direct connection, whereas r_0 and r_3 have an undirect connection, therefore, it is generalized in the GEANT order that r_i,r_i+1= direct, otherwise, the connection is undirect .

In the case of ICN, a router represents the overall connectivity, i.e., intersection and inter-connectivity. Hence a router's degree can be represented as [39]:

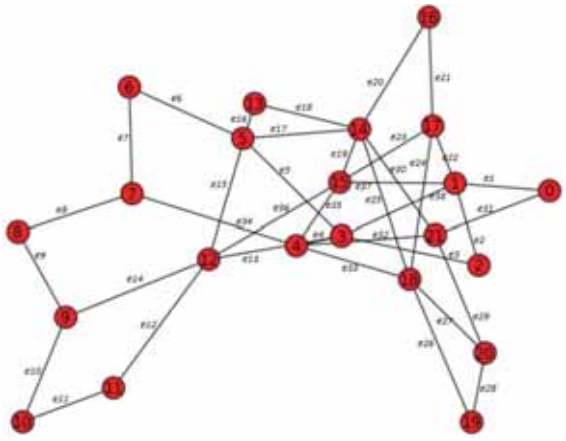


Figure 2. GEANT topology

$$d_r = \{In - degree + Out - degree\} \tag{5}$$

For example, in Figure 2 $d(r_4)$= 6 and $d(r_{14})$= 6 for 6 interconnected routers on the edge.

According to the definition of the graph, the maximum degree of a network G is represented by ΔG.

To prove our model using GEANT topology, the maximum router degree is ΔT =6.

Therefore, the overall router degree, which is defined as maximum cache capacity of ICN router is given by

$$\Delta T = \sum_{i=1}^{n}(r_i). \tag{6}$$

To know the idea of ICN routers and connections relationship, assume the routers' membership in T, as described in [35, 39], if r_i represents routers and C_i the cache size, it implies that each router r_i can cache a content. Then the cache size of router r_i can be

$$C_{size} = \sum_{i=1}^{n}(C_i). \tag{7}$$

This develops a network topology T with routers and connections as

$$T = \{r_1, r_2, ..., r_n\}, \tag{8}$$

where $r \epsilon C$.

Table 1. Simulation Scenario.

Cache Size	1GB-10GB
Catalog Size	10^8
Zipf probability (α)	0.7, 1.0
Topology	GEANT and DTelekom
Social Network Topology	Facebook [40–42]
Simulator	SocialCCNSim [43, 44]
Simulation Runs	10 times

5. ANALYSIS

To know the accuracy of the system model and the simulation, the simulations are performed in SocialCCNSim [43] - ICN caching simulator, according to the parameters presented in Table 1, while the analysis is done in Maple 18 for cache hit according to the parameters presented in Tables 2 and 3. For the analysis of cache hit, we consider a real Facebook topology [40–42] which consists of 4,039 nodes. We assume that each node in the network is placed at a constant distance: in our assumption, this distance is 25 meters. Each time when a content is downloaded, the hop decrement increases 100 hops. This assumption is made on the basis of our ordinary topology in Figure 1, where initially a subscriber is 4 hops away (on any path) from the node (i.e., Server) having the desired content. As hit occurs and the content popularity reaches the threshold value (i.e., 10), it is cached at all *on-path* routers. However, according to our proposed strategy, the content may be evicted from all routers (UR_1 to UR_n) but it will stay at the router having maximum outgoing interfaces (Router R1 in the given figure), which is 3 hops away from the user(s).

When $0 < \alpha < 1$, the asymptotic cache hit ratio, H_c, is calculated as [45, 46]:

$$H_c = C^{1-\alpha} \tag{9}$$

where *C* is the cache size that caches chunks (each chunk is of 10MB size) and α=0.7. Looking at the analysis and simulation results, presented in Table 2, the average result of analysis is 16% while it is 14.3% for simulation. The average difference is 1.7% and hence the accuracy is 98.3% . The resultant graph is shown in Figure 3(a). The same variables, i.e., cache size and chunk size, are used for the scenario when α=1.0. Now, if α=1.0, the asymptotic cache hit ratio, H_c, is calculated as [45, 46]:

$$H_c = lnC. \tag{10}$$

The analysis and simulation results are presented in Figure 3(b) and Table 3. The average analysis and simulation results are 37.6% and 36.1% , respectively. The average difference is 1.5% and therefore the accuracy is 98.5% . It is observed that when the cache size is small then the difference of analysis and simulation results (with α=1.0) is high, however, with the increase of cache size this difference approaches 0.

Similarly, for the analysis of stretch results, we consider the same scenario with a real Facebook topology [40–42]. The numerical results can be calculated as [47]:

$$S = \frac{\alpha n}{2\gamma} \tag{11}$$

where *S* represents the stretch, *n* is the total number of network nodes, γ is the distance between two hops, and α is the Zipf probability parameter, such that $0 < \alpha < 1$.

In addition, the numerical results for α=1.0 can be calculated as:

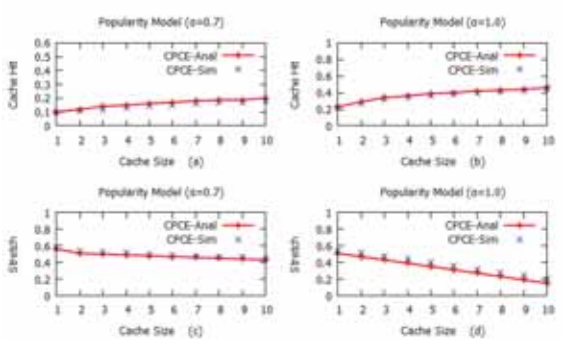


Figure 3. CPCE analysis vs. simulation

Table 2. Cache Hit: Analysis vs. Simulation with $\alpha = 0.7$ and Chunk Size 10MB.

Cache Size	Chunk	Anal	Sim	Difference
1	$1(10^2)$	0.10	0.10	0.00
2	$2(10^2)$	0.12	0.11	0.01
3	$3(10^2)$	0.14	0.12	0.02
4	$4(10^2)$	0.15	0.13	0.02
5	$5(10^2)$	0.16	0.14	0.02
6	$6(10^2)$	0.17	0.15	0.02
7	$7(10^2)$	0.18	0.16	0.02
8	$8(10^2)$	0.19	0.17	0.02
9	$9(10^2)$	0.19	0.17	0.02
10	$10(10^2)$	0.20	0.18	0.02

Table 3. Cache Hit: Analysis vs. Simulation with $\alpha = 1.0$ and Chunk Size 10MB.

Cache Size	Chunk	Anal	Sim	Difference
1	$1(10^2)$	0.23	0.22	0.01
2	$2(10^2)$	0.29	0.27	0.02
3	$3(10^2)$	0.34	0.32	0.02
4	$4(10^2)$	0.36	0.34	0.02
5	$5(10^2)$	0.39	0.37	0.02
6	$6(10^2)$	0.40	0.39	0.01
7	$7(10^2)$	0.42	0.40	0.02
8	$8(10^2)$	0.43	0.42	0.01
9	$9(10^2)$	0.44	0.43	0.01
10	$10(10^2)$	0.46	0.45	0.01

$$S = \frac{\alpha n}{2\gamma} - (D + h) \tag{12}$$

where *n* is the number of total network nodes, γ is the distance between two hops, *D* is the hop decrement, and *h* is a constant. We assume that *D*=γ and the value of *h*=4. This assumption is based on our ordinary proposed topology, shown in Figure 1. The achieved analysis results are presented in Figure 3(c) and Table 4 for α=0.7, and Figure 3(d) and Table 5 for α=1.0. The average difference in α=0.7 is 2.2% while it is 4.8% when α=1.0.

Table 4. Stretch: Analysis vs. Simulation with $\alpha = 0.7$ and Chunk Size 10MB.

Cache Size	Chunk	Anal	Sim	Difference
1	$1(10^2)$	0.56	0.58	0.02
2	$2(10^2)$	0.51	0.54	0.03
3	$3(10^2)$	0.50	0.53	0.03
4	$4(10^2)$	0.49	0.52	0.03
5	$5(10^2)$	0.48	0.50	0.02
6	$6(10^2)$	0.47	0.49	0.02
7	$7(10^2)$	0.46	0.48	0.02
8	$8(10^2)$	0.45	0.47	0.02
9	$9(10^2)$	0.44	0.46	0.02
10	$10(10^2)$	0.43	0.44	0.01

Table 5. Stretch: Analysis vs. Simulation with $\alpha = 1.0$ and Chunk Size 10MB.

Cache Size	Chunk	Anal	Sim	Difference
1	$1(10^2)$	0.51	0.54	0.03
2	$2(10^2)$	0.47	0.52	0.05
3	$3(10^2)$	0.43	0.48	0.05
4	$4(10^2)$	0.39	0.44	0.05
5	$5(10^2)$	0.35	0.40	0.05
6	$6(10^2)$	0.31	0.36	0.05
7	$7(10^2)$	0.27	0.32	0.05
8	$8(10^2)$	0.23	0.28	0.05
9	$9(10^2)$	0.19	0.24	0.05
10	$10(10^2)$	0.15	0.20	0.05

6. TOPOLOGICAL EFFECT ON RESULTS

Topology has a direct impact on the simulation results [48] and according to the ICN baseline scenarios [49, 50], there is no general agreement on the topology selection. Therefore, for a fair evaluation, apart from the GEANT topology, we have also simulated all the strategies on the DTelekom topology. To evaluate the performance of ICN caching, we believe that cache hit and stretch ratio are considered the most prominent metrics because cache hit combines the properties of other metrics, such as content eviction rate, cached element rate, and cached miss rate, while the stretch ratio also affects the hop decrement as well as the content retrieval delay. Therefore, we present the simulation results for cache hit and stretch ratio on the mentioned two topologies in the following subsections.

6.1. Cache Hit on GEANT and DTelekom

Figure 4(a,b) and Figure 5(a,b) represent the cache hit ratio on GEANT and DTelekom topologies, respectively. In a low popularity scenario, i.e., when α is 0.7, the ratio of cache hit is low on both GEANT and DTelekom topologies with cache sizes 1GB to 10GB, but with the increase of cache size as well as the increase of popularity value, i.e., α=1.0, a higher hit ratio was achieved. On the other hand and for the GEANT topology, the recorded hit ratio was 15%, 16%, and 18% for CEE, Betweenness-Centrality, and CPCE, respectively,

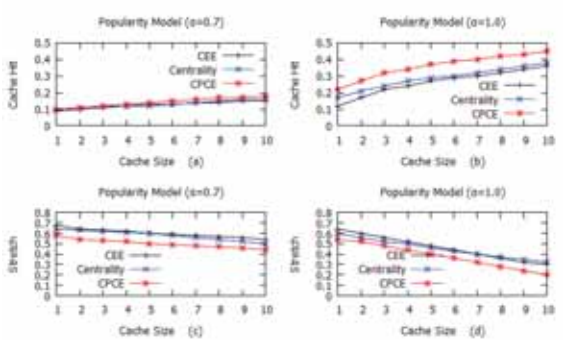


Figure 4. Result on GEANT topology

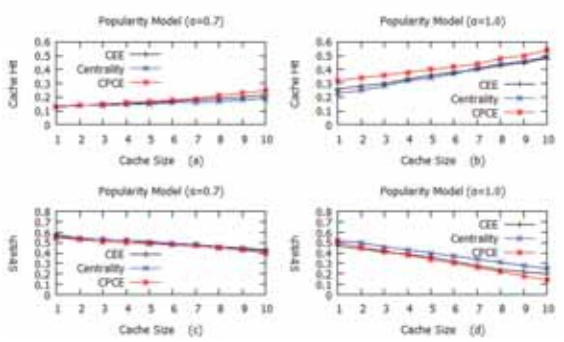


Figure 5. Result on DTelekom topology

with popularity value of 0.7. However, with the increase of popularity model, the hit ratio reached 45% with α = 1.0 in CPCE as compared to 38% of Betweenness-Centrality and 36% of CEE. Similarly, on the DTelekom topology, the achieved simulated results for α=0.7 were as follows: CEE = 22%, Betweenness-Centrality = 19%, and CPCE = 25%. While with the increase of popularity model, the hit ratio reached 54% with α = 1.0 in CPCE as compared to 49% of Betweenness-Centrality and 48% of CEE. In all figures, Betweenness-Centrality is represented by *Centrality*.

6.2. Stretch on GEANT and DTelekom

Figure 4(c,d) and Figure 5(c,d) show the stretch ratio exhibited by CEE, Betweenness-Centrality, and CPCE on GEANT and DTelekom topologies, respectively. The stretch was almost the same for all strategies with low popularity model and lower cache size on DTelekom topology, however, there was some difference on the GEANT topology. When the popularity model α was increased from 0.7 to 1.0, the recorded stretch ratio was as follows: on the GEANT topology it ws 64%-30% in CEE with cache size 1GB-10 GB, 60%-32% in Betweenness-Centrality, and 54%-20% in CPCE. While on the DTelekom topology the recorded stretch with the same parameters was as follows: CEE = 47%-20%, Betweenness-Centrality = 52%-25%, and CPCE = 50%-14%, respectively.

7. DISCUSSION

Throughout this study, the performance of CPCE is assessed with respect to different parameters, such as popularity model variation, cache size, metrics measurement, and topological impact, i.e., cache deployment. The CPCE caches popular contents, and according to Cisco visula networing index [51] multimedia applications are popular contents because video on demand (VoD) only from YouTube and Netflex attract more than half of the overall traffic, therefore, when the skewness of the distribution (which is measured through α) increases, popular contents attain maximum portion of the Internet traffic.

In addition, the current available DRAM is 10GB (maximum), therefore, after some times the memory becomes full and content eviction starts for the accommodation of new arrived contents. Consequently, the cache hit rate is reduced and subsequent requests for the evicted contents are forwarded to the server, which increases the content retrieval delay. To add more, with the eviction of contents, the hop decrement is affected and thus stretch ratio increases. Similarity, CEE and Betweenness-Centrality also cache diverse kinds of contents (as they are not particularly designed for popular contents) and the popular ones (i.e., VoD) will not have enough space to be cached. Moreover, in case of memory overflow, there is no eviction policy in CEE and Betweenness-Centrality but they simply delete the contents which affect the mentioned metrics.

The Cache size is another factor of the CPCE result supremacy, i.e., when the network becomes stable then CEE and Betweenness-Centrality cache sizes overflow and thus eviction operation starts. On the other hand, unlike CEE, the CPCE does not cache every content and hence still has space for accommodating new contents rather than deleting the cached ones. Besides, Betweenness-Centrality stores lesser contents than the CEE but because of having no eviction policy for Betweenness-Centrality, the popular contents are evicted and subsequent requests for those contents are forwarded to the original server. This leads to reduce the performance of Betweenness-Centrality. While in the CPCE the contents are not deleted but cached at other routers, as discussed earlier.

8. CONCLUSION

In the modified draft of ITU - T recommendation [52] it is requested that every network segment in data aware networking (DAN) is recommended to support a caching component and be additionally ready to assess subscriber requests that pass through it with the goal that it can make a decision on subscriber requests and respond using the cached data objects. To address the ITU-T recommendations, this paper proposes a flexible caching strategy, named CPCE, for popular content caching to improve the performance of ICN caching in terms of cache hit rate and stretch ratio during content downloading. The performance of CPCE is compared with two other strategies, i.e., the default ICN strategy, known as Cache Everything Everywhere (CEE), and Betweenness-Centrality. In simulations, the CPCE strategy outperformed both CEE and Betweenness-Centrality in terms of cache hit rate and stretch ratio.

REFERENCES

[1] G. M. de Brito, P. B. Velloso, and I. M. Moraes, *Information Centric Networks: A New Paradigm for the Internet*. John Wiley & Sons, 2013.

[2] V. Jacobson, D. K. Smetters, J. D. Thornton, M. F. Plass, N. H. Briggs, and R. L. Braynard, "Networking named content," in *Proceedings of the 5th international conference on Emerging networking experiments and technologies*. ACM, 2009, pp. 1–12.

[3] I. M. Moraes, M. E. M. Campista, L. H. M. Costa, O. C. M. Duarte, J. L. Duarte, D. G. Passos, C. V. N. de Albuquerque, and M. G. Rubinstein, "On the impact of user mobility on peer-to-peer video streaming," *IEEE Wireless Communications*, vol. 15, no. 6, pp. 54–62, 2008.

[4] T. Plagemann, V. Goebel, A. Mauthe, L. Mathy, T. Turletti, and G. Urvoy-Keller, "From content distribution networks to content networks-issues and challenges," *Computer Communications*, vol. 29, no. 5, pp. 551–562, 2006.

[5] T. Koponen, M. Chawla, B.-G. Chun, A. Ermolinskiy, K. H. Kim, S. Shenker, and I. Stoica, "A data-oriented (and beyond) network architecture," *ACM SIGCOMM Computer Communication Review*, vol. 37, no. 4, pp. 181–192, 2007.

[6] J. Kurose, "Content-centric networking: technical perspective," *Communications of the ACM*, vol. 55, no. 1, pp. 116–116, 2012.

[7] M. Berman, J. S. Chase, L. Landweber, A. Nakao, M. Ott, D. Raychaudhuri, R. Ricci, and I. Seskar, "Geni: A federated testbed for innovative network experiments," *Computer Networks*, vol. 61, pp. 5–23, 2014.

[8] FIA, "Future internet assembly (fia)," Tech. Rep. [Online]. Available: http://www.future-internet.eu/home/future-internet-assembly.html

[9] AsiaFI, "Asiafi," Tech. Rep. [Online]. Available: http://www.asiafi.net/meeting/2010/summerschool/

[10] G. Garcia, A. Beben, F. J. Ramon, A. Maeso, I. Psaras, G. Pavlou, N. Wang, J. Sliwinski, S. Spirou, S. Soursos *et al.*, "Comet: Content mediator architecture for content-aware networks," in *Future Network & Mobile Summit (FutureNetw), 2011*. IEEE, 2011, pp. 1–8.

[11] FP7 Scalable and Adaptive Internet Solutions: SAIL project. [Online]. Available: http://www.sail-project.eu/

[12] L. Zhang, D. Estrin, J. Burke, V. Jacobson, J. D. Thornton, D. K. Smetters, B. Zhang, G. Tsudik, D. Massey, C. Papadopoulos *et al.*, "Named data networking (ndn) project," *Relatorio Tecnico NDN-0001, Xerox Palo Alto Research Center-PARC*, 2010.

[13] W. K. Chai, D. He, I. Psaras, and G. Pavlou, "Cache less for more in information-centric networks (extended version)," *Computer Communications*, vol. 36, no. 7, pp. 758–770, 2013.

[14] I. Psaras, W. K. Chai, and G. Pavlou, "Probabilistic in-network caching for information-centric networks," in *Proceedings of the second edition of the ICN workshop on Information-centric networking*. ACM, 2012, pp. 55–60.

[15] Z. Li and G. Simon, "Time-shifted tv in content centric networks: The case for cooperative in-network caching," in *Communications (ICC), 2011 IEEE International Conference on*. IEEE, 2011, pp. 1–6.

[16] S. Eum, K. Nakauchi, Y. Shoji, and N. Nishinaga, "CATT: Cache aware target identification for ICN," *Communications Magazine, IEEE*, vol. 50, no. 12, pp. 60–67, 2012.

[17] S. Eum, K. Nakauchi, M. Murata, Y. Shoji, and N. Nishinaga, "CATT: Potential based routing with content caching for ICN," in *Proceedings of the 2nd ACM SIGCOMM workshop on Information-centric networking*. ACM, 2012, pp. 49–54.

[18] G. Zhang, B. Tang, X. Wang, and Y. Wu, "An optimal cache placement strategy based on content popularity in Content Centric Network," *Journal of Information & Computational Science*, vol. 11, no. 8, pp. 2759–2769, 2014.

[19] J. Wang, J. Ren, K. Lu, J. Wang, S. Liu, and C. Westphal, "An optimal cache management framework for information-centric networks with network coding," in *Networking Conference, 2014 IFIP*. IEEE, 2014, pp. 1–9.

[20] K. Cho, M. Lee, K. Park, T. T. Kwon, Y. Choi, and S. Pack, "Wave: Popularity-based and collaborative in-network caching for content-oriented networks," in *Computer Communications Workshops (INFOCOM WKSHPS), 2012 IEEE Conference on*. IEEE, 2012, pp. 316–321.

[21] C. Bernardini, T. Silverston, and O. Festor, "Mpc: Popularity-based caching strategy for content centric networks," in *Communications (ICC), 2013 IEEE International Conference on*. IEEE, 2013, pp. 3619–3623.

[22] M. Ong, M. Chen, T. Taleb, X. Wang, and V. Leung, "Fgpc: fine-grained popularity-based caching design for content centric networking," in *Proceedings of the 17th ACM international conference on Modeling, analysis and simulation of wireless and mobile systems*. ACM, 2014, pp. 295–302.

[23] G. Carofiglio, V. Gehlen, and D. Perino, "Experimental evaluation of memory management in content-centric networking," in *Communications (ICC), 2011 IEEE International Conference on*. IEEE, 2011, pp. 1–6.

[24] I. Ud, Din, S. Hassan, and A. Habbal, "A content placement scheme for Information Centric Networking," *Advanced Science Letters*, vol. 21, no. 11, pp. 3484–3486, 2015.

[25] S. K. Fayazbakhsh, Y. Lin, A. Tootoonchian, A. Ghodsi, T. Koponen, B. Maggs, K. Ng, V. Sekar, and S. Shenker, "Less pain, most of the gain: Incrementally deployable icn," in *ACM SIGCOMM Computer Communication Review*, vol. 43, no. 4. ACM, 2013, pp. 147–158.

[26] Y. Thomas and G. Xylomenos, "Towards improving the efficiency of icn packet-caches," in *Heterogeneous Networking for Quality, Reliability, Security and Robustness (QShine), 2014 10th International Conference on*. IEEE, 2014, pp. 186–187.

[27] L. Zhang, J. Zhao, and Z. Shi, "LF: A caching strategy for named data mobile ad hoc networks," in *Proceedings of the 4th International Conference on Computer Engineering and Networks*. Springer, 2015, pp. 279–290.

[28] V. G. Vassilakis, M. F. Al-Naday, M. J. Reed, B. Alzahrani, K. Yang, I. D. Moscholios, M. D. Logothetis *et al.*, "A cache-aware routing scheme for Information Centric Networks," in *Communication Systems, Networks & Digital Signal Processing (CSNDSP), 2014 9th International Symposium on*. IEEE, 2014, pp. 721–726.

[29] J. Garcia-Luna-Aceves, "A fault-tolerant forwarding strategy for interest-based information centric networks," in *IFIP Networking Conference (IFIP Networking), 2015*. IEEE, 2015, pp. 1–9.

[30] Y. Li, H. Xie, Y. Wen, C.-Y. Chow, and Z.-L. Zhang, "How much to coordinate? optimizing in-network caching in Content-Centric Networks," *Network and Service Management, IEEE Transactions on*, vol. 12, no. 3, pp. 420–434, 2015.

[31] I. Ud, Din, S. Hassan, and A. Habbal, "A content eviction mechanism for Information Centric Networking," *ARPN Journal of Engineering and Applied Sciences*, vol. 11, no. 5, pp. 3233–3235, 2016.

[32] D. Perino and M. Varvello, "A reality check for content centric networking," in *Proceedings of the ACM SIGCOMM workshop on Information-centric networking*. ACM, 2011, pp. 44–49.

[33] G. Rossini, D. Rossi, M. Garetto, and E. Leonardi, "Multi-terabyte and multi-gbps information centric routers," in *IEEE INFOCOM 2014-IEEE Conference*

on Computer Communications. IEEE, 2014, pp. 181–189.

[34] S. Arianfar, P. Nikander, and J. Ott, "Packet-level caching for information-centric networking," in *ACM SIGCOMM, ReArch Workshop*, 2010.

[35] G. Gallo, G. Longo, S. Pallottino, and S. Nguyen, "Directed hypergraphs and applications," *Discrete applied mathematics*, vol. 42, no. 2, pp. 177–201, 1993.

[36] V. I. Voloshin, *Introduction to graph and hypergraph theory*. Nova Science Publ., 2009.

[37] A. Bretto, "Hypergraph theory," *An introduction. Mathematical Engineering. Cham: Springer*, 2013.

[38] M. Xiao, "Finding minimum 3-way cuts in hypergraphs," *Information Processing Letters*, vol. 110, no. 14, pp. 554–558, 2010.

[39] I. Abdullahi, S. Arif, and S. Hassan, "Cache-less redundancy using hypergraph in information-centric network," *Advanced Science Letters*, vol. 21, no. 11, pp. 3546–3549, 2015.

[40] J. Leskovec and A. Krevl, "SNAP Datasets: Stanford large network dataset collection," Jun. 2014. [Online]. Available: http://snap.stanford.edu/data

[41] J. Leskovec and J. J. Mcauley, "Learning to discover social circles in ego networks," in *Advances in neural information processing systems*, 2012, pp. 539–547. [Online]. Available: http://cs.stanford.edu/people/jure/pubs/circles-nips12.pdf

[42] Stanford Network Analysis Project (SNAP). [Online]. Available: http://snap.stanford.edu/data/egonets-Facebook.html

[43] C. Bernardini. SocialCCNSimm. [Online]. Available: https://github.com/mesarpe/socialccnsim

[44] I. Ud, Din, S. Hassan, and A. Habbal, "SocialCCNSim: A simulator for caching strategies in Information Centric Networking," *Advanced Science Letters*, vol. 21, no. 11, pp. 3507–3511, 2015.

[45] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, "Web caching and zipf-like distributions: Evidence and implications," in *INFOCOM'99. Eighteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*, vol. 1. IEEE, 1999, pp. 126–134.

[46] M. Gallo, B. Kauffmann, L. Muscariello, A. Simonian, and C. Tanguy, "Performance evaluation of the random replacement policy for networks of caches," in *ACM SIGMETRICS Performance Evaluation Review*, vol. 40, no. 1. ACM, 2012, pp. 395–396.

[47] M. Won and R. Stoleru, "A low-stretch-guaranteed and lightweight geographic routing protocol for large-scale wireless sensor networks," *ACM Transactions on Sensor Networks (TOSN)*, vol. 11, no. 1, p. 18, 2014.

[48] D. Rossi and G. Rossini, "Caching performance of Content Centric Networks under multi-path routing (and more)," *Relatorio tecnico, Telecom ParisTech*, 2011. [Online]. Available: http://netlab.pkusz.edu.cn/wordpress/wp-content/uploads/2011/10/

[49] J. Seedorf, M. Arumaithurai, A. Tagami, K. Ramakrishnan, and N. Blefari-Melazzi, "Using ICN in disaster scenarios," Internet-Draft-work in progress 02, IETF, Tech. Rep., 2014.

[50] K. Pentikousis, B. Ohlman, D. Corujo, G. Boggia, G. Tyson, E. Davies, A. Molinaro, and S. Eum, "Information-Centric Networking: Baseline scenarios," Tech. Rep., 2015. [Online]. Available: http://www.rfc-editor.org/info/rfc7476

[51] Cisco, "Cisco visual networking index: forecast and methodology: 2013-2018," Tech. Rep. [Online]. Available: http://www.intel.com/content/www/us/en/communications/internet-minute-infographic.html

[52] ITU-T. ITU-T Y.3033 Recommendation, ITUT Framework of data aware networking for future networks. [Online]. Available: https://www.itu.int/rec/T-REC-Y.3033-201401-I/en

MULTI-PATH CHUNKED VIDEO EXCHANGES OVER OF@TEIN SDN CLOUD PLAYGROUND

Phyo May Thet[1], JongWon Kim[2], Chaodit Aswakul[1]

1. Wireless Network and Future Internet Research Group, Department of Electrical Engineering, Faculty of Engineering, Chulalongkorn University, Thailand
2. Networked Computing Systems Laboratory, School of Information and Communications, Gwangju Institute of Science & Technology, Korea

phyo.ma@student.chula.ac.th, jongwon@nm.gist.ac.kr, chaodit.a@chula.ac.th

ABSTRACT

Recent explosion of SDN (software-defined networking) paradigm is steadily facilitating the open evolution of legacy protocol-biased networking. In this paper, we design and develop middle-box splitting functionalities for chunked video exchanges (i.e., file transfer and streaming) over multiple concurrent paths over South Korea, Malaysia, and Thailand parts of OF@TEIN SDN-Cloud playground. To reduce middle-box processing delay and to enable multi-path capacity leverage, we propose to combine the multi-path file-transfer function and Tsunami protocol. This combination allows the multi-path-based pre-fetch transfer of 4K-video chunks from Tsunami server in South Korea to Thailand's SmartX Box. The experiment results show that the proposed scheme can reduce 23~72 seconds in transferring 10-minute 4K-video file by effectively utilizing the available international multi-path capacity.

Keywords— OpenFlow software-defined networking, middle-box coordination, multi-path file transfer and streaming.

1. INTRODUCTION

Software-defined networking (SDN) [1] has emerged in the networking industry and standardization bodies as the new paradigm for facilitating the evolution of networking protocols. SDN brings the flexible centralized network programmability and controllability to the network operators for managing the network configuration easily. SDN introduces the functions of separating the control and data planes by open access protocol called OpenFlow [2]. SDN can lead various promising benefits to network operators, vendors and users such as centralized network management, lower capital expenses, ability to control multi-vendor networking devices, increasing rate of reliability and security of networks via programmability with open interfaces.

SDN comes up with network function virtualization and network orchestration which are very popular technologies for implementing future internet infrastructures. Moreover, SDN becomes a major requirement for cloud computing service providers since SDN eliminates the need for manual configuration of distributed switches, routers and hardware devices by using a centralized management system. In SDN, everything is controlled by software through a single or multiple SDN controllers. To build a SDN cloud computing infrastructure, OpenStack [3] becomes a popular cloud computing application due to its open source and fast emerging functionalities support for user requirements. Cloud services such as cloud video streaming, file transferring are popular applications among users.

To evaluate file-transfer experiments over TCP with SDN technology, a multi-path controller for GridFTP transfer over SDN has been introduced in [4] by testing in both virtual and real global-scale networks. However, file-transferring over TCP can lead to unnecessarily congested network conditions and can cause transmission delay due to retransmission and multiple acknowledgement requests. Moreover, although multi-path TCP file-transfer method over large scale network has been introduced in [4], no investigations have been made for how much transmission delay can be decreased due to their proposed multi-path TCP file-transferring over international large-scaled network. This challenge has motivated us to conduct research on UDP file-transferring with SDN multi-path technologies over an international large-scaled network. File-transfer protocols such as UDP-based data transfer protocol (UDT) for high-speed wide area networks and Tsunami file-transfer protocol via UDP and TCP without using SDN have been implemented in [5] and [6].

For video streaming over SDN, multi-path chunked video streaming over emulated OpenFlow network by using splitting and middle-box functionalities have been introduced in our previous studies [7,8]. Unlike in [7,8], in this paper we propose the solutions for addressing file-

This research has been supported by the Special Task Force for Activating Research (STAR) Funding in Wireless Network and Future Internet Research Group, Chulalongkorn University, Thailand, and is based on the collaborations of OF@TEIN project. The first author wishes also to acknowledge the scholarship for International Graduate Students in ASEAN Countries, Chulalongkorn University.

transfer delay over actual international gateway environments by using splitting and middle-box functionalities. In order to perform experiments over international gateway environments, we have selected OpenFlow at Trans-Eurasia Information Network (OF@TEIN) which is connected to 11 sites in 9 countries (Korea, Indonesia, Malaysia, Thailand, Vietnam, Philippines, Pakistan, India and Taiwan) over TEIN4 [9]. OF@TEIN is a large-scaled international cloud playground testbed. In this study, we have chosen three international nodes which are located in South Korea, Malaysia and Thailand to set up the smallest test network with two alternative paths available for data transmissions.

The main contributions of this paper are as follows:

- To design and develop middle-box splitting functionalities for chunked video exchanges (i.e., file transfer and streaming) by using POX controller, OpenStack, Open vSwitches, KVM based Middle-box and SmartX boxes
- To reduce middle-box processing delay and to enable multi-path capacity leverage by introducing the combination of multi-path file transfer function and Tsunami protocol over OF@TEIN SDN cloud playground.

The objective is to design and develop middle-box splitting functionalities for chunked video exchanges (i.e., file transfer and streaming) over multiple concurrent paths of OF@TEIN SDN cloud playground which has plenty of available bandwidth but non-guaranteed and uncertain path capacities. TCP file-transferring over international links is limited in terms of transfer duration and bandwidth capacity due to the firewall in each node and the distance between nodes. Moreover, in transferring via TCP, all packet losses are counted as congestion and multiple retransmission requests decrease the bandwidth capacity.

In order to solve out that kind of issues with TCP, we have selected Tsunami [6] file-transfer protocol in these multi-path chunked video exchanges experiments. Tsunami [6] is a combination of UDP and TCP file-transfer protocols in which bulk data are transferred via UDP and control data are transferred via TCP. In this paper, to reduce middle-box processing delay and to enable multi-path capacity leverage, we propose to combine the multi-path file-transfer function and Tsunami protocol over OF@TEIN SDN cloud playground. In addition, after the completion of file-transfer experiments, we would investigate how downloaded video file can be streamed out smoothly within the local area network.

2. DESIGN OF MULTI-PATH CHUNKED VIDEO EXCHANGES

In this multi-path chunked video exchanges (i.e., file transfer and streaming) experiments over OF@TEIN SDN cloud playground, we have selected three countries: Thailand, Malaysia and South Korea. The following SmartX boxes have been used in this multi-path chunked video exchanges experiments: GIST-B (Gwangju Institute of Science and Technology, Korea), MY (University of Malaya, Malaysia) and CHULA (Chulalongkorn University, Thailand). The overall and detailed architectures of multi-path chunked exchanges over OF@TEIN SDN cloud playground are depicted in Figures 1 and 2.

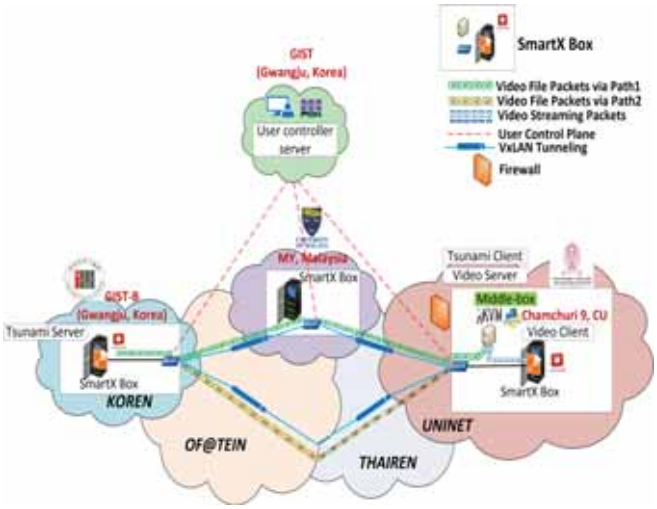


Figure 1: Overview of multi-path chunked video exchanges over OF@TEIN SDN cloud playground.

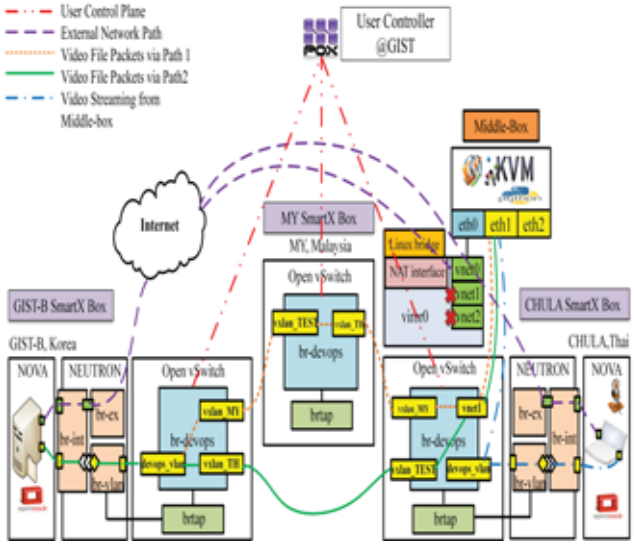


Figure 2: Architecture of multi-path chunked video exchanges over OF@TEIN SDN cloud playground.

As described in Figures 1 and 2, we have created OpenStack [3] VM running Ubuntu 14.04 OS with 1 VCPU and 1GBytes of RAM in GIST-B OpenStack compute node and OpenStack VM running Ubuntu 14.04 OS with 2 VCPUs and 4GBytes of RAM in Thailand OpenStack compute node for serving as a Tsunami video file-transfer server and a VLC video streaming client respectively. The middle-box with 1 VCPU and 2GBytes of RAM running Ubuntu 12.04 OS which has been implemented by using Kernel virtual machine (KVM) in the Thailand SmartX box for serving as a Tsunami video file-transfer client and a VLC video streaming server. OpenStack Nova [10] and Neutron [11] functions are

initiating the OpenStack instances and OpenStack network services, respectively. VxLAN tunneling has been used to connect between three br-devops in three SmartX boxes (GIST-B, MY and CHULA) from different countries.

Unlike in the old OF@TEIN SDN cloud playground architecture [9], FlowVisor [12] has not been used in this version of OF@TEIN SDN cloud playground architecture to support both OpenFlow 1.0 and OpenFlow 1.3 functions into OF@TEIN cloud playground. The reason is that FlowVisor does not support for OpenFlow 1.3 functions. Therefore, only one user has been allowed to run experiments over new OF@TEIN architecture during the requested time schedule. Therefore, developers need to reserve the timing for running experiments over OF@TEIN. As a first trial of multi-path experiments over OF@TEIN cloud playground, we have selected OpenFlow 1.0 functions as in our previous studies [7,8]. The OVS bridge (br-devops) has been created instead of br1, br2 and brcap which have used in the old OF@TEIN SDN cloud playground architecture [9]. VxLAN tunneling has been used to connect between three br-devops in three SmartX boxes (GIST-B, MY and CHULA) from different countries. There is no separation of user and admin controllers in this OF@TEIN SDN cloud playground architecture. The POX [13] controller has been used for adding flow entries into OVS bridges (br-devops) at GIST-B, MY and CHULA.

Firstly, the Tsunami video file-transfer server sends out the chunked video files into multiple concurrent paths via Path 1 (via GIST-B>MY>CHULA) and Path 2 (via GIST-B>CHULA). The OVS (br-devops) at GIST-B is responsible for splitting chunked video files into Path 1 and Path 2 periodically. The OVS (br-devops) at CHULA is responsible for combining the video file packets from two paths and then forwarding video streaming packets from the middle-box to the video client in CHULA OpenStack VM. The middle-box serves as a Tsunami video file-transfer client and a VLC video streaming server. The IP address of eth1 in the middle-box have been configured to be the same network and the same VLAN ID (111) as in OpenStack VMs. VLAN (802.1q) configuration program (vconfig) [14] have been used for adding VLAN ID into eth1 of the middle-box. The MTU size of all eth1 interfaces for GIST-B VM, CHULA VM and the middle-box have been configured as 1410bytes in order to be able to transmit the iPerf traffic. The flow entries of each SmartX boxes have been used in these experiments are shown in Tables 1-4.

Table 1: Flow entry of GIST-B SmartX box (OVS: br-devops) for transmission via path 1 (GIST-B>MY>CHULA) [7,8]

Header field	Action	Timeout
in_port = 1 (devops_vlan)	output:2 (vxlan_MY)	hard_timeout = *m*
in_port = 2 (vxlan_MY)	output:1 (devops_vlan)	hard_timeout = *m*

Table 2: Flow entry of GIST-B SmartX box (OVS: br-devops) for transmission via path 2 (GIST-B>CHULA) [7,8]

Header field	Action	Timeout
in_port = 1 (devops_vlan)	output:3 (vxlan_TH)	hard_timeout = *n*
in_port = 3 (vxlan_TH)	output:1 (devops_vlan)	hard_timeout = *n*

As shown in Tables 1 and 2, the video file packets are periodically chunked into smaller files at GIST-B (br-devops) by specifying the *m* s to parameter hard timeout of the flow entry for packet transmission via Path 1 and *n* s to parameter hard timeout for that via Path 2. The ratio *m:n* then determines the chunk size ratio in this splitting function as in [7,8].

Table 3: Flow entry of MY SmartX box (OVS: br-devops)

Header field	Action
in_port = 1(devops_vlan)	output:3(vxlan_TH)
in_port = 3 (vxlan_TH)	output:1(devops_vlan)

The flow tables at MY OVS (br-devops) as shown in Table 3 are simply that when the packets arrive from vxlan_TEST (GIST-B), those packets will be forwarded to vxlan_TH. When the packets arrive from vxlan_TH, those packets will be forwarded to vxlan_TEST.

Table 4: Flow entry of CHULA SmartX box (OVS: br-devops)

Header field	Action
in_port = 3 (vxlan_MY)	mod_dl_src:c2:ff:3e:21:81:d4, output:5(vnet1)
in_port = 2 (vxlan_TEST)	mod_dl_src:f2:a8:b7:ea:85:63, output:5(vnet1)
in_port = 5 (vnet1)	ALL
in_port = 1 (devops_vlan)	output:5(vnet1)

The flow entries of br-devops in CHULA SmartX box are shown in Table 4. The br-devops in CHULA SmartX box is responsible for combining the video file packets from two paths and forwarding the video streaming packets from the middle-box to the video client in CHULA OpenStack VM. When the packets arrive from the ingress vxlan_MY and vxlan_TEST ports, the packet headers are set to be c2:ff:3e:21:81:d4 for packets coming from vxlan_MY and f2:a8:b7:ea:85:63 for packets coming from vxlan_TEST and then forwarded to the vnet1 for receiving Tsunami video file client in the middle-box. After completely receiving the video file packets in the middle-box, the middle-box serves as a VLC streaming server. For the packets arrive from the vnet1 ingress port,

all packets will be forwarded to all available ports at br-devops except the ingress port vnet1. Therefore, the network of the middle-box is reachable to both GIST-B and CHULA OpenStack VMs. The packets arrive from devops_vlan will be forwarded only to vnet1 so that CHULA OpenStack VM data path network is reachable to the middle-box. The reason is that we do not need to transmit any traffic between GIST-B OpenStack VM and CHULA OpenStack VM in this experiment. Therefore, we cut out the routing between GIST-B OpenStack VM and CHULA OpenStack VM in order to save the unnecessary bandwidth usages.

Another implementation to solve the fast time responsiveness of GUI applications in accessing remote OpenStack VMs over OF@TEIN SDN cloud playground have been introduced in this paper. The limited access of GUI applications with X11 display over OF@TEIN OpenStack VMs is limited for streaming video with large video resolution size. In order to solve out those issues, we have implemented the X11 Desktop Environment in OpenStack VMs by using light weight X11 desktop environment (LXDE) [15] which supports fast desktop performance for easy access GUI applications in the cloud. Moreover, in order to access remotely to the implemented Desktop Environments, we have installed an open source remote desktop protocol (rdp) server called xrdp [16], a free remote control software called tightvnc [17] and an open source implementation of the X Window system called xorg [18]. There are three access methods for OpenStack VMs: (1) access via xrdp without requiring port information, (2) access via tightvnc with requiring 5901 port access and (3) access via ssh with X11 display. For accessing method 1 and 2, we can use either remote desktop connection with Window OSs or rdesktop with Linux/ Ubuntu OSs.

3. RESULTS AND DISCUSSION

The Big Buck Bunny, a 4K-animation video with a total file size of 843Mbytes, resolution 3840 x 2610 and the duration of 10-minute has been used in this multi-path chunked video exchanges (i.e., file transfer and streaming) experiments. Tsunami file-transfer application [6] has been installed on both GIST-B OpenStack VM and middle-box VM in CHULA SmartX box. Since the purpose of this experiment is to investigate the effects of combining our proposed multi-path file-transfer function and traditional Tsunami file-transfer protocol, we have not modified the default codes of Tsunami application. The Tsunami server in GIST-B OpenStack VM starts the Tsunami server by using the command “tsunamid –port 46224 filename”. The middle-box VM in CHULA SmartX box receives the video file by using the command “tsunami set rate xxxM connect serveripaddress get filename". The detailed parameter settings of video file used for file transferring experiments are shown in Table 5.

Table 5: Detailed parameters of video file for file transferring experiments

Video File Detailed Parameters	
Video Format	mp4
Duration	10 minutes
Total File Size	843Mbytes
Resolution	3840 x 2610

The parameters for Tsunami file-transfer protocol are the default block size: 1024bytes (how large UDP blocks to use) and buffer size: 20Mbytes (size of ring buffer in RAM) in this experiment. Tsunami protocol allows a client to choose many parameters such as block size, buffer size, target file transfer rate, error threshold, and inter-packet delay. However, in this experiment, we vary only target file transfer rates in order to investigate the effects of varying target file transfer rates on transfer duration.

After completely received the video file in the middle-box, Tsunami protocol generates the analytic results of file transferring. Among various output results from Tsunami analytic results, transfer duration, file data, throughput and final file rate (actual file transfer rate) have been used for analyzing our multi-path-based pre-fetch transfer of 4K-video experiments. After completely transferring a video file in the middle-box, the middle-box serves as a VLC streaming server to stream out the downloaded video towards CHULA OpenStack VM in order to investigate the performance of the downloaded video stream. iPerf UDP test has been performed to check the available bandwidth between GIST-B, MY and CHULA nodes. The bandwidth capacities of around 422Mbps between OpenStack VM at GIST-B and middle-box VM at CHULA via Path 1 (GIST-B>MY>CHULA) and around 456Mbps via Path 2 (GIST-B>CHULA) have been obtained from the test. The round trip time (RTT) via the MY route is approximately 125ms and the RTT delay from GIST-B to CHULA via the direct route is about 105ms.

Three scenarios have been tested for this multi-path-based pre-fetch transfer of 4K-video experiments. They are (1) multi-path with chunk size ratio (1:2sec) which is periodically splitting by transmitting 1 second via Path 1 and transmitting 2 seconds via Path 2 (2) using Path 1 alone (GIST-B>MY>CHULA) (3) using Path 2 alone (GIST-B>CHULA). In order to investigate the effects of target file transfer rate on the transfer duration, we have varied the target file transfer rates of Tsunami protocol to be 100, 200, 300, 400 and 500Mbps and have tested three times with the same parameter settings for all three selected scenarios. The file transfer duration can be computed from the following formula.

$$Transfer\ Duration = \frac{Total\ File\ Size}{Actual\ File\ Transfer\ Rate}$$

Where:
Transfer Duration = File transfer duration between server and client in seconds (s),

Total File size = File size in bits (bits) and
Actual File Transfer Rate = Rate of file transmission in bits per second (bps).

Transfer duration results of transferring 843 Mbytes (6744Mbits) 4K-video file by using three scenarios: multi-path (1:2sec), using Path 1 alone (GIST-B> MY>CHULA) and using Path 2 alone (GIST-B>CHULA) are shown in Table 6 and Figure 3. The transfer duration results are the average results of three times tested with the same parameter settings. According to the tested results, file transfer duration decreases with upper bound associated to the available link bandwidth capacity when the target file transfer rate of Tsunami protocol increases for all three scenarios. That result trend can be seen when comparing the cases of target file transfer rates (100, 200, 300 and 400Mbps). The required transfer duration of using Path 2 alone (GIST-B>CHULA) and multi-path (1:2sec) are similar until the target file transfer rate is up to 300 Mbps when the link bandwidth capacity is enough to carry the whole video file traffic. As for using Path 1 alone (GIST-B>MY>CHULA), the required file transfer duration is higher than the other two scenarios, although there is enough link capacity to carry the whole traffic. The reason is that the longer round trip time (RTT) of 125ms is required for transmitting via Path 1 (via MY route) while RTT delay via Path 2 (via a direct link to CHULA) would require 105ms. Moreover, the RTT delay of multi-path (1:2sec) is periodically switching between 125ms and 105ms. Therefore, the RTT delay is an important factor to consider in order to obtain the lower transmission delay.

In the case of target file rate 400 and 500Mbps, when the link becomes congested, the file transfer duration results of our proposed multipath (1:2sec) case outperform the cases using Path 1 alone and Path 2 alone. However, the target file transfer rate of 500Mbps is not recommended with our multi-path function because it requires a longer delay than those of 400Mbps rate. In that case, UDP packet losses increase due to the overloaded links. The reason that the file transfer duration of our proposed multi-path (1:2sec) case cannot outperform in the cases of target file transfer rate of 100, 200, 300Mbps is that using Tsunami file-transfer protocol limits the maximum transfer rate. Therefore, the transfer rate cannot be more than the specified target rate when using our multi-path function. The tested results of file transfer duration confirm that using multi-path splitting function achieves the lowest file transfer duration time when the links are congested and do not have enough capacity to carry the whole traffic by using a single path.

Table 6: Results of average file transfer duration over OF@TEIN

Testing Scenarios	File Transfer Duration (s)				
Target Transfer Rate(Mbps)	100	200	300	400	500
Multi-path(1:2sec)	72.2	36.0	25.7	23.5	26.5
Single-path(GIST-B>CHULA)	71.0	35.7	24.9	30.4	33.0
Single-path(GIST-B>MY>CHULA)	74.3	50.2	25.1	44.3	35.0

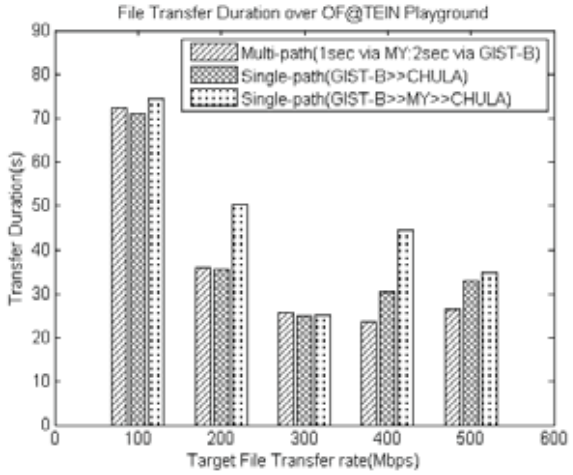


Figure 3: File transfer duration over OF@TEIN SDN cloud playground.

Table 7: Results of average throughput over OF@TEIN

Testing Scenarios	Throughput (Mbps)				
Target Transfer Rate (Mbps)	100	200	300	400	500
Multi-path(1:2sec)	94.2	188.3	262.9	341.0	394.7
Single-path(GIST-B>CHULA)	94.9	188.8	274.9	334.7	354.3
Single-path(GIST-B>MY >CHULA)	91.2	134.9	268.7	226.2	303.6

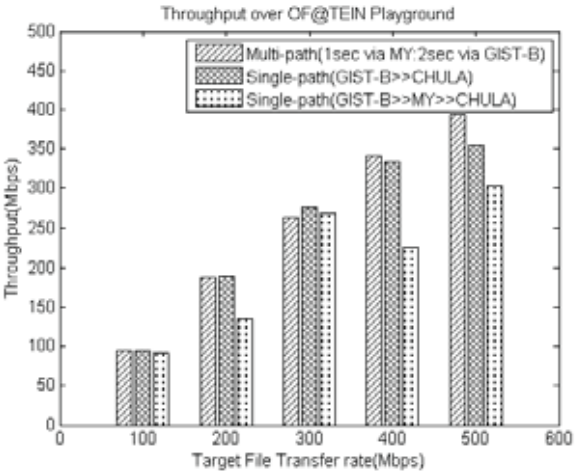


Figure 4: Throughput over OF@TEIN SDN cloud playground.

Tables 7 and 8 show the results of file transfer throughput and actual file transfer rate of varying target file transfer rate for three scenarios and those results are average results from tested three times with same parameter settings. Figures 4 and 5 depict the file transfer throughput and actual file transfer rate for three scenarios with various target file transfer rates. Those throughput results confirm the remarks as discussed above. The file transfer throughput shown in Table 7 and Figure 4 are the available throughput during the file transferring period. According to the file transfer throughput results in the case of 400 and 500Mbps target file transfer rates, the throughput obtained by using multi-path function is higher than that by using single-path alone. However, those file

transfer throughput results have not been used for calculating file transfer duration results.

In order to calculate the file transfer duration results, the average actual file transfer rate results as shown in Table 8 and Figure 5 have been used. According to the actual file transfer results, the highest actual file transfer rate achieves in the case of using 400Mbps target file rate with multi-path scenario which combines the bandwidth of two paths periodically. In the cases of 100, 200, 300Mbps, we recommend to use only Path 2 (GIST-B>CHULA) in order to reduce the cost of multi-national network links. Comparing to the file transfer rates of the cases 100, 200, 300Mbps and the cases 400, 500Mpbs, the single-path file transferring function outperforms the multi-path function in the former cases due to the maximum file transfer limit of Tsunami application. However, multi-path function is recommended to use in order to obtain the highest file transfer rate when the single-path alone is not enough to carry out the whole video file traffic. We can see that trend from the case of target file transfer rate 400, 500Mbps.

Table 8: Results of average actual file transfer rate over OF@TEIN

Testing Scenarios	Actual File Transfer Rate (Mbps)				
Target Transfer Rate (Mbps)	100	200	300	400	500
Multi-path(1:2sec)	93.4	187.3	264.8	288.2	255.7
Single-path(GIST-B>CHULA)	94.9	188.8	271.4	224.2	193.9
Single-path(GIST-B>MY >CHULA)	91.2	134.5	268.7	169.8	207.4

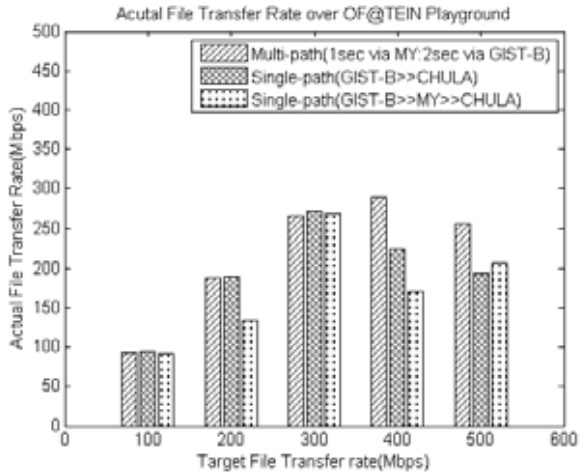


Figure 5: Actual file transfer rate over OF@TEIN SDN cloud playground.

Table 9: Video streaming experimental parameters

Video Streaming and Transcoding Detailed Parameters	
Video Format	H. 264
Duration	10 minutes
Total File Size	843Mbytes
Resolution	3840 x 2610
Audio and Video Mean Bit Rate	11133kbits/sec

After completely received the video file in the middle-box VM, the middle-box VM serves as a VLC streaming server for CHULA OpenStack VM. This experiment is in order to evaluate the performance of downloaded video file and to stream out smoothly within CHULA SmartX box network. The Big Buck Bunny H. 264 video codec with video and audio mean bit rate of total 11133kbits/sec with resolution 3840x2610 and duration of 10-minute as shown in Table 9 has been used for streaming from the middle- box VLC server to the video client in CHULA OpenStack VM. The VLC player from the video server has been configured to stream out video by using RTP mode. The VLC GUI interface has been used in order to stream out with original video tracks. The video packets are captured at eth1.111 of the middle-box VM and eth1 of CHULA OpenStack VM by using tcpdump packet analyzer. To investigate the performance of video streaming, we have run three times for the same video streaming session. For the detailed dynamics of packet generation arriving at the video server and the received packet stream departing at the video client, we include here the arrival and departure curves in Figure 6. The number of packets shown in the graph is the number of packets transmitted during video streaming. The number of packet losses is zero in all tests. This curve confirms that 4K RTP video streaming within CHULA SmartX box network can be streamed out without any significant extra delay.

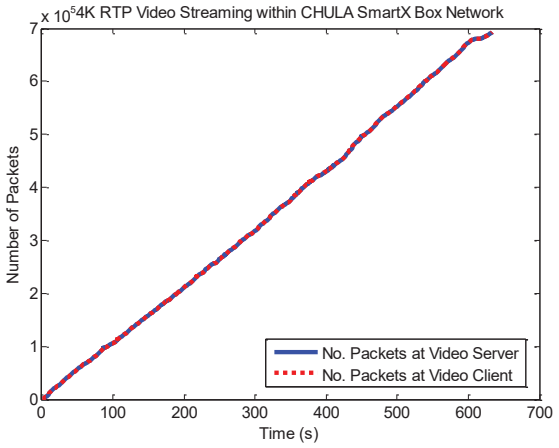


Figure 6: 4K RTP video streaming within CHULA SmartX box network.

4. CONCLUSION AND FUTURE WORK

In this multi-path chunked video exchanges experiments, we have tested the combination of traditional Tsunami protocol and proposed multi-path file-transfer function. In addition, we have implemented the X11 desktop environment and access method for OpenStack VMs in order to use the GUI applications with fast access. According to the tested results with three scenarios: multi-path (1:2sec), using Path 1 alone (GIST-B>MY>CHULA) and using Path 2 alone (GIST-B>CHULA), the proposed multi-path file-transfer method can transfer with the

minimum transmission delay when individual paths are congested and do not have enough capacity to carry the whole video file traffic. However, this multi-path file-transfer method is not recommended when the main path capacity already suffices for carrying out the incoming packets of video file traffic and the main path has a lower RTT delay than other available paths due to the results of target file transfer rates (100, 200, and 300Mbps) by using Path 2 alone. Moreover, the tested local video streaming results confirm that our implemented testbed can stream out the 4K-video within CHULA SmartX box network by using X11 desktop environment. However, we observe one fact that the VLC application itself has some limitations to play back video with very high resolution, even in normal playing back without a streaming session. Therefore, when streaming out 4K-video over cloud playground, it is recommended to adjust the video resolution scale to be around 50% (eg. resolution:1630x937) lower than the normal 4K-video resolution scale. As for future study, we plan to evaluate multi-path scenarios over OF@TEIN by using OpenFlow 1.3 group functions.

REFERENCES

[1] ONF Market Education Committee, and others, "Software-defined networking: The new norm for networks," *ONF White Paper.*, Palo Alto, US, Open Networking Foundation, April, 2012.

[2] N. McKeown, T. Anderson, H. Balakrishnan, G. Parulkar, L. Peterson, J. Rexford, S. Shenker, J. Turner, "OpenFlow: enabling innovation in campus networks," *ACM SIGCOMM Computer Communication Review 38.*, pp. 69-74, April, 2008.

[3] OpenStack [Online]. Available from : https://www.openstack.org/ [2016, July]

[4] C. Huang, C. Nakasan, K. Ichikawa, and H. Iida, "A multipath controller for accelerating GridFTP transfer over SDN," *in Proc. of 2015 IEEE 11th International Conference on IEEE.*, pp. 439-447, August 31, 2015.

[5] Y. Gu, and R.L. Grossman, "UDT: UDP-based data transfer for high-speed wide area networks," *in Proc. of Computer Networks 51.*, pp. 1777-1799, May 16, 2007.

[6] M.R. Meiss, "Tsunami: A high-speed rate-controlled protocol for file transfer," *Indiana University*, 2004.

[7] P.M. Thet, P. Panwaree, J. Kim, and C. Aswakul, "Design and functionality test of chunked video streaming over emulated multi-path OpenFlow network," *in Proc. of Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI_CON), 12th International Conference, IEEE.*, pp. 1-6, June 24-26, 2015.

[8] P.M. Thet, N. Tientrakul, J. Kim, and C. Aswakul, " Emulated OpenFlow based experimental study on middle-box buffering effect for multi-path chunked video streaming," *in Proc. of 30th International Technical Conference on Circuit/Systems Computers and Communications (ITC-CSCC).*, pp. 184-187, June 29- July 2, 2015.

[9] A.C. Risdianto, N.L. Kim, J. Shin, J. Bae, M. Usman, T.C. Ling, P. Panwaree, P.M. Thet, C. Aswakul, N.H. Thanh, A. Iqbal, U. Javed, M.U. Ilyas, and J. Kim, "OF@TEIN: A community efforts towards open/shared SDN-Cloud virtual playground," *in Proc. of the Asia-Pacific Advanced Network 40.*, pp. 22-28, August 10-14, 2015.

[10] OpenStack Compute, Nova [Online]. Available from : https://wiki.openstack.org/wiki/Nova [2016, July]

[11] OpenStack Networking, Neutron [Online]. Available from : https://wiki.openstack.org/wiki/Neutron [2016, July]

[12] R. Sherwood, G. Gibb, K.K. Yap, G. Appenzeller, M. Casado, N. McKeown, and G. Parulkar, "FlowVisor: A network virtualization layer," [Online], 2009 October. Available from : http://openflowswitch.org/downloads/technicalreports/openflow-tr-2009-1-flowvisor.pdf [2016, July]

[13] POX Wiki [Online]. Available from : https://openflow.stanford.edu/display/ONL/POX+Wiki [2016, July]

[14] VLAN (802.1q) configuration program (vconfig) [Online]. Available from : http://linux.die.net/man/8/vconfig [2016, July]

[15] LXDE [Online]. Available from : http://lxde.org/ [2016, July]

[16] Xrdp [Online]. Available from : http://www.xrdp.org/ [2016, July]

[17] Tightvnc [Online]. Available from : http://www.tightvnc.com/ [2016, July]

[18] Xorg [Online]. Available from : https://www.x.org/wiki/ [2016, July]

SESSION 5

SERVICES AND IMPLEMENTATION-RELATED ISSUES

S5.1 Implementation of tele-rehabilitation system combined with video call center.*

S5.2 Intricacies of implementing an ITU-T X.1303 cross-agency situational-awareness platform in Maldives, Myanmar, and the Philippines.*

S5.3 A community-driven information system to develop next generation collaborative and responsive rural community (NCoRe).

S5.4 Toward authenticated caller ID transmission: the need for a standardized authentication scheme in Q.731.3 calling line identification presentation.*

IMPLEMENTATION OF TELE-REHABILITATION SYSTEM COMBINED WITH VIDEO CALL CENTER

Kotaro Suzuki†, Yoshitoshi Murata††

† Graduate School of Software and Information Science, Iwate Prefectural University Graduate School
Takizawa, Iwate, 020-0693 Japan
††Faculty of Software and Information Science, Iwate Prefectural University
Takizawa, Iwate, 020-0693 Japan

ABSTRACT

Japan's low birthrate and rapidly aging population are causing medical expenses to take up ever more of the national budget and leading to a shortage of young medical professionals. As a result, rehabilitation therapy is being shifted from hospital-care to home-care. Several other countries will also face the same situation in the near future. Thus, we propose a tele-rehabilitation system combined with a video call center to make up for the shortage of rehabilitation therapy done by visiting physiotherapists. A video call center operator coaches a patient instead of a physiotherapist, and a physiotherapist supervises multiple operators. The system focuses on cerebrovascular patients who have a home-visit rehabilitation or an outpatient one and uses Microsoft KINECT to measure strain of the upper body. In this paper, implementation of this system is mainly described.

Keywords— Rehabilitation, Tele-rehabilitation, KINECT, low birthrate, rapidly aging population

1. INTRODUCTION

Japan's low birthrate and rapidly aging population are causing medical expenses to take up ever more of the national budget and leading to a shortage of young medical professionals. To suppress this increase in medical expenses, medical treatments, including rehabilitation, are being shifted from hospital-care to home-care. The amount of rehabilitation therapy in a home done by a visiting physiotherapist is limited by law and is insufficient for patients to recover completely. Several other countries will also face the same situation in the near future.

Thus, we proposed a concept of tele-rehabilitation system combined with a video call center (TRS) to make up for the shortage of rehabilitation done by visiting physiotherapists [1]. A call center operator coaches a patient instead of a physiotherapist; and a physiotherapist manages multiple operators as a supervisor.

Forty-four percent of rehabilitation patients suffer from cerebrovascular diseases as shown in Figure 1 [2]. These diseases also have the longest rehabilitation term as shown in Figure 2 [2]. We focus mainly for cerebrovascular patients who have one side paralysis and are in outpatient or home-visit rehabilitation.

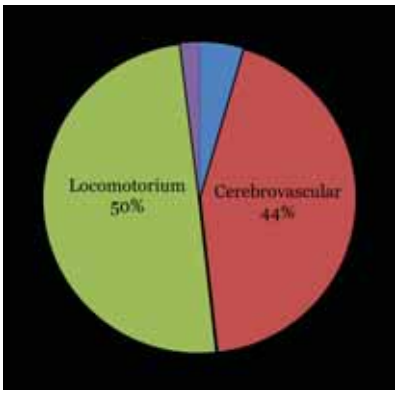


Figure 1. Ratio of disease for rehabilitation

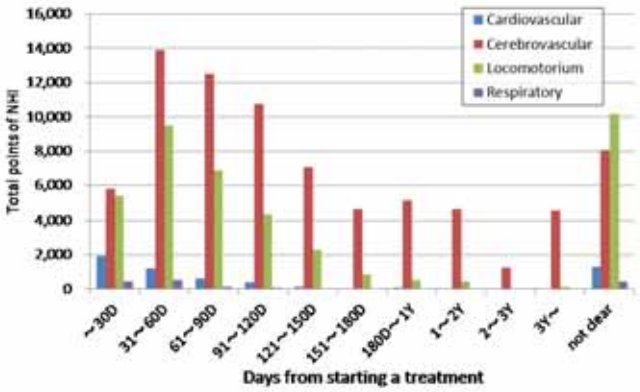


Figure 2. Total points of the national healthcare insurance (NHI) according to the duration of feeding period from treatment start date in Japan

It is very difficult for patients to continue the self-rehabilitation at home, so our system has two features to help them continuously:

- A patient can check data to see the effect of rehabilitation. This time, we use Microsoft KINECT [3] to measure strain of the upper body.
- A call center operator guides patients through the therapy and encourages them with conversation through the Internet.

We believe that patients should see practical data showing them getting better and hear a person's voice to improve their morale and to motivate them to continue rehabilitation.

In our proposed tele-rehabilitation system, a call center operator guides patients instead of a physiotherapist. Hence, our system has supervising functions for a physiotherapist to coach a call center operator like supervising functions in existing voice call center.

After describing related works in Section 2, we introduce outlines and functions of TRS in Section 3. Implementation is described in Section 4. We consider effect for shortage of physiotherapist and problems in Section 5. The key points in this paper are summarized in Section 6.

2. RELATED WORK

In this section, we introduce existing remote rehabilitaion systems and Microsoft KINECT usage applications adopted in rehabilitation.

2.1 Remote rehabilitation

Traditionally, tele-rehabilitation has been administered between a therapist and a patient through a video conference system or video phones, without using measuring and monitoring applications [4]. In accordance with the evolution of remote monitoring applications, robotics and virtual reality technologies have been combined with video conference systems. Holden et al. applied virtual reality technologies to their tele-rehabilitation system [5]. Carignan reported a rehabilitation system for which robotics was applied including remote rehabilitation [6]. Bradley et al. reported investigations of the design, control and implementation of a form of the intelligent exoskeleton, web-based strategies and robotics for remote rehabilitation [7]. In this research, therapists directly guide or coach patients through their systems. Therefore, existing tele-rehabilitation systems can shorten convey time for a visiting therapist. However, these systems are insufficient to make up for the shortage of therapists.

2.2. Therapy contents adopted Microsoft KINET to rehabilitation

Garrido et al. applied Microsoft KINECT rehabilitations for patients who have trouble with their sense of balance with cerebrovascular disease [8].

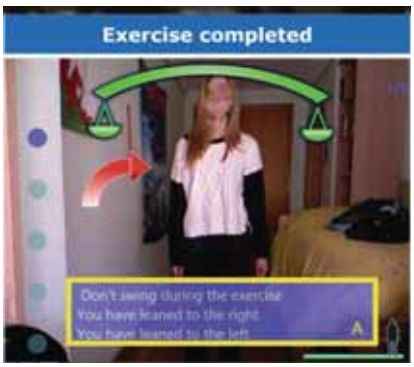


Figure 3. Screen shot of the rehabilitation system for balance disorder patients described in [8]

They express the lean of the body by an image of the balance scale and show arrows to correct a patient's posture as shown in Figure 3.

There are also many video games for rehabilitation that use Microsoft KINECT [9][10][11]. These contents can measure and record performance of a virtual game and show it to a patient to motivate him or her to continue rehabilitation.

3. CONCEPT OF TRS WITH VIDEO CALL CENTER

Our TRS is based on the following ideas:

- Practical data that shows the patients getting better will effectively encourage them to continue rehabilitation, more than simply giving them vague information such as "you are a little better than yesterday".
- Hearing a person's voice is likely to cheer patients up.

Additionally, we plan to employee non-professionals as operators who work in their home instead of physio-therapists to hold down operation costs and compensate for a shortage of physiotherapists.

We introduce roles of a physiotherapist and operator, and necessary functions to realize above concepts.

3.1. Roles of a physiotherapist and operator

Roles of a physiotherapist are as follows:

- **Teaching** operators how to guide patients through rehabilitation and **supervising** the operators.
- **Deciding** and **changing** therapy programs on the basis of diagnostics data and measured data.

Roles of an operator are as follows:

- **Monitoring** motions of a patient and **measuring** joint angles by the measuring applications.
- **Coaching** a patient in how to move his or her body using the administration tools and therapy contents on the basis of therapy programs.

3.2. Providing features

As shown in Figure 4, this system comprises following components;

- Administration tools: An operator uses these tools to guide patients.
- Measuring tools: An operator uses these tools to measure conditions of a patient including joint angles. Kinds of measuring tool that are kind of sensors are easily added to the TRS as necessary.
- Supervising tools: A physiotherapist uses these tools to coach operators. A physiotherapist accesses the patient database to check measured

data and coaching video, instructs an operator. And also, a physiotherapist revises therapy programs based on measured data.

- Communication exchange program: This application connects a patient with an operator. And, a supervisor can monitor an operator coaching a patient with this program.
- Therapy contents: Presentation contents to explain how to train, or training content such as video games for rehabilitation.
- Patient database: Patient data that include profile data, measured data, therapy programs, and coaching video are stored and managed by this database. The access permission policy for this database has to be decided by the management organization of this system.

The supervising tools and communication exchange program are newly added to introduce operators to the tele-rehabilitation system. The administration and measuring tools are used by an operator, and the supervising tools are provided to a physiotherapist as a supervisor.

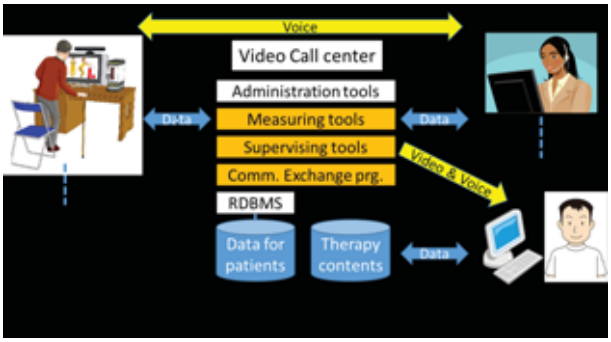


Figure 4. System concept of the remote rehabilitation

4. IMPLEMENTATION

We have already developed TRS proposed in Section 3 but not its administration tools. In this section, we explain how to implement the above components other than the administration tools.

4.1. Design policy

Patients are spread over a wide area, and some live in rural areas. The operators are also spread over wide area. Therefore, it is difficult to maintain patients' PCs and operators' PCs by visiting them. Thus, we adopt the Web application as possible to realize the above tools. And also, the Web application can store personal information safely. Moreover, we adopt as many open source programs as possible to suppress system cost.

4.2. Communication exchange program

Both a voice communication path and a data communication path have to be established between a patient-PC and an operator-PC for an operator to monitor movements of a patient and coach her or him verbally. A sequence flow establishing these two paths is shown in Figure 5.

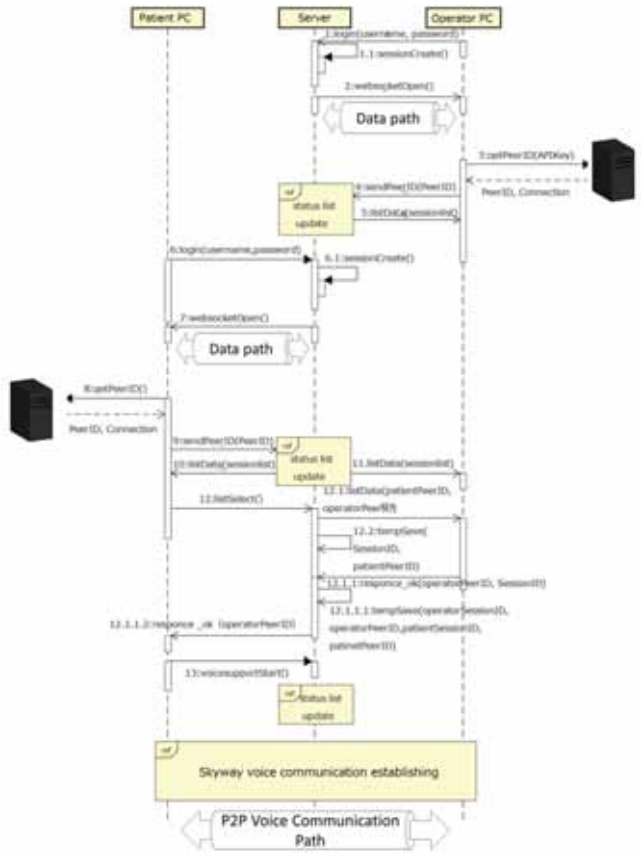


Figure 5. Sequence flow to establish paths between a patient and an operator

Since it is better for a patient to select an operator than vice versa, operators log in to the call center server prior to a patient choosing an operator. A patient then clicks a button to choose one of the logged in operators. After an operator logs in to the video call center server, the data path is established between the operator-PC and the video call center server. This data path between a patient-PC and the video call center server is established the same way.

As explained in the next subsection, we use Microsoft KINECT to measure strain of the upper body of a patient. Since Microsoft KINECT is difficult to control by using a Web-browser, we developed a KINECT control application as a native application, not a Web application. Data measured by Microsoft KINECT are sent to the video call center by using the Web socket and converted into Web contents.

We use WebRTC [12] and SkyWay server [13], which is a WebRTC platform, to establish a voice path between a patient-PC and an operator-PC. Each PC for a patient and an operator accesses the SkyWay server to obtain a peer ID and sends it to the video call center server before

establishing a voice communication path between a patient-PC and an operator-PC. After a patient selects an operator, the video call center server matches the operator_PeerID with the patient_PeerID. The SkyWay client_libraries on each PC establish a voice communication path between a patient-PC and an operator-PC in combination with the SkyWay server.

4.3. Measuring tool

After data and voice paths are established between a patient and an operator, the operator requests the patient to start the Microsoft-KINECT v2 application so that the operator can monitor the patient and data can be measured. This sequence flow is shown in Figure 6.

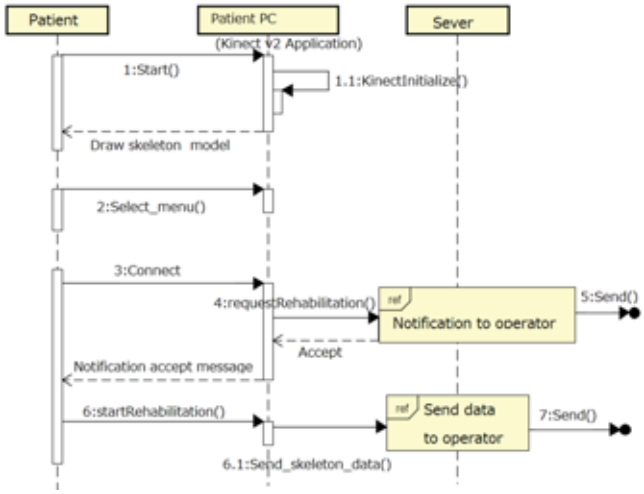


Figure 6. Sequence flow to start Microsoft KINECT

Since video images and skeleton data are independently output from Microsoft KINECT, both have to be synchronized to enable monitoring at an operator's office as shown in Figure 7. We thus serialize skeleton data in the JSON format and add them to each item of video frame data, which is resized, transferred to the JPEG format, and then transferred to the Base 64 text data at the application on a patient's PC, as shown in Figure 8. These data are sent to a Web server in the video call center. They are separated again and individually drawn on the canvas of the html monitoring page, as shown in Figure 9. Video image data are drawn on a lower layer and skeleton data are drawn in our presentation format on a transparent upper layer, as shown in Figure 10.

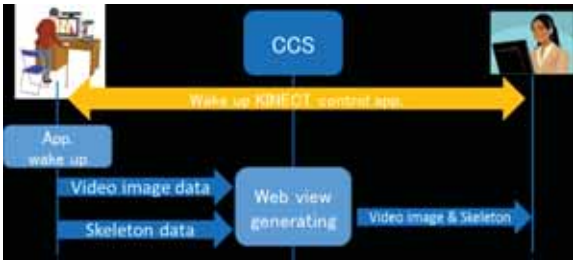


Figure 7. Monitoring scheme of patient's motion

We applied this system to the stand-up and sit-down training, as shown in Figure 11. Both a patient's display and an operator's one when a patient is doing stand-up and sit-down training are shown in Figure 12. We present strain of the upper body by using three view formats on an operator's display: the front view, the side view, and the top view.

An operator selects a view format and clicks a button to measure data. These data are temporally stored as a file. After an operator decides to save them on the server, the file is stored as an archived file. This sequence flow is shown in Figure 13. An operator can also bookmark selected frames.

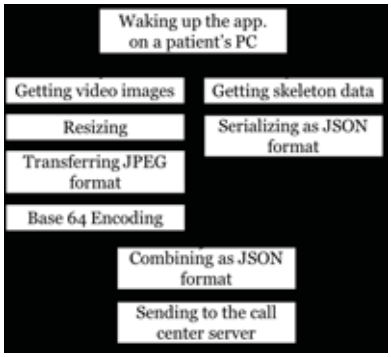


Figure 8. Flow diagram for aggregating video images and skeleton data

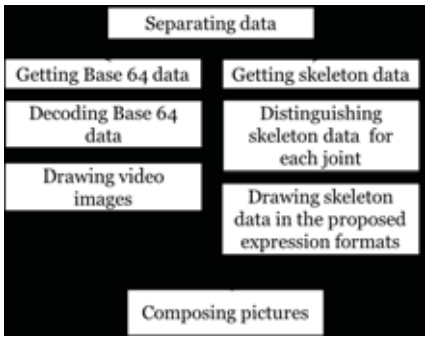


Figure 9. Flow diagram for composing pictures

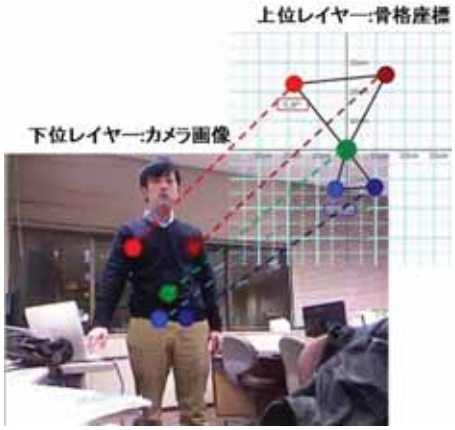


Figure 10. Example of composing video image and skeleton data

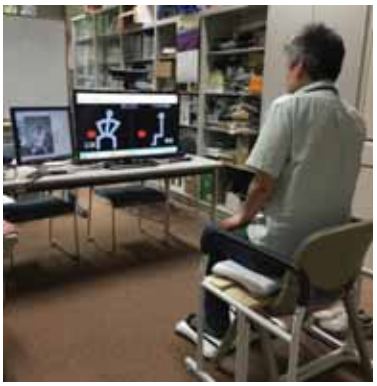


Figure 11. Stand-up and sit-down training

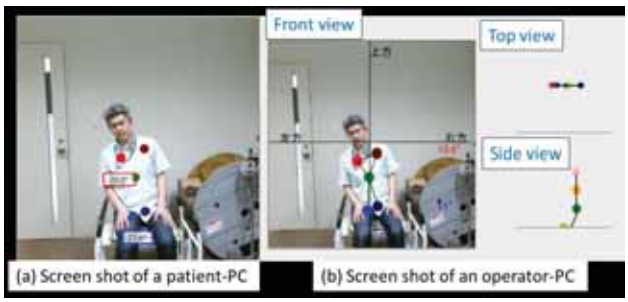


Figure 12. Screenshots from a patient-PC and an operator-PC

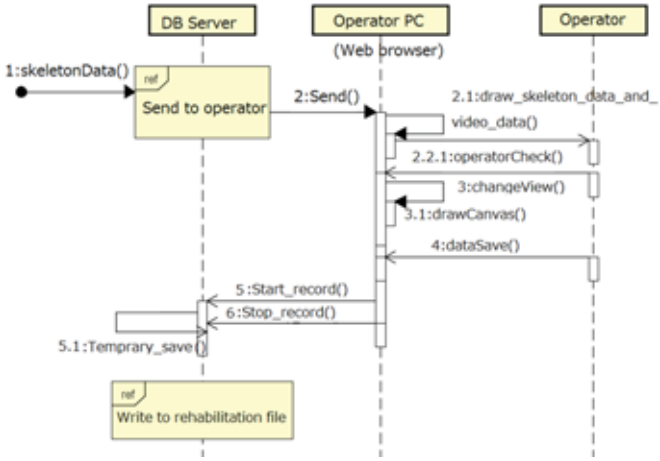


Figure 13. Sequence flow to measure data

4.4. Supervising tools

Since physiotherapists do not coach patients directly using this tele-rehabilitation system, we have developed the following four supervising tools to help physiotherapists to supervise operators;

- Monitoring tool: This tool enables a supervisor (physiotherapist) to watch video images on an operator's display and listen to conversations between a patient and an operator to find out how the operator is coaching the patient.
- Recording tool: This tool records video images on an operator's display and conversations between the patient and operator and stores them in a database.
- Searching tool: This tool enables an operator to search for objective measured data and recorded video images and conversations.
- Replaying tool: This tool replays selected video images and conversations.

4.4.1. Monitoring tool

A supervisor selects an operator and patient pair from the coaching list shown in Figure 14 and clicks the join button. A supervisor can see which operator is coaching which patient by looking at this list. A voice path is established between a supervisor and an operator with the sequence flow shown in Figure 5. After a supervisor logs in, the monitoring tool on the supervisor-PC obtains a peer ID from the SkyWay server and sends it to the video call center server. The video call center server sends the operatorPeerID to the supervisor-PC in accordance with the request from the supervisor-PC. After that, the monitoring tool on the supervisor-PC establishes a voice path to an operator-PC in cooperation with the SkyWay server and displays the same screen that an operator is viewing (e.g. Figure 12 (b)).

Figure 14. Supervisor's page after login

4.4.2. Recording tool

Video images and conversation of an operator coaching a patient are automatically stored in the video call center server as archived files. Directory paths to all files are managed by the database management system. The structure of this database management system is described in the next sub-section.

4.4.3. Searching tool

Since measurements of the patient, video images on an operator's display, and conversations are large data, they are stored as files. Moreover, directory paths to all files are managed by the database management system. Measured data and archived data are featured by the following keys:

- Auto_ID.
- Stored date: time stump for storing a file.
- Patient_ID.
- Operator_ID.
- Comment: An operator writes.
- Measured/Archived.
- Directory_path.

Detailed information about an operator and a patient is stored in separate tables. Keys of the operator table are as follows:

- Operator_ID.
- Employment_date.
- Date_of_Birth.
- Address.

- Mail_address.
- Phone_number.

The patient table has almost all the same keys as the operator table, except operator_ID is changed to patient_ID.

Frames in a file that were bookmarked by an operator are managed with the following keys:

- Auto_ID.
- Bookmark_ID.
- Start frame number.
- End frame number.

An E-R diagram for these tables is shown in Figure 15. A supervisor can search for measured/archived files with the patient_ID and/or the stored_date.

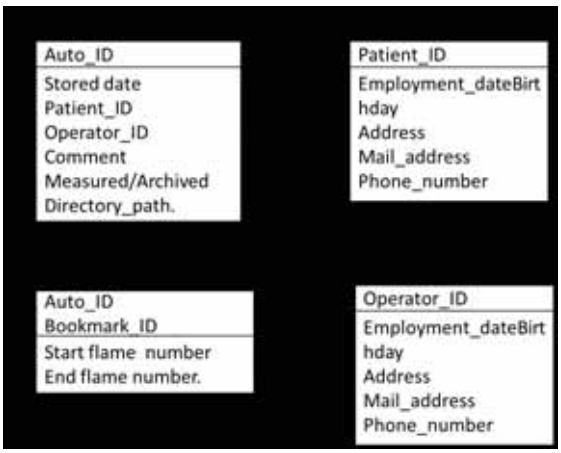


Figure 15. E-R diagram to manage files

4.4.4. Replaying tool

A supervisor can replay the content of selected files by clicking them. It is possible to select view items such as camera view, line between joints, and joint points. We made a seek bar in addition to a replay button, a stop button, and arrow buttons, as shown in Figure. 16. A supervisor can easily check measured data in the selected frame.

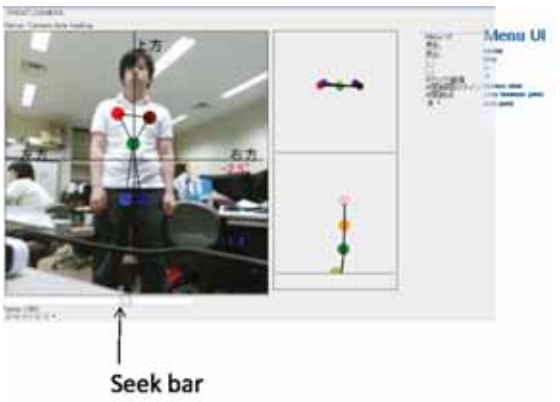


Figure 16. Screenshot of the replaying tool

5. CONSIDERATION

We consider how our system makes up for a shortage of physiotherapists. Since TRS is useful for patients who undergo home-visit rehabilitation, we compare TRS with home-visit rehabilitation.

In TRS, a patient is coached by an operator instead of a physiotherapist, and a physiotherapist supervises multiple operators. Moreover, an operator's salary is a fraction of a physiotherapist's one. Therefore, a physiotherapist's salary can be used to pay multiple operators to coach multiple patients in TRS.

The number of patients that can be coached using a physiotherapist's salary in TRS Nps is as follows:

$$N_{ps} = N_m * N_o / (1 + N_m * S_o/S_p) \quad (1)$$

where N_m is the number of operators that a physiotherapist supervises, N_o is the number of patients that an operator coaches in one hour, S_p is a physiotherapist's salary, and S_o is an operator's salary.

We assume that a physiotherapist and an operator coach a patient for 90 and 45 minutes, respectively. Therefore:

$$N_o = 2 * N_p \quad (2)$$

N_p is the number of patients that a physiotherapist coaches in one hour. The ratio between N_{ps} and N_p is as follows:

$$N_{ps}/N_p = 2 * N_m/ (1 + N_m * S_o/S_p) \quad (3)$$

N_{ps}/N_p is calculated and shown in Figure 17. When a physiotherapist supervises five operators and his or her salary is five times that of an operator, five times as many patients are coached in TRS as in home-visit rehabilitation.

TRS is very useful for overcoming a shortage of physiotherapists. However, a patient has to pay about $1500 to buy a PC and install Microsoft KINECT. It is difficult to evaluate whether this is expensive or not because a PC can be used for other tasks, and a physiotherapist needs an automobile to visit a patient for home-visit rehabilitation.

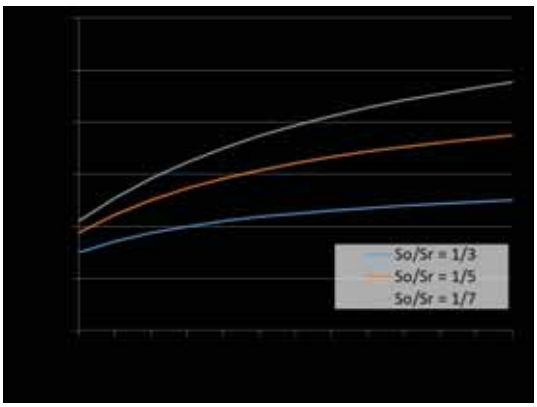


Figure 17. Effect for shortage of physiotherapist in TRS

6. CONCLUSION

We developed a tele-rehabilitation system that enables a video call center operator to coach a patient in place of a physiotherapist and a physiotherapist to supervise multiple operators. We believe the proposed system will help to suppress the increase of medical expenses and make up for the shortage of young medical professions caused by low birthrates and rapidly aging populations.

However, there are several problems to putting TRS on the market.

- For healthcare insurance to cover it, TRS has to be approved by the government department in charge of healthcare.
- The national healthcare insurance in Japan covers only face-to-face rehabilitation, not tele-rehabilitation. Thus, TRS will need to be licensed for the national healthcare insurance to cover it.
- If a license cannot be obtained, a new business model will be needed for business operators to continue this business.
- Criteria for evaluation are needed to maintain service quality. At the very least, criteria an operator has to learn are needed

Technically, we think TRS is a platform system for the tele-rehabilitation service. Therefore, many kinds of training contents or tools will be connected to TRS in a future. To connect such tools to TRS, standard protocols will be needed. We also think an operator has to speak a certain way to patients to improve their morale and to motivate them to continue rehabilitation. We plan to develop a function for evaluating emotions an operator expresses while speaking.

Because we have not found any partners to evaluate our system, we will continue looking for evaluation partners.

REFERENCES

[1] Kotaro Suzuki, Yoshitoshi Murata, and Kazuhiro Yoshida, "Remote Rehabilitation System for Cerebrovascular Patients Combined with Video Call Center," IARIA, International Journal on Advances in Life Sciences, vol. 7 no 3 & 4, pp. 132 – 1442015, 2015.

[2] Investigation into the current state of rehabilitative care, Ministry of Health, Labor and Welfare, http://www.mhlw.go.jp/shingi/2007/11/dl/s1128-14h.pdf (in Japanese), [retrieved: October, 2016].

[3] Meet Kinect for Windows, https://developer.microsoft.com/en-us/windows/kinect, [retrieved: October, 2016].

[4] David M. Brennan, Sue Mawson, and Simon Brownsell, "Telerehabilitation: Enabling the Remote Delivery of Healthcare, Rehabilitation, and Self Management," IOS Press, Advanced Technologies in Rehabilitation, 2009.

[5] Maureen K. Holden, Thomas A. Dyar, and Lilian Dayan-Cimadoro, "Telerehabilitation Using a Virtual Environment Improves Upper Extremity Function in Patients With Stroke," IEEE, Transactions on neural system and rehabilitation engineering, vol. 15, no. 1, pp. 36-42, 2007.

[6] Craig R. Carignan and Hermano I. Krebs, "Telerehabilitation robotics: Bright lights, big future?," JRRD vol. 43, no. 5, pp. 695–710, 2006.

[7] D.A. Bradley et al., "Nexos: Remote Rehabilitation Using an Intelligent Exoskeleton," Gerontechnology Vol. 3, No. 4, pp. 71-80, 2005.

[8] Juan Enrique Garrido, Irma Marset, Victor M. R. Penichet, and Maria D. Lozano, "Balance Disorder Rehabilitation through Movement Interaction," IEEE, 7th International Conference on Pervasive Computing Technologies for Healthcare and Workshops, 2013.

[9] Shih-Ching Yeh et al., "A study for the application of body sensing in assisted rehabilitation training," IEEE, 2012 International Symposium on Computer, Consumer and Control, 2012.

[10] Belinda Lange et al., "Interactive Game-Based Rehabilitation Using the Microsoft Kinect," IEEE, Virtual Reality 2012, 2012.

[11] Marientina Gotsis et al., "Mixed Reality Game Prototypes for Upper Body Exercise and Rehabilitation," IEEE, Virtual Reality 2012, 2012.

[12] WebRTC, http://www.webrtc.org/, [retrieved: October, 2016].

[13] SkyWay, https://nttcom.github.io/skyway/en/index.html, [retrieved: October, 2016].

INTRICACIES OF IMPLEMENTING AN ITU-T X.1303 CROSS-AGENCY SITUATIONAL-AWARENESS PLATFORM IN MALDIVES, MYANMAR, AND THE PHILIPPINES

Biplov Bhandari[1], Angga Bayu Marthafifsa[1], Manzul Kumar Hazarika[1], Francis Boon[2], Lutz Frommberger[2], and Nuwan Waidyanatha[2]

1. Geoinformatics Center, Asian Institute of Technology, Klong Luang, Pathumthani, Thailand
2. Sahana Software Foundation, 350 S Figueroa ST STE 437, Los Angeles, CA 90071, USA

ABSTRACT

Maldives, Myanmar, and the Philippines are vulnerable to natural disasters [1]. Sendai Framework[1] of Action calls for risk reduction by implementing early warning systems [2]. A prevailing challenge is for authorities to coordinate warnings across disparate communication systems and autonomous organizations [3]. Cross-Agency Situational-Awareness platforms and the ITU-T X.1303 Common Alerting Protocol (CAP)[2] interoperable data standards presents themselves as solution for diluting the inter-agency rivalries and interconnection disparities [4]. The Sahana Alerting and Messaging Broker (SAMBRO) was designed to overcome these issues by providing a Common Operating Picture and a platform for all Stakeholders to share early warnings. To that end, the CAP-on-a-MAP project is implementing SAMBRO and the CAP standard along with the policies and procedures in the Maldives, Myanmar and Philippines. The project is applying an agile development methodology with a design, build, test, and redesign strategy for implementing the cross-agency situational-awareness and warning system in the respective countries. This paper discusses the country context implementation challenges and discusses strategies fostered through the introduction of the CAP content standard for warning system designers to consider for overcoming similar challenges.

Keywords— interoperability, electronic services, early warning, common alerting protocol, situational awareness

1. INTRODUCTION

Disasters are a major problem worldwide and a serious threat to sustainable development. The rapid and often unplanned expansion of human settlements, especially in cities, is exposing more people and economic assets to the risk of disasters and the effects of climate change [5]. Coastal cities are made vulnerable by the low-lying land that are often built upon and as such are susceptible to flood, storm surge, tsunami, and sea-level rise [6]. Many coastal cities in Asia and the Pacific region are found in tropical areas with hot and humid climates and low-lying land, both of which heighten their vulnerability to extreme events.

As a consequence of climate change, the world is facing an increasing threat of extreme events. Especially in developing countries, this heavily affects equal access to opportunities and development and is a main reason for poverty. As a result the Sustainable Development Goals (SDGs)[3] emphasis on the need to reduce poverty and inequalities. It also emphasizes on strengthening climate actions and sustainable economies, communities and cities [7]. Early warning systems for disaster response are critical to fighting the consequences of climate change contributing to the factors that influence the SDGs.

Disaster Risk Management interventions such as alerting/early warning, evacuation planning, and coastal zone management are important for addressing the challenges faced by the coastal communities. The project titled "CAP on a Map" was designed to improve the institutional responsiveness to coastal hazards in Maldives, Myanmar, and the Philippines. It would augment the capabilities of the National Disaster Management Organizations (NDMOs), National Warning Centers (NWCs), line-agencies and other relevant stakeholders in disaster management to interchange and share early warning information.

A prevailing challenge is for NDMOs, NWCs, line-agencies and relevant stakeholders to coordinate and interchange alerts and warnings. To overcome this dilemma, the CAP on a Map project introduced a Cross-Agency Situational Awareness platform for coordinating alerts and warnings and the Common Alerting Protocol (CAP) content standard for interchanging warning messages across disparate systems. These are the underpinning design concepts of the Sahana Alerting and Messaging Broker (SAMBRO). In this

[1] "Sendai Framework for Disaster Risk Reduction - UNISDR." 2015. 11 Jul. 2016 <http://www.unisdr.org/we/coordinate/sendai-framework>

[2] "Common Alerting Protocol - Oasis." 2015. 11 Jul. 2016 <https://docs.oasis-open.org/emergency/cap/v1.2/CAP-v1.2-os.html>

The "CAP on a Map" project was made possible through the United Nations Economic and Social Commission for Asia and the Pacific Trust Fund.

[3] "Sustainable development goals - United Nations." 2015. 11 Jul. 2016 <http://www.un.org/sustainabledevelopment/sustainable-development-goals/>

paper we discuss the Maldives, Myanmar, and Philippine case studies, their warning requirements and the strategy for customizing SAMBRO and CAP for them.

This work was made possible through the United Nations Economic and Social Commission for Asia and the Pacific Trust Fund for Tsunami, Disaster, and Climate Preparedness.

2. TECHNOLOGY

The technology provides an overview of the CAP version 1.2, the Sahana software suite, and the SAMBRO web and mobile applications.

2.1 Common Alerting Protocol

CAP is an OASIS (Organization for the Advancement of Structured of Information Standards) advocated Emergency Data Exchange Language (EDXL) content standard. CAP is designed for all-hazard all-media warnings; a standard that is recommended by the ITU-T (2008) documented as X.1303. The World Meteorological Organization and the International Federation of Red Cross and Red Crescent are also key advocates of the standard. Google Crisis Response offers to publish NWC generated CAP feeds through their products including Google Public Alerts. Federation of Internet Alerting is a consortium of online advertising agencies that have extended the service of rendering alerts on their online ad-spaces. Meteo-alarm and Accuweather are other, among several, online services that help Nations publicize CAP messages.

Figure 1 shows the CAP Document Object Model. It is essentially a XML document that inherits the interoperability aspects of the XML technology. The data structure consists of a main node element <Alert> and its sub elements (node) <Info>, <Area>, and <Resources>. Each of the sub elements are composed of several other elements. The ITU-T X.1303 document defines each of these elements with respect to semantic interoperability. X.1303, is broadly recognized internationally as the key standard to achieve the goal of all hazards, all-media public alerting.

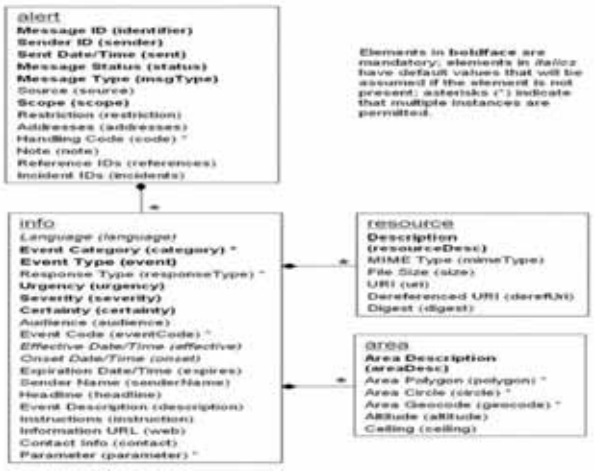


Figure 1: CAP Document Object Model

2.2 Sahana Eden Platform

Sahana is a collection of disaster management modules that work as a platform for integrating multi-organizational response efforts in providing critical information and communication needs. It advocates international data standards, it is internationalized and localized [8]. Sahana software is a wrapper around the Python Web2Py software development framework. Sahana strictly follow the HTTP standard and RESTful concept, making it is easier for third party application to add/edit/delete the information. Sahana follows a Model, View, Controllers (MVC) architecture. The code-base is hosted in GitHub and free to use and edit under the MIT license.

2.3 Sahana Alerting and Messaging Broker (SAMBRO)

SAMBRO, is a specialized Sahana solution (i.e. a Sahana template), designed with CAP version 1.2 as the underlying interoperable data standard to serve as a warning and situational-awareness tool. The original SAMBRO (version 1.0) design, as described in [9], was transformed with newer features. The current SAMBRO version 2.0 builds on a decade of action research as discussed in [10], [11], and [12]. SAMBRO is designed to serve as a CAP message publisher, aggregator, and a disseminator.

2.3.1. Control

Implementers define the metadata and ready the system defining the event types, warning classification, predefined alerting areas, CAP message templates, and grants user permissions. Publishers are authorized users who creates CAP warning messages in consent of their seniors and approve for their dissemination. Subscribers are authenticated users with minimum roles allowing them to subscribe to receive warnings of their choice.

The access control and permission of SAMBRO allows us to control the access of the application at different level. The permission access can be controlled at the module levels, particular table levels, functional levels, and/or individual record level. This is a much needed use-case for the warning, as there are many data which are only shared at organization level or between some closed user group and not to public.

SAMBRO's Audit trail is another distinguishing feature in the system. The system keep records of who logged into the system, who created the alert, when it was created, when it was submitted for approval, who approves it, when it was approved and many of the essential information that are needed for the audit purpose. After an alert is issued, a snapshot of the record independent of all the external references are kept in another table.

2.3.2. Features

All workflows are guided by the initial selection of the event type. Cyclone, Flood, Earthquake, Mass Movement, Civil Unrest, are examples of event types. They are the high

level category of disaster. The system adopts a set of rules for each event type, such as filtering the set of relevant CAP message templates, warning classification, predefined alert areas, and message recipients. The warning classification are for message recipients to summarize the severity, certainty, and urgency CAP properties of the warning message. The warning classification is not to be confused with the scale of the hazard such as the Saffir-Simpson Hurricane wind scale[4].

CAP message templates are predefined CAP messages. It allows for the Implementers to generate consistent messages that can be changed to suit the context at the time of issuing the warning. SAMBRO CAP templates are generally defined for Cyclone, Flood, Heavy Rain, Earthquake, Tsunami, and Landslide, so on and so forth. While most of the CAP elements would carry predefined text some elements such as the CAP message "description" would contain blanks that need to be filled in based on the event specifics.

2.3.3. Workflows

There is a two stage workflow for warning message dissemination. The Editor (or CAP message author) creates/edit the alert, and requests for an Approver to verify the message and approve for dissemination. Approvers receive an Email and a SMS with a URL pointing to the message awaiting approval. The Approver does not need to be at an office table to use a Personal Computer. They can be anywhere and use their Smartphone, provided they have Internet connectivity, to approve the warning message.

Relying warnings is a common practice among NDMOs and Response Organizations. CAP makes it easy to implement a warning message relay workflow with SAMBRO. When a NWC issues a bulletin (alert) or a warning, the NDMO can relay that message. SAMBRO is capable of subscribing to National and International CAP feeds. These are, typically, through RSS feeds but has the option of integrating specific APIs. When a message is received through an external source, SAMBRO offers a feature to relay the message allowing the Authority to make changes to the original message before disseminating through the SAMBRO communications engine.

2.3.4 Communications

The single entry of a CAP message can be disseminated, to the end recipients, through multiple channels. Short Message Service (SMS), Email, File Transfer Protocol (FTP), Real Simple Syndication (RSS), Google Cloud Messaging (GCM), Websites, and social media like Facebook, twitter are the available communication channels. SAMBRO uses a simple Extensible Stylesheet Transformation (XSLT) to produce the text outputs for the various media channels.

[4] "Saffir-Simpson Hurricane Wind Scale - National Hurricane Center." 2012. 5 Jul. 2016 <http://www.nhc.noaa.gov/aboutsshws.php>

SAMBRO also acts as an Alert Hub. The advantage of using any protocol is the standardization of the disseminated message from the system. So any CAP 1.2 implemented system can interact with SAMBRO system. Each published CAP compliant warning from other organization can be imported into the local database. More exciting is that those warnings coming from other organization can be relayed to the respective organization. Here SAMBRO has bridge the gap between the inter-organizational communications.

2.3.5. SAMBRO Web

Figure 3 show the Common Operating Picture gives an overview of "What is happening and Where and When". The GIS enabled mapping provides more interaction with the audience. The warnings can be filtered to move into the area of interest. Each warning has its own main profile that overview the associated alerting qualifiers, detailed information relating to warning, any instructions, descriptions and many more.

The SAMBRO Web Application is localized allowing the user to choose their choice of the Graphic User Interface (GUI) language. The localization is fairly easy under Sahana because of the multi-language capabilities the Web2Py platform offers.

Figure 2: SAMBRO Mobile App for the first responders

2.3.6. SAMBRO Mobile

The SAMBRO Mobile can run on both Android and iOS Smartphones. The mobile application was developed with the Cordova based "PhoneGap" with HTML, CSS and JavaScript; allowing the mobile-app to be independent of the operating system: Android, iOS or Windows. The mobile-app adopt GCM for pushing messages in real-time

onto the phones. This requires a dedicated internet connection. Future version is looking at transporting the information over SMS.

The SAMBRO Mobile is for local Alerting Authorities to issue localized alerts. For example, be able to quickly warn a local village of a chemical spill or a large factory fire opposed to waiting for central national authority to issue the same. It was also designed to serve as a wakeup call with an audible siren to wakeup first-responders in the middle of the night or to get their attention if they were at distant from the device.

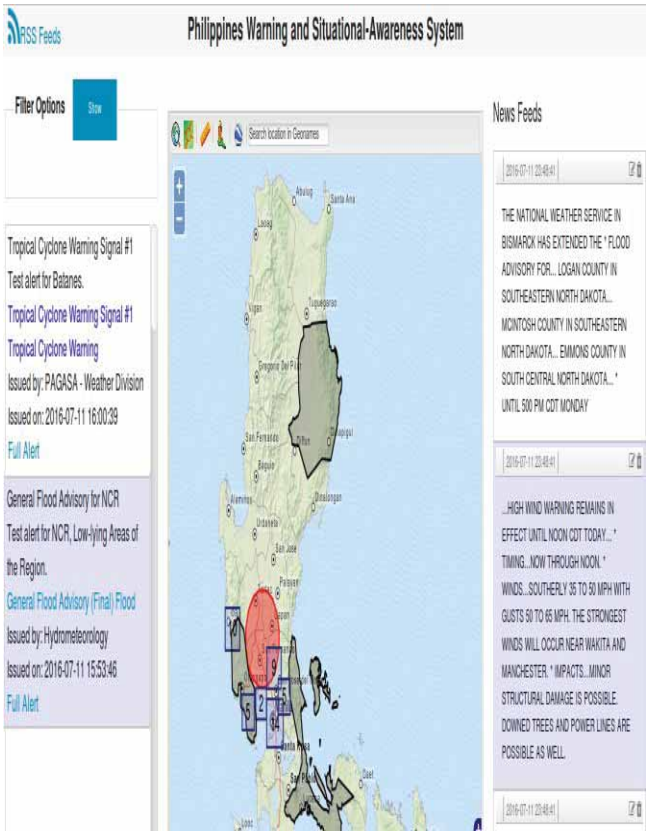


Figure 3: SAMBRO Web Application for Disaster Management

3. METHODOLOGY

The design and implementation methodology considered a scientific approach. [13] discuss the intricacies of applying a waterfall method with well-defined elicitation and documentation of complete requirements, followed by an architectural and high-level design development and inspection. Given the dilemma of introducing CAP to novices who haven't been exposed to the content standard but are engaged in early warning, it is cumbersome to follow a plan-driven method because they can become frustrating to the users and the implementers. We identified four related areas of study and practice within the broad field of information system design. These four areas are (1) User-Centered Design (UCD); (2) rapid prototyping; (3) agile software design (SCRUM); and (4) action research.

UCD is "a general term for a philosophy and methods which focus on designing for and involving users in the design of computerized systems." [14]. Rapid prototyping "involves creating a realistic model of a product's user interface to get prospective customers involved early in the design of the product. With rapid prototyping, the user model, the work flows and information needs without investing the time and labor required to write actual code [15]. Thereafter, revise the prototype to address the user's comments and keep iterating the process until we agree on the design parameters before creating the product [16].

Agile software design is defined as "a lightweight software engineering framework that promotes iterative development throughout the life-cycle of the project, close collaboration between the development team and business side, constant communication, and tightly-knit teams." With the Scrum, the light-weight process framework for agile development, initially the user stories[5] are collected by involving the user. These are the wish-lists that client like to have at the end of the product development. These user stories goes into the backlog catalogue, which are placed under the order of priority. The software development begins with small chunks of the backlog catalog called as sprints. With scrum, using sprints we can build pieces of software and the client can experience each part and determine what to do next.

Action research is typically regarded as an approach to research that balances knowledge generation with planned action. According to Hearn & Foth [17], "action research not only aims to understand the problem, it aims to provoke change through actionable outcomes." It is also known by other related terms such as participatory action research, collaborative inquiry, emancipatory research, or action learning [18].

The information, gathered from the preliminary stakeholder rapid prototyping interactions, was compiled to produce a set of customer requirements. The requirements were transformed to a set of User Experience (UX) and User Interface (UI) designs to customize SAMBRO. Selected participants from the three countries were invited to trial the preliminary release of SAMRBO.

Subsequently all stakeholders from each country were invited to experiment with SAMRBO. This, second iteration of the rapid prototyping, was, once again, used to further enhance SAMBRO. Throughout this cycle the lead Organizations engaged in weekly and monthly interactions with the development team to discuss their requirements and test the system. The lead Organization liaised with the stakeholders to update them on the new developments as well gather information on their requirements.

[5] User stories are documented in the Sahana Wiki Blueprints for: Maldives (http://eden.sahanafoundation.org/wiki/BluePrint/CAPBroker/Maldives), Myanmar (http://eden.sahanafoundation.org/wiki/BluePrint/CAPBroker/Myanmar), and Philippines (http://eden.sahanafoundation.org/wiki/BluePrint/CAPBroker/Philippines)

4. DISCUSSION

4.1 Stakeholder Engagement

The project partnered with a National Government Organization in each country to lead with the multi-stakeholder participatory approach with the design and implementation. They were the Maldives National Disaster Management Center (NDMC), Myanmar Department of Meteorology and Hydrology (DMH), and the Philippines Atmospheric Geophysical and Astronomical Administration Service (PAGASA). Table 1 identifies the list of SAMBRO Cross-Agency Situational-Awareness platform primary beneficiaries.

4.1.1. Warning Practices

Currently, the warning dissemination practices in the three countries are similar. Typically, the NWC issues a bulletin that is received by the NDMO and other focal agencies. The NDMO would use a phone tree to disseminate the messages from the National nodes to the branches and then eventually to the leaf nodes. The method is reliable but slow and laborious because they use telephones, fax, and VHF radios. Some social media and SMS is used. Such practices are acceptable for slow onset hazards such as disease outbreaks, cyclones and floods that provide a long warning horizon. Present day warning dissemination practices in the three countries are inefficient in serving rapid and sudden onset hazards such as dam burst, tsunami, or storm surge, with very short warning horizons. SAMBRO removes the laborious paper and manual hierarchical tree structure and provides a hub and spoke architecture. The SAMBRO software works much faster than human-based procedures involving multi-phase message relays.

Maldives NWC: Maldives Meteorological Service	**Maldives NDMO**: National Disaster Management center
Maldives Response Orgs: Maldives National Defense Force (Coast Guard and Fire Search and Rescue Department); Maldives Red Crescent Society; Local Atoll Councils; Local Island Councils; Maldives Police; Maldives Red Crescent Society; Department of Health; Ministry of Tourism Arts and Culture; Ministry of Education	
Myanmar NWC: Department of Meteorology and Hydrology	**Myanmar NDMOs**: Relief and Resettlement Department; General Administration Department;
Myanmar Response Orgs: Department of Irrigation; Department of Health; Department of Agriculture; Department of Fisheries; Department of Inland Transportation; Fire Services Department; Myanmar Red Cross Society	
Philippines NWC: PAGASA; Philippines Institute of Volcanology and Seismology	**Philippines NDMO**: Office of the Civil Defence (OCD)
Philippines Response Orgs: Department of Social Welfare and Development; Local Government Units, Disaster Risk Reduction and Management Councils	

Table 1: National beneficiary Organizations

Another inherent characteristic is the National EWS is that they have not developed an integrated approach to EWSs to foster an all-hazard approach. Each Organization is developing their own early warning dissemination mechanism forcing the public and Response Organizations to interface with multiple data feeds. For example, each agency will host their own twitter and Facebook accounts to share their information; opposed to a single source publishing warnings on all-hazards; i.e. a "one stop shop" approach.

Any entity that has privileges to edit CAP messages, can relay a received CAP message allowing them to change the content (e.g. description and instructions) and rebranding the message with qualifying elements associating them as the message sender. The relay function extends beyond messages issued through the SAMBRO system. For example, SAMBRO can receive any CAP message published by any other system in the world and offer the relay feature to the users of the SAMBRO instance. All the National stakeholders might consider collaborating to make full use of such features offered by SAMBRO that brings efficiencies to coordinating early warnings.

4.2 CAP Implementation Strategy

The common practice is for NDMOs to lead and be mandated with warning dissemination. The NWCs are mandated with detection and monitoring. Meteorological, Hydrological, and Seismological agencies, serving as NWCs, would feed daily information bulletins and hazard event information of significant interest to the NDMOs. The NDMOs would, then, transform and disseminate those messages to the closed user group alerts or public. SAMBRO was designed to serve such a workflow. Realizing, this important relationship and workflow, SAMRBO was designed for NDMOs to own and operate SAMBRO and for NWCs to interface with SAMBRO through CAP information feeds.

Even though it is in their mandate and best interest to foster such a platform, the project learned that NDMOs (Myanmar RRD and Philippines OCD) expressed interest but were less inclined to operationalizing a CAP-enabled situational-awareness platform. Maldivian NDMC proved the contrary

with operationalizing SAMBRO before MMS. NWCs were more inclined to adopting the CAP standard as in the case of PHIVOLCS and PAGASA in the Philippines and DMH in Myanmar. A primary factor might be is that NWCs realize the importance of interoperability and practice monitoring/detection and alerting on a daily basis. A second realization was the NWCs already had competent technical capacity who were able adapt to the introduced technologies. The World Meteorological Organization's program on advocating CAP might be a third factor. Although NDMOs are mandated with warning dissemination, given their detachment from practicing on a daily basis, might be the fourth factor.

To that end, the project administered a strategy that combined the NWCs in leading the project implementation alongside the NDMOs. Thereafter, the NWCs would gradually transfer the technology over to the NDMO to own and operate. In the Philippines PAGASA would support NDRRMC, Myanmar's DMH would support RRD/GAD. Maldives has now established technical competency and does not require the support of MMS.

4.3 Organizational Interactions

SAMBRO web and mobile software applications were perceived, by the users, to be useful tools for creating, disseminating, and sharing early warnings. SAMBRO was designed to dilute the inter agency rivalries and bureaucratic barriers by introducing the CAP interoperable standard for the siloed organizations to interchange lifesaving information. However, a challenge faced by the beneficiary countries is coordinating the implementation and operationalization the systems involving all relevant stakeholders. Government bureaucratic layers require formal procedures for engaging the stakeholders. Often Government Officials are reluctant or are discouraged by these bureaucratic formalities to pursue the project objectives of operationalizing a Cross-Agency Situational-Awareness platform for improving institutional responsiveness to hazards. This situation relates to the chicken and egg causality dilemma.

To overcome the dilemma the project was persistent with engaging the stakeholders through workshops. Using workshops as a platform reduces the need for bureaucratic formalities. However, the downfall to this is that without a formal Memorandum of Understanding (MOU) between the stakeholders they do not invest their time, take ownership, and make the project a priority. The participants active in the design, build, test, and re-design process were junior level staff. These staff members were unable to convey the utility of SAMBRO and the project to their Directors and Decision-makers at the root of the organizational chain. Therefore, these participants treated the exercise simply as another ad-hoc activity. Nevertheless, the lead Organizations have shown a keen interest in operationalizing the system. Strategy is to prove its utility over a period of time, then the other Stakeholders would want come on board.

4.4 Technology related challenges

In this section we discuss the challenges of transforming the country context warning requirements, including the workflows and procedures and how those were mapped to design parameters involving the CAP content standard and the process variables involving SAMBRO feature. Generally, technology is perceived as the least challenging component in implementing any system. Relatively, the organizational challenges are leaps and bounds. We discuss some of the challenges CAP on a Map faced.

4.4.1. Short Message Service

Acquiring a bulk SMS service for a Government entity is difficult than for a private entity. In all three countries, the lead Organization was reluctant to pay for the service and was looking for the Mobile Operator to provide the service in-kind. Such an arrangement requires an enormous effort of weaving the wave through the Government bureaucracy. Myanmar Post and Telecommunications (MPT), now falls under the same Ministry as DMH; i.e. the Ministry of Transportation and Communication. However, it took over six months for DMH to get MPT to agree to provide a SMS Gateway free of charge. The next challenge was that MPT had never offered such a service to an external entity it was a news service. Commercial providers such as Clickatell and Text Magic do not service Myanmar. PAGASA was reluctant to undergo the procedures for getting one of the local mobile operators: Globe or Smart to provide a bulk SMS package. Instead the project had to purchase the service from Clickatell. Going across the ocean add more uncertainties and delays to the SMS alert dissemination. PAGASA could leverage the "free mobile disaster alerts law" [19].

The project facilitated for the Communications Authority of Maldives (CAM) to Mobile Operators to provide an SMS gateway. CAM negotiate a discounted rate for alerting first-responders. In the event the entire Nation needs to be notified of a Tsunami, for example, the SMS would be free. However, the deal did not come through. Instead, NDMC worked out a public private partnership with a local Bank to purchase the SMS service.

4.4.2. Email

In Myanmar we realized that many First-Responders did not have an Email account. Therefore, they are unable to receive email alerts. Email also serves as the means for verifying and activating a new user's registration and requesting for resetting the password. In the Myanmar case, the project had to create dummy email accounts such as somename@example.com. However, this account would not provide the aforementioned features related to a user's login.

4.4.3. Mobile App

During the silent-tests we realized some of the Smartphone to block the audible siren. These were caused by Smartphone applications that kill applications that idle for a long period of time. Also certain Smartphones had not provided the mobile SAMBRO app with permissions to forward notifications.

The first mobile SAMBRO app was developed for Android Smartphones. In Myanmar we observe nearly everyone to use Smartphones. Some were still using Android versions below 4.4. GCM and other SAMBRO feature do not function well on Android phones with an operating system version below 4.4. The app was also migrated to the Apple iOS platform. This decision was made when CAM presented that sixty percent of the phones were Smartphones and of them there was a fifty-fifty market split between Android and Apple phones. As a result the project made investments in adapting to the Apple market as well.

Following the CAP standard 1.2 which states the scope of the alert. The 'Public' alerts are disseminated to the public. In SAMBRO, the implementer can create different groups responsible for different activities during the disaster and issue Restricted or Private alerts for this purpose. Similarly, the previous alert can be updated, clear, cancel or error easily according to the CAP standard 1.2

4.4.4. Predefined Alerting Areas

The project realized that there is a need for developing predefined alerting area polygons to enhance warning efficiencies. To address this need SAMBRO has introduced a mapping tool that allows Risk-Analysts and Warning Practitioners to develop a set of predefined polygons. Identifying the level of risk by geographical area allows for defining impact-based alerting. Integrating risk maps, with SAMBRO, allows differentiating community that might be at a greater risk to a hazard event over another. In order to save the lives and livelihoods, alerts can be issued to higher impacting communities first for the responders to attend that community's needs first. Thus, removing the burden of optimizing the response resources.

Naturally, it is difficult to predefine alert areas for a forest fire a tropical cyclone but can be defined for volcanic, tsunami, and floods. The National Stakeholders, in the three countries, faced difficulties acquiring any kind of risk map as Vector or Rasta GIS data. Some cases they had very limited risk maps, confined to a few targeted regions or townships, but as still images; of no use to SAMBRO to offer interactive analysis and mapping capabilities.

A dilemma with Island nations (or Archipelagos) was including or excluding Islands that are not visible on a map. For example, Maldives wanting to develop predefined alert areas for each of the Atolls, when zooming in to an Atoll could not see some of tiny Islands to realize, when they were drawing a polygon, was including or excluding those unseen Islands.

Geocodes presented themselves as an innovating solution for overcoming the dilemma of differentiating Islands that belong various jurisdictions. Philippines had implemented geocodes to define their administrative areas. CAP provides means for defining an alerting area associated with the respective geocodes. The lesson learned from the Philippines was valuable for requesting Maldives to follow the same, seeking assistance from the Maldives Land and Survey Authority.

5. CONCLUSION

EWSs, in the region, are gradually evolving to their required potential. However, the concepts of moving beyond a top-down approach to a peer-to-peer approach using software services is yet to mature. SAMBRO has realized those gaps and has presented itself to server in this capacity. National policies and strategies must be put into practice to further strengthen these concepts. A growing challenge faced by National initiatives is integrating the early warning dissemination and coordination with all relevant stakeholders. The authors have realized the intricacies of implementing cross-agency situational-awareness platforms in the Region. Lessons learned from the current experience provides inputs to shaping SAMBRO strategies for operationalizing such systems for improving institutional responsiveness to all-hazards.

REFERENCES

[1] United Nation International Strategy for Disaster Reduction (UNISDR), "HFA Progress in Asia Pacific", 2009-2011

[2] A. Zia, and C.H. Wagner, "Mainstreaming early warning systems in development and planning processes: Multilevel implementation of Sendai framework in Indus and Sahel", International *Journal of Disaster Risk Science*, *6*(2), pp. 189-199., 2015

[3] S. Smith and D. Bunker, "Disaster Management and Community Warning Systems: Inter-Organizational Collaboration and ICT Innovation", *Pacific Asia Conference on Information Systems (PASIC) Proceedings*, Hyderabad, India, 2009

[4] J. Pagotto and D. O'Donnell, "Canada's Multi-Agency Situational Awareness System - Keeping it simple", *The 9th International ISCRAM Conference. Proceedings,* Vancouver, Canada, April 2012

[5] WEF - World Economic Forum, "City Limits: The Risks of Rapid and Unplanned Urbanization in Developing Countries", *10th Edition of the Global Risks Report,* 2015

[6] R.J. Nicholls, P.P. Wong, V.R. Burkett, J.O. Codignotto, J.E. Hay, R.F. McLean, S. Ragoonaden and C.D. Woodroffe, "Coastal systems and low-lying areas", *Climate Change 2007: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change,* Cambridge University Press, Cambridge, UK, pp 315-356, 2007

[7] D. Griggs, M. Stafford-Smith, O. Gaffney, J. Rockström, M.C. Öhman, P. Shyamsundar, I. Novel et.al, "Policy: Sustainable

development goals for people and planet", *Nature*, Macmillan Publishers Limited, pp 305-307, 2013

[8] M. Prustalis and C. De Silva, "The Sahana Free and Open Source Disaster Management System in Haiti", *ICT for Disaster Risk Reduction, United Nations Asian and Pacific Training Centre for Information and Communication Technology for Development (UN-APCICT/ESCAP)*, pp. 110-124

[9] M. Careem, D. Pradeeper, M. Kaluarachchchi, G. Gow, C. Sampath, M. Ganesan, N. Janakiraman and N. Waidyanatha, "Sahana Alerting Software for Real-Time Biosurveillance in India and Sri Lanka", Proceedings of the 1st IEEE International Conference on Computer and Information Applications (ICCIA), Tianjin, China, pp. 37-373, Dec. 03-05, 2010

[10] N. Waidyanatha, G. Gow, and P. Anderson. "Common alerting protocol message broker for last-mile hazard warnings in Sri Lanka: An essential component". *Proceedings of the 2nd International Information Systems for Crisis Response and Management (ISCRAM),* Harbin, China, pp 59-65, Aug. 26-27, 2007.

[11] G. Gow and N. Waidyanatha, "Using common alerting protocol to support a real-time biosurveillance program in Sri Lanka and India", *Biosurveillance: Methods and case studies*. Chapman and Hall,.pp. 268-288, 2010

[12] K. Perera, T. Wilfred, N. Waidyanatha, and M. Silva, "Automation Complexities in Integrating Voice and Text for Emergency Communications in Sri Lanka", *Annual Technical Conference*, University of Moratuwa, Sri Lanka, 2011.

[13] D. Wells and L. (Eds) Williams, "Extreme Programming and Agile Methods-XP/Agile Universe 2002", *Second XP Universe and First Agile Universe Conference Chicago,* Proceedings, Vol. 2418, Springer, Chicago, IL, USA, August 4-7, 2002.

[14] C. Abras, D. Maloney-Krichmar, J. Preece, "User-Centered Design", *Berkshire Encyclopedia of Human-Computer Interaction: When Science Fiction Becomes Science Fact*, Vol.2. pp. 763-768, 2005

[15] L. J. Najjar, "Rapid prototyping (TR 52.0020)", .Atlanta, GA: IBM Corporation, 1990

[16] C-K Chau, K-F Leong, and C-S Lim, "Rapid Prototyping Principles and Applications", World Scientific Publications, 2nd Edition, 2003.

[17] G. Hearn, and M. Foth, "Action research in the design of new media and ICT systems", *K. Kwansah-Aidoo (Ed.), Topical issues in communications and media research,* Nova Science, New York, pp. 79–94, 2005

[18] R. O'Brien, "An Overview of the Methodological Approach of Action Research", *Roberto Richardson (Ed.), Teoria e Prática da Pesquisa Ação [Theory and Practice of Action Research]*, Universidade Federal da Paraíba, João Pessoa, Brazil, 2001

[19] G de la Cruz, "Will the free mobile disaster alerts law finally be implemented?" *Rappler,* Philippines, 2015

A COMMUNITY-DRIVEN INFORMATION SYSTEM TO DEVELOP NEXT GENERATION COLLABORATIVE AND RESPONSIVE RURAL COMMUNITY (NCORE)

Jayanta Basak, Rishikesan Parthiban*, Siuli Roy ǂ, Somprakash Bandyopadhyay**

*Social Informatics Research Group, Indian Institute of Management Calcutta, India
ǂ Heritage Institute of Technology, Kolkata, India
@ Corresponding author (somprakash@iimcal.ac.in)

ABSTRACT

Much of the ICT interventions for rural transformation are exogenous in nature (development from outside), in the sense that they use a "push" approach towards development, without considering the nature and problems of an individual member of the rural community. Information and knowledge transactions, especially with dis-empowered people and groups, are a complex process and ICT needs to be appropriated and used in a way that helps resolve daily concerns. With this perspective in mind, this paper proposes NCoRe, an interactive community-driven information system platform to harness the potential of community participation in governance. In a digitally-connected global society, each individual in a community of people is not only a consumer of information but also a producer of information: a potential contributor in many ways to build a better community. NCoRe exploits the potential of community knowledge, making them available to the community and empowering the communities to interact, collaborate and participate in the development of society and transforming the way they live, learn and work. NCoRe is an ongoing initiative to build next-generation collaborative and responsive community by empowering the rural community of India with an ICT-enabled "capability framework" involving the self-help groups (SHG): the micro-communities within a village community.

Keywords— Self Help Group (SHG), Rural community, Interactive information system, Capability framework

1. INTRODUCTION

Large chunks of the population of rural India are marginal farmers, landless laborers, petty traders and rural artisans who are socially and economically backward (including the tribal population). They are faced with problems related to poverty, illiteracy, lack of skills, health care etc. Mainstream institutional efforts are inadequate to provide financial support and livelihood options to the rural masses of India. Also, these are problems that cannot be tackled individually but can be better solved through group efforts. With this objective in mind, small groups are formed within a village with the purpose of solving their common problems. These groups, known as **Self-help groups (SHG),** are micro-communities (with 10-to 15 members) within a village community and they have now become the vehicle of change for the poor and marginalized [1, 2]. In other words, Self-help group is a method of organizing the poor people and the marginalized to come together to solve their problems. The SHG method is used by the government, NGOs, and other non-government agencies worldwide.

SHG movement in India involves voluntary association of people (mostly women) in small groups to address their lives and livelihood issues. SHGs in most cases are groups that have been formed by economically and socially deprived a class of the society (particularly women) who have come together to discuss and raise their voice together. Savings and credit activities act as binding forces for them in most cases. However, SHGs have been conceived not just to promote savings and provide credit, but also to act as institutions of change and aid in human development in order to empower its members. They have great potential to address social, gender and women empowerment issues and help to strengthen local governance [1]. In India, there are three major streams through which SHGs have been promoted. The first one is NABARD's SHG Bank Linkage Program (BLP) and the second one is through the Swarnajayanti Gram Swarojgar Yojana (SGSY), a Self-Employment program launched by the Government of India to promote micro-entrepreneurship and provide sustainable income to poor people. The third stream is through the NGOs. In BLP alone, about 94 million poor villagers, linked with banks through 7.5 Million SHGs, have mobilized an amount of Rs. 33,000 crores (US$5.5 Billion) as savings and issued loans to the tune of Rs 66,000 crores (US$11 Billion) of which Rs 43,000 crores (US$7 Billion) is an outstanding credit mobilized from banks. The poor women of these SHGs in India collectively control the financial business with an annual turnover of Rs 100,000 crores (US$17 Billion), much larger than many multi-national corporations in India [3, 4].

In spite of this huge investment and volume of people involved in upliftment and livelihood enhancement of the rural community, the success is still limited [5]. One of the major problems (discussed in section 3) is a lack of inter-group, group-to-agency (NGO, Bank or Government) and group-to-external world communication and coordination.

For example, agencies are giving money for micro-entrepreneurship development, expecting SHGs to produce goods and services, but they usually do not provide any market linkages. Similarly, agencies are organizing training for SHGs without considering their need and capabilities. Agencies are focusing on frequency of meetings among SHG members and loan repayment issues but pay less attention to their day to day problems. Thus, the entire approach is ***exogenous*** in nature (development from outside), in the sense that it uses a "push" approach towards development, without considering the nature and problems of *individual* SHG. This approach overshadows the ***endogenous*** *model of development* (development from inside) [6] that focus more directly on human beings and their resources and aspirations.

In this context, the power of ICTs can be fully exploited to promote local development. For many remote communities, ICTs are a means for communicating with the outside world. Intelligent use of ICT can improve involvement of everyone in the community, especially young people. It may help to build networking with agencies and help administration with digital documentation for monitoring and evaluation. There are several Government / Non-Government initiatives in India to include ICT in a rural community in general and SHGs in particular. The major focus is to give access to internet connectivity and value-added mobile services. For example, the Common Services Center Scheme has started as a part of the ambitious National e-Governance Plan (NeGP) of Government of India. The Scheme envisages the setting up of 100,000+ IT-enabled access points (internet kiosk) to act as outlets for two basic services – reaching an IT infrastructure to all Indian villages and developing an organizational system for delivering services over this infrastructure employing the profit incentive [7]. The 'EShakti' is another initiative by NABARD, specially designed project for e-bookkeeping of SHG records and related Management Information System (MIS) on a real time basis. This is in tune with the GOI's mission for creating a Digital India. To begin with, two districts viz. Ramgarh (Jharkhand) and Dhule (Maharashtra) are being covered in pilot mode [8].

However, making ICTs available is not enough to ensure that people have access to right kind of information; it is more important that ICTs are appropriated and used in a way that helps resolve daily concerns. With this perspective in mind, this paper proposes **NCoRe,** an interactive community-driven information system platform to harness the potential of community participation in governance. In a digitally-connected global society, each individual in a community of people is not only a consumer of information but also a producer of information: a potential contributor in many ways to build a better community. They can collaborate for a social mission, participate in local governance, respond to the emergencies, communicate their needs and wants and share their knowledge and expertise to help other underprivileged communities. With this notion, NCoRe offers an interactive platform to exploit the potential of community information and knowledge, making them available to the community and empowering the communities to interact, collaborate and participate in the development of society and transforming the way they live, learn and work. NCoRe is an ongoing initiative to build next-generation collaborative and responsive community by empowering the rural community of India with an ICT-enabled "capability framework" involving the self-help groups (SHG) at the different block, district, and state levels. Currently, it collects data from the target communities through interactive SMS, as internet accessibility and affordability is still a problem in rural India. However, NCoRe can easily be upgraded to the internet-enabled interactive system to build next-generation collaborative and responsive communities.

2. DEVELOPMENT OF RURAL COMMUNITIES THROUGH ICT

ICTs have the potential to make a huge impact on developing countries. Not only can ICTs be used to integrate rural communities into wider economic and social development, but digital technologies can also be used to enhance and preserve the knowledge and culture of rural communities [9]. However, as mentioned earlier, this assertion is derived from an ***exogenous model of development*** that underpins many of the interventions aimed at employing ICTs to meet poverty reduction goals. It comprises development from outside and it overshadows the ***endogenous model of development*** (development from inside) model that focuses more directly on human beings and their resources and aspirations. [6, 10]

2.1 The Exogenous Model

The exogenous model assumes that the necessary technology (in terms of hardware, software, and services) already exists in the world, as does the considerable experience of its use. Therefore, the development task is to encourage the acquisition and application of the technology, support training of its use and promote the type of regulatory changes, as needed [6]. This assumption is based on traditional theories of modernization, in which technologies are 'transferred' from 'developed' countries to less developed ones. "The exogenous model (and indeed some versions of the endogenous model), cloaks the interests of investors in the global 'North' whose principal ambition is profited from the sale of digital technologies and the content that is hosted on or circulated through them" [10]. An interesting recent example is Facebook's Free Basics, where Facebook offers free internet access to users only for a few selected sites including Facebook.

It is beyond doubt that through this approach, quite a few significant developments have taken place. ICTs are acquired and used; telecommunication infrastructures are improved and the cost has decreased; connections to the outside world are made. There are significant progress in several countries like Africa, Srilanka, Thailand, Bangladesh, India and several South-East Asian and east European countries in the domain of 1) agriculture and health 2) infrastructure, communication, and community

informatics / knowledge, 3) Economic empowerment and small scale entrepreneurship, 4) Policies, strategies and e-governance [9, 11].

However, ICT interventions for rural transformation are more than just making people "literate" in ICT and making it accessible. Björn-Sören Gigler shows [12] that there is a gap between information and communications technologies (ICTs) and socio-economic development. In this paper, he has shown that, if information is critical to development, then ICTs, as a means of sharing information, are not simply a connection between people, but a link in the chain of the development process itself. ICTs can enhance the functioning of markets because it can properly integrate and bind the floating market components into static contents in order to provide a sustainable model.

Much of the 'ICT promise' for rural transformation has been expressed in terms of the power of information and knowledge. However, information and knowledge transactions, especially with dis-empowered people and groups, are a complex process. It is generally not amenable to across-the-counter productization and monetization. Only some kinds of information – like agriculture price information, health related information, etc. can be delivered usefully through a rural kiosk-based model. Most other information and knowledge transactions are much more human interaction intensive and require to be done in an altruistic and community-minded spirit [7]. A simple example is a poor destitute woman seeking help against domestic violence. There is no digital platform for her using which she can seek help and get advisory support from the external world.

2.2 The Endogenous Model

The endogenous approach is based on an understanding that the most important impact related to ICTs and their use are caused not by the technology, but by the new forms of informational behavior they facilitate. These new behaviors enable new connections to be made, which may offer the potential of new value and transformative change to be created in the social, political, and cultural spheres, as well as in the economic [10].

As concluded in the UNRISD workshop [13]: "it is a serious mistake to assume that they constitute a uniform process globally or share a common destination, rather than a variety of new processes each influencing and being influenced by the society in which they are taking place." ICT for development, therefore, in this approach is about finding solutions to the informational challenges faced in the process of development. The core developmental challenge here is to make it community driven: encourage and support informational developments within the communities themselves.

2.3 The Digital Divide

Almost three in five of the world's people are still not connected to the Internet. This digital divide hampers economic and social progress. Broadband markets that price Internet access out of reach for the majority of people are neither socially nor economically efficient. Liberalizing the telecommunications industry is not enough; the state also has to facilitate strategic investments, subsidizing access for underserved communities and implementing effective and transparent regulations, including open access to subsidized infrastructure [14]. In several developing countries, "Free Basics" by Facebook or similar initiatives provide people with access to useful services on their mobile phones for free without data charges. However, in India, these kinds of services are banned to favor net-neutrality. Hence, active participation of all stakeholders in hammering out a concrete plan of action is perhaps the single most important step to move from high prices and low uptake to low prices and high demand [14].

3. PROBLEMS FACED BY SHG

In order to design an information system following an endogenous model of development, we should move towards a more citizen and service-centric approach, beginning with a needs analysis: what are the needs, who for, how to develop solutions, where the services need to be delivered and how they can be accessed. In this context, we have conducted both primary and secondary studies to identify the problems faced by SHGs in order to derive system specification of our NCoRe system.

Our secondary study [15, 16] reveals the following set of problems faced by SHGs in India:

1. **Ignorance of Members/Participants**: Majority of the group are unaware of the schemes of assistance offered to them. Many are Ignorant about the scheme.
2. **Inadequate Training Facilities**: The training given to the members of SHGs are inadequate
3. **Problems Related to Raw Materials**: Normally each SHG procures raw materials individually from the suppliers. They purchase raw materials in smaller quantities and hence they may not be able to enjoy the benefits of large-scale purchase like discount, credit facilities etc. Most of the SHGs are ignorant about the major raw material suppliers and their terms and conditions. All these causes high cost of raw materials.
4. **Problems of Marketing**: Following are the major problems relating to marketing.
   (a) Lack of sufficient orders.
   (b) Lack of linkage with the marketing agencies.
   (c) Lack of adequate sales promotion measures.
   (d) Lack of permanent market for the products
   (e) Absence of proper brand name.
   (f) Poor/unattractive packing system.
   (g) Poor quality of products due to lack of quality consciousness
   (h) Lack of a well defined and well-knit channel of distribution for marketing
5. **Weak Financial Management**: The return from the business is not properly managed for further investment due to lack of proper financial planning
6. **Low Return**: The return on investment is not attractive in certain groups due to inefficient

management, the high cost of production, the absence of quality consciousness etc.

7. **Noncooperative Attitude of the Financial Institutions**: The Financial Institutions do not consider SHGs seriously while providing finance and other help.
8. **Inadequate Support from Line Department**: For obtaining assistance and support, the group members have to approach the line officers, who are not always very co-operative with the SHGs.

To empirically analyze the dynamics of the SHGs, a primary study was conducted in a state in India, namely Tamil Nadu. 100 semi-structured in-depth interviews were conducted in six villages spanning across three districts, over a four month period. The interviews covered the group leaders, the group representatives, and the general group members. Each interview lasted from half an hour to over an hour. To gain insights into the dynamics of the SHGs in order to assess their strengths and weaknesses (and identify possible gaps where ICT can pitch in), specific questions were posed to the members of the group. Additionally, the questions were carefully designed to bring out factors that would impact the performance of the SHGs. The questions were carefully translated to the native language by the author.

Table 1: Analysis of Gaps from Information sharing perspective

Themes	**Analysis of Gaps in SHG developments from Information sharing perspective**
Networking and External Links	Limited Inter-Group Interactions; No knowledge about SHGs outside the group's locality; referential (word of mouth) growth of SHGs; enrolment of new members only through member references
Market Awareness	Personal selling of group's products; No knowledge of market outside the locality.
Supporting Agencies	Low interactions with the NGOs; No knowledge about other supporting agencies; harassment by the banking system
Government Schemes	Limited knowledge about the benefits of government schemes
Health	Lack of knowledge about procurement of health insurance; limited awareness about seasonal epidemic diseases
Miscellaneous findings	Lack of task interdependence within the group; Regular Manual maintenance of multiple record books, which is cumbersome

To bring in objectivity to the data captured through the interviews, they were transcribed and qualitatively analyzed for dominant patterns and themes. Since many of the participants were not able to answer the survey themselves, the person administering the survey read out the questions to the participants on many occasions, who then went on to fill the answers. The responses were used in two ways. Firstly, the average scores of all participants for each module and sub-module were calculated to obtain insights about the dynamics of the SHGs and identify gaps where in ICT could play a role. Additionally, questions for which the responses were very high or very low were filtered out for finer analysis.

Table 1 provides the highlights of the findings from the information sharing perspective. It would help in the analysis of gaps where ICTs could play a role.

4. NCoRe SYSTEM MODEL

Based on the above observations, NCoRe (**N**ext Generation **Col**laborative and **Re**sponsive Rural Community) has been designed and developed as an Interactive Community driven Information System that helps the day-to-day functioning of the SHGs, who are acting as micro-communities within a village community. The objectives of NCoRe are to

a. empower SHGs to interact and collaborate among themselves and with other rural development authority's (sub-clusters, clusters, government agencies, NGOs etc.) to share their problems and search for solutions
b. enable the involvement of SHGs to participate in their local governance
c. support them to access various skill development programs, entrepreneurship training to improve their livelihood
d. guide them to establish market linkage to sell their products, get raw materials and to get financial assistance
e. engage them to form a resilient community for better disaster management

As indicated earlier, in India, apart from infrastructure, a major impediment in Internet usage is still internet costs. In several other developing countries, "Free Basics" by Facebook or similar initiatives provide people with access to useful services on their mobile phones for free without data charges. However, in India, these kinds of services are banned to favor net-neutrality. Still, taking advantage of the fact that cell-phone penetration and SMS usage in India is very high; NCoRe collects data from the target communities (SHGs) through interactive SMS in native language [17, 18]. However, NCoRe can easily be upgraded to the internet-enabled interactive system to build next-generation collaborative and responsive communities, transforming the way they live, learn and work. A block diagram of NCore System is shown in fig 1.

NCoRe System collects and collates information from SHGs regarding their needs and wants and connects them electronically to different government and non-government agencies dealing with livelihood enhancement and community well-being programs involving SGHs. NCoRe has the following features:

1. Keep a record of static information of SHGs. This part will be recorded in the system only once on registration of new SHG in the system

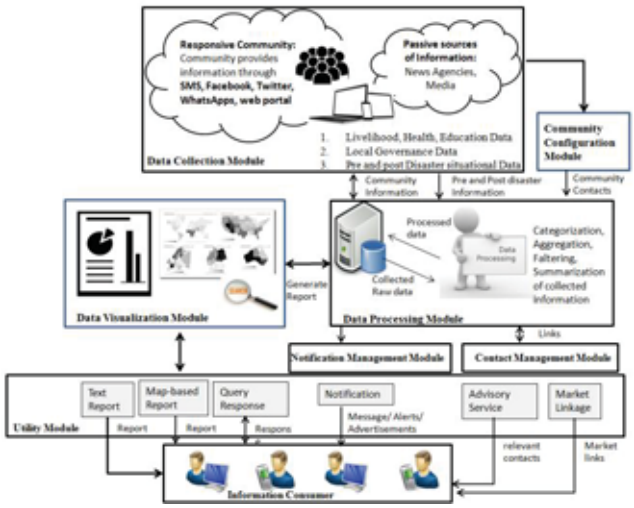


Figure 1: NCoRe system block diagram

2. The administrative agencies (federation, government agencies) will periodically interact with SHG through NCoRe to know how the SHGs can be supported in different sectors such as the need for vocational training, market linkage, advisory services, etc. Some examples are:

- **Vocational Training**
  - i. SHG members will express their need for various livelihood related vocational training and the location specific summary of their needs will be generated by the system
  - ii. The agencies will conduct specific topic-based training at suitable time and venue on getting sizeable training demand and advertize that on the system as well as send messages to the members who expressed their interest for the said training
- **Market linkage**
  - i. System will interact periodically with SHG members to know if they have any product to sell and the seller can send their product specification, the preferred selling price, quantity and contact details
  - ii. Category-wise Product list of sellable items will be displayed along with contact details in the web portal
  - iii. Prospective buyers may also register in the web portal for buying specific products along with their location and contact details
  - iv. Buyers can check the category wise product list and can contact the seller directly
- **Advisory Service**
  - i. Counseling related to
    - Health and hygiene
    - further studies,
    - job oriented studies
    - mental disorder
  - ii. Contacts for available Medical support/ rehabilitation centers in nearby areas for drug addicts physically and mentally challenged persons
  - iii. Contacts for micro financers
  - iv. Contacts for legal advice
  - v. Push messages in mobiles of SHG member regarding health and hygiene in regional language

NCoRe software framework contains four broad modules; i) Manage Collaboration, ii) Manage Response, iii) Manage Reports and iv) Manage Notifications and Contacts which is illustrated in fig 2. The components, responsibilities, and functions of each module are described below.

Figure 2: NCoRe System Framework

Manage Collaboration contains four sub-modules; i) Configure community, ii) Configure questionnaire, iii) Manage interaction and iv) Configure User access. Tasks of each sub-module are discussed below.

- ***Configure community:*** Since basic objective of NCoRe is to empower the community, it is the, therefore, necessary to feed the system with the basic contact details of the members of the community so that system can initiate automatic interaction with the community on different subjects as and when needed.
- ***Configure questionnaire*****:** The questionnaire for interaction with the community is framed in native language and is added to the system through this section. The questionnaire may be divided into multiple subcategories so that community member may choose any category for which he/she is interested in providing information. He will then get multiple choice based questions from selected category only. Interaction with relevant community associated with a particular domain makes it easier to gather meaningful and authentic information of particular category, i.e., community members associated with farming may feel more comfortable to interact on agriculture and farming related matters in his locality than on any other topics. The questionnaire is organized in such a way that next question is picked based on the response to the earlier question on a

particular category. Categories and questions can be changed based on the information need of the user.

- ***Manage Interaction:*** This module is responsible for the collection of different types of information regarding the regular activities and needs of a community from the community itself. We have used the concept of interactive information crowdsourcing [17, 18] and integrated an *interactive SMS system* in this module that enables the system to initiate a dialogue with interested members of the community on specific domain selected by the members from a given list of options. *Interactive SMS system* has a built-in intelligent query manager that automatically selects suitable next question based on the earlier response obtained from the member so that a comprehensive picture about a specific domain can be framed. Currently, topics are categorized as livelihood, health, education, local governance, pre and post disaster situational updates. However, this list may be further augmented based on community need.

Manage Response is essentially data processing module that offers functionalities like response validation, location, and topic based aggregation, filtering and summarization of collected responses. Data processing is a multi-step method where data sources are authenticated, data aggregation is done, data duplication and inconsistency are removed and finally topic-specific content analysis and summarization is done.

Manage Report is data visualization module with two sub-modules: configure report format and configure report viewer.

- ***Configure report format***: This enables the administrator to input user-specific report format in the system as that may differ from user to user.
- ***Configure report viewer:*** Information is made available to the prospective agencies like cooperatives, village authorities, district headquarters and state Government in their desired format. Each agency may opt to view the report in text format or tabular format or on the Map. User specified viewing options are fed to the system through this section.

Manage Notifications and Contacts has three sub modules: Configure Notification, Push Notification, and Publish advertisements

- ***Configure Notification:*** Any emergency or administrative notifications are added to the system through this section by the administrator. Since any notification has time bound validity, the system also offers flexibility to withdraw earlier notification if validity is over.
- ***Push Notification:*** Administrator may push any domain-specific notifications to the relevant section of the community only. Otherwise, mass notifications may be viewed on the web portal or location-based notification SMS may be sent by the system automatically to the desired community. It also includes an advisory service where contacts and information about special needs of a community like, specialized medical treatment, career counseling, higher education and vocational training related queries, exhibition, trade fare related queries are answered and corresponding responses are sent as SMS to the members who raised the query.
- ***Publish Advertisement:*** A collaborative platform is implemented here that enables community members to establish a link to the market beyond the local boundary. Community members can advertise the details of their manufactured products for getting suitable market links. On the other hand, prospective buyers can publish their contact details and purchase interests in different products manufactured by the community. The location and product-specific buyer and seller details are organized by the system and made available in the Web portal so that buyer and seller can establish a direct selling channel if their interest matches.

5. A FIELD TRIAL WITH NCoRe

5.1 Data Collection: Methodology

To test our Information System, a field trial was organized with 100 SHG members at remote villages in the Namkhana region of West Bengal, India. Internet connectivity in this location is very poor and, even if it is available, it is not affordable to the majority of people. Community members are using low-end mobile phones to fulfill their daily communication needs. Establishing Internet-based communication with community members is a big challenge in that place.

To cope with this challenge, we use our ***Interactive SMS System***, which will initiate an automated conversation with the selected SHG members of a community on the specific topic. As discussed in the earlier section, multiple choice based queries are picked up from the topic-specific question set preloaded in the system and automatically sent as SMS to the selected SHG members. The SHG members need to choose the proper option to indicate his answer to that query. As a result, SHG members having a low-end phone can also participate in the interaction and answer the query with single key stroke. Based on the earlier response, query manager module sends the next query to the members to get more in-depth information on a specific topic. From the collected data, our system will be able to identify the basic needs of local communities in terms of i) Market Linkage ii) Training Needs and iii) Advisory Services, etc..

Within this framework, we have investigated key factors that have to be met to enable the underprivileged community to have real and meaningful access to ICTs and allow them to appropriate these technologies as an instrument for their own development. Questions are prepared based upon Reserve Bank of India's guidelines [19] on National Rural Livelihoods Mission launched by the Ministry of Rural Development, Government of India**.** A sample set of questions (not exhaustive) is given in

Table 2, which was translated in native language and was used in the form of interactive SMSs during the field study.

Table 2: Sample Questions (translated in native language) used for interactive SMS during the field study

Q. No.	Question in Category: **Training**
1	Are you interested to participate in any type of training program organized by NGO / Gram Panchayat? **a.** Yes **b.** No
2	What type of training programs is suitable you? **a.** Agriculture **b.** Cooking **c.** Poultry **d.** Tailoring
3	Do you have access to regular training / workshops? **a.** Yes **b.** No
4	Do you have previous experience in order to create M.I.P. (Micro Investment Plan)? **a.** Yes **b.** No
5	How many people from your team want to be associated with this type of training programs? **a.** Less than 5 **b.** Greater than 5
6	From where do you get product development training? **a.** From block **b.** From NGO **c.** not at all
Q. No.	Question in Category: **Market Linkage**
1	Where do you sell the product? **a.** City Market **b.** Cooperatives **c.** Local Market
2	To whom do you sell the product? **a.** Wholesaler **b.** Retail **c.** Co-operative **d.** Government
3	What is the main problem for selling your products? **a.** Less Market Price **b.** Low demand **c.** Monopoly Business **d.** Transportation **e.** Sale price unknown
4	How much profit do you earn? **a.** Good **b.** Inconsiderable **c.** No profit
5	Did you know the exact market value of your product? **a.** Yes **b.** No
6	From where, raw materials will be collected? **a.** Local market **b.** City market
Q. No.	Question in Category: **Advisory / Counseling Service**
1	What kind of support do you want? **a.** Health **b.** Education and training **c.** Legal
2	What kind of medical-related supports you want? **a.** Heart **b.** Pregnant women **c.** Children **d.** Mental illness **e.** Cancer **f.** Addiction treatment
3	What kind of advice do you want? **a.** Job **b.** Higher Education
4	Need legal assistance? **a.** Family Violence **b.** Property Related Issue
5	Need any type of Vocational Training? **a.** Yes **b.** No

5.2. Data Processing

The responses are collated and classified according to different categories to document their day to day needs and wants. These responses form the basic pointers to the respective authorities / agencies to take appropriate actions and orient their services in a more *customized* fashion. The insights generated in this field trial suggest that NCoRe will be useful both for purposes of effective program development and policy design. It promotes interactions and provides direct connectivity among:

- SHG-to-SHG: sharing knowledge and experiences across SHGs
- SHG-to-Government : making government agencies aware of the needs and wants of SHG in particular and the community at large
- SHG-to-Bank: Enabling banks to identify financial requirements of SHGs before executing formal procedure
- SHG-to-market: Enabling SHG members to get market awareness and market linkage directly bypassing middleman; enabling them to identify sources of raw material procurement and channels for finished product selling
- SHG-to-Advisory Service Provider: enabling the providers to know the needs of SHGs and act accordingly (mediated by NGOs)
- SHG-to-Training Organizations: enabling the providers to know the needs of SHGs and act accordingly (mediated by NGOs)

5.3 Challenges Faced and Possible Solutions

✓ Most of the SHG members are from indigent communities of the village with very low literacy skills. They use only low-end feature phones for calling purposes and are not used to SMS. Additionally, all feature phones do not provide native language support in firmware level. So, we cannot interact with some of them using SMS in the native language. Some members have difficulties in reading native language text also. In future, *IVR (Interactive Voice Response)* [20] service will be incorporated with NCoRe system, through which NCoRe system will be able to interact with the community members through voices instead of SMS text. In both cases, community training on phone usage will be needed.

✓ Presently, SMS cost in India is high. It is an impediment to communicating with indigent communities through Interactive SMS. At present *Toll-free SMS* in India is not possible. But, TRAI (Telecom Regulatory Authority of India) issue licenses [21] to some mobile operators for providing MVNO (Mobile Virtual Network Operator) [22] service in the India. Under this technique, toll-free SMS will be possible and senders need not pay for their SMSs. This service will be incorporated with NCoRe in future.

✓ We are also planning to provide a low-cost smart phone to a group of SHGs on a trial basis. This scheme is already implemented in some villages of Tamil Nadu, India. A customized smart phone has been distributed to 4,626 SHGs in the district under NRLM [23] scheme.

6. CONCLUSION

Currently, NCoRe system collects data from the target communities through interactive SMS, as internet accessibility and affordability is still a problem in rural India. However, NCoRe can easily be upgraded to the internet-enabled interactive system to build next-generation collaborative and responsive communities. We are planning to conduct a field trial with 100 SHGs by providing the leader of each SHG a smart phone and free internet connectivity for six months. We will train them so that they can use NCoRe directly (in their native language) as well as perform group chats in a voice as well as text mode in their native language using the WhatsApp-like platform. Eventually, NCoRe will become the digital platform of the SHGs, empowering them to interact, collaborate and participate in the development of the community. We conclude with a recommendation that free/ subsidized internet access for selective services is a must to serve the underprivileged in developing countries in order to create a vibrant rural community.

REFERENCES

[1]. VOICE (Voluntary Operation in Community and Environment). "Success and Failure of SHG's in India – Impediments and Paradigm of Success". *A Report*. Planning Commission, Government of India, 2008. http://planningcommission.gov.in/reports/sereport/ser/ser_shg3006.pdf

[2]. N. Sreedhar, "Self Help Groups Performance in India" *Arth Prabandh: A Journal of Economics and Management* Vol.1 Issue 8, November 2012.

[3]. Narender Kande, "The Self Help Groups movement in India", *impatient optimists*, Bill & Melinda Gates Foundation, January 08, 2015. http://www.impatientoptimists.org/Posts/2015/01/The-Self-Help-Groups-movement-in-India#.V3-I8dJ9600

[4]. NABARD Report. "Status of microfinance in India (2014-15)". https://www.nabard.org/Publication/SMFI-2015r-new.pdf

[5]. Savitha V and H. Rajashekar, "Evaluation of Major Problems Faced by the Members of Self Help Groups: A Study of Mysore District". *International Journal of Research in Applied, Natural and Social Sciences (IMPACT: IJRANSS)* Vol. 2, Issue 6, Jun 2014, 59-64

[6]. International Telecommunication Union (ITU). "The Role of ICT in Advancing Growth in Least Developed Countries: Trends, Challenges and Opportunities", *Report, 2011* http://www.itu.int/pub/D-LDC-ICTLDC.2011-2011

[7]. IT for Change. "ICTs for community development in India - Going beyond the basic CSC model". *Position paper, 2009* http://www.itforchange.net/gov-ict/73-governance/280-csc-model.html

[8]. NABARD. E-Shakti portal. https://eshakti.nabard.org/Aboutus.aspx, 2015

[9]. Rebecca Nthogo Lekoko and Ladislaus M. Semali "Cases on Developing Countries and ICT Integration: Rural Community Development", *IGI Publication*. September, 2011. ISBN13: 9781609601171

[10]. Mansell, R. (2010). Power and interests in developing knowledge societies: exogenous and endogenous discourses in contention. *IKM Working Paper* No. 11, 2010. p. 7. http://wiki.ikmemergent.net/files/IKM_Working_Paper-11-Robin_Mansell-July2010-final-pdf.pdf

[11]. International Telecommunication Union (ITU). (2005). Tunis agenda for the Information Society. http://www.itu.int/wsis/docs2/tunis/off/6rev1.htm

[12]. Björn-Sören Gigler, "Informational Capabilities'- The Missing Link for the Impact of ICT on development" *Working Paper Series* No.1, The World Bank, March, 2011

[13]. UNRISD. (2005). *Understanding Informational developments: a reflection on key research issues*. *Conference Report. p. 2.* http://www.unrisd.org/

[14]. Alliance for Affordable Internet (A4AI) (2013). "The Affordability Report." http://a4ai.org/wp-content/uploads/2014/01/Affordability-Report-2013_Final-2.pdf

[15]. T. Thileepan and K. Soundararajan, "Problems and Opportunities of Women SHG Entrepreneurship in India" *International Research Journal of Business and Management – IRJBM* Volume VI, December 2013.

[16]. Neha Misra "Problems Faced by SHGs and Suggestion to Minimize It" http://www.yourarticlelibrary.com/india2/selfhelpgroup/problemsfacedbyshgsandsuggestiontominimizeit/66720/

[17]. Chen C., Huang Y., Liu Y., Liu C., Meng L., Sun Y., Bian K., Huang A., Duan X., Jiao B. "Interactive Crowdsourcing to Spontaneous Reporting of Adverse Drug Reactions". In: *IEEE ICC 2014*, Sydney, 2013.

[18]. Anubrata Das, Neeratyoy Mallik, Somprakash Bandyopadhyay, Sipra Das Bit, Jayanta Basak: "Interactive information crowdsourcing for disaster management using SMS and Twitter: A research prototype". PerCom Workshops 2016: Sydney, Australia

[19]. Reserve Bank of India. "Priority Sector Lending- Special Programmes" June 2013. https://www.rbi.org.in/Scripts/NotificationUser.aspx?Id=8075&Mode=0

[20]. https://en.wikipedia.org/wiki/Interactive_voice_response

[21]. http://thetechportal.com/2016/03/29/telecom-commission-approves-mobile-virtual-network-operators/

[22].https://en.wikipedia.org/wiki/Mobile_virtual_network_operator

[23]. http://www.thehindu.com/news/cities/Madurai/amma-mobile-phones-for-shg-trainers/article8302327.ece

TOWARD AUTHENTICATED CALLER ID TRANSMISSION: THE NEED FOR A STANDARDIZED AUTHENTICATION SCHEME IN Q.731.3 CALLING LINE IDENTIFICATION PRESENTATION

Huahong Tu, Adam Doupé, Ziming Zhao, and Gail-Joon Ahn
Arizona State University
{tu, doupe, zzhao30, gahn}@asu.edu

ABSTRACT

The rising prevalence of phone fraud is hurting consumers and businesses. With about a half million reports each year in the United States, phone fraud complaints have more than doubled since 2013. In the current calling line identification presentation scheme, the caller ID is trivially spoofed. Scammers are using spoofed caller IDs to trick their victims into answering unwanted calls and further a variety of scams. To provide a solution to this problem, this paper proposes an authentication scheme that provides the possibility of a security indicator for the current Q.731.3 calling line identification presentation supplementary service. The goal of this proposal is to help prevent users from falling victim to phone impersonation scams, as well as provide a foundation for future defenses to stop unwanted calls based on the caller ID information. This work will help to guide the future development of a standardized scheme in authenticating SS7 identities.

Keywords— Caller ID, calling line identification, spoofing, fraud, scam, authentication, verification, standardization

1. INTRODUCTION

With the introduction of IP access to the Public Switched Telephone Network (PSTN), today the PSTN is rife with telephone spam, namely voice, voicemail, and SMS spam. Voice phishing, vishing, or phone fraud is a significant and rapidly growing problem in many countries, including the US [1] and UK [2].

To deal with this issue, governments, including the US [3] and UK [4], have enacted laws to restrict most forms of unwanted telephone calls. Furthermore, some governments have established regulatory agencies and telephone number registries that allow consumers to explicitly opt out of unwanted calls [5, 6].

In addition to government efforts, there are also consumer and business products that are made to defend against unwanted calls. In the consumer market, there are physical call-blocking devices for landline telephones, and various smartphone apps, that can block unwanted calls from offending caller IDs. Among business and network operators, there is a also supplementary network feature known as MCID (Malicious Call Identification) that allows the destination operator to request identification of the offending calling party.

Despite various efforts to reduce telephone spam, scam and robocalls, complaints on illegal calls has been making record numbers in recent years. According to a recent US government report, the number of phone fraud complaints in the US more than doubled in just a matter of two years from 2013 to 2015 [1]. The rise of phone scam is troubling, as billions are lost to phone scams each year [7]. In the US, more than 75% of the reported fraud and identity theft attempts are made over the phone [1]. Today, the US government receives about 200,000 robocall complaints every month, and the total number of reported complaints on illegal calls totaled more than 3.5 million in 2015 [8].

Clearly, all these countermeasures have so far failed at reducing the growth of telephone spam. According to a recent research [9], illegal callers today have access to various technologies aimed at circumventing call blockers and preventing identification. Among them, a practice known as *caller ID spoofing* is particularly effective at defeating call blockers, avoiding identification, and further a variety of scams.

To show an example of how caller ID spoofing is used in phone scams, one type of phone fraud that occurs frequently is the credit card verification scam, where the spammer spoofs the caller ID of a bank, and uses audio recorded directly from the credit card issuer to scam his recipients. The audio recording tells the recipients that their credit cards have been suspected of fraud, and is in need of verifying their personal information to reactivate their account. Of course, the true motive of this scam is to steal the recipients' credit card and personal information.

Furthermore, caller ID spoofing can also frame true owners of spoofed caller IDs with illegal behavior. When a malicious caller spoofs a known number to commit crimes, such as making scam calls or illegal purchase orders, or deceiving police into raiding a compound [10], true owners of spoofed caller IDs often end up questioned by law enforcement, and receive unfriendly calls for wrongdoings that have nothing to do with them.

The telephone number in North America and many other regions follows a numbering format that identifies the region code, central office code, and subscriber number [11]. If the telephone number is spoofed, law enforcement would lose key information that could identify and locate the offender. As most telephone spam defenses today (including law enforcement) rely on user feedbacks, caller ID spoofing has

made identification and user feedbacks completely irrelevant.

2. THE RISE OF CALLER ID SPOOFING

The caller ID is a generic name for a supplementary service offered by the called party's telephone company that presents the calling party's telephone number to the called party's user equipment during an incoming call. It helps the called party to decide whether to answer a call based on the caller's phone number, and, to call back the caller if the call could not be answered. Since its introduction in the 1990s, the caller ID service has now become ubiquitous in almost every telephone service. Today, the caller ID number is also used in other telephony services, such as the SMS and MMS, and, with the prevalence of smartphones, many smartphone apps and services also rely on the caller ID for identification.

However, because the PSTN was traditionally regarded as a closed trusted network, it was designed with little security in mind. Telephone companies rely on the trust in other operators to play by the rules. In the process of providing the caller's telephone number, the originating exchange can control what caller ID number is sent on a call-by-call basis.

Traditionally, a caller would need to gain control of a SS7 switch in order to have the capability to customize the caller ID. In consumer telephony services, the caller ID is typically managed by the caller's Local Exchange Carrier (LEC), preventing general users from spoofing the caller ID. It was also prohibitively expensive for individuals and small businesses to gain switch level access to the SS7 network, which kept the number of people with caller ID spoofing capability small.

However, with the recent rise of IP access to the PSTN, cheap IP-based client protocols (such as SIP [12]) are replacing the expensive traditional bulk telephone services (such as ISDN). Cheap and accessible Voice-over-IP (VoIP) bulk telephony services are now becoming the norm.

The PSTN is also moving toward being carried by the IP infrastructure (such as SIGTRAN [13]), however, the core SS7 signaling protocols have not changed to ensure compatibility with legacy systems. Telephone companies still relied upon trust in other switch operators to play by the rules. With growing IP access to the PSTN, the SS7 network is no longer exclusive to traditional telephone carriers. Today, there are now many internet telephony service providers (ITSPs) that provide bulk telephony services over an Internet connection. With the popularity of the cloud business model, access to SS7 switch level capability is becoming more available to untrusted parties. Some ITSPs *sell customizable caller ID as a service feature*, along with mass distribution technologies such as voice broadcasting, voicemail broadcasting, and SMS broadcasting, all provided over an Internet connection.

Further complicating matters, the Internet provides plenty of opportunities for a malicious caller to evade law enforcement through geography and technology. With an Internet connection, a spammer can now cost-effectively distribute outbound calls from an overseas location, beyond the jurisdiction of law enforcement. To further prevent identification, the spammer can hide behind virtual private networks (VPNs) and Tor networks to distribute the calls anonymously.

The PSTN has transformed from a closed national ecosystem to an open global ecosystem, therefore mutual trust and local laws can no longer be relied upon to materially guard against the abuse of SS7's inherent insecurities. There is a lack of accountability in phone identities. This is why we advocate for a standardized caller ID authentication scheme for the PSTN. By securing the caller ID, not only would consumers benefit from being able to distinguish between verified and unverified caller IDs, it provides a foundation for many telephony spam defenses (including law enforcement).

With the growing prevalence of phone fraud, calls from billing, government, and banking institutions would also greatly benefit from providing authenticity of their caller IDs, such that their customers would feel greatly assured doing business over the phone. Authenticated caller IDs may also be useful for immediate customer identity verification, without relying on (possibly stolen or guessable answers of) security questions to verify the identity of customers. As there are also scam calls that spoof the caller IDs of existing customers, which the malicious callers then trick the institution into emptying their customers' bank account [14].

However, for any viable deployment of such feature, it requires ITU-T standardization to ensure mutual interoperability. Therefore, standardization is key to building a PSTN ecosystem that could rely on the trust of caller IDs.

3. HOW CALLER ID SPOOFING WORKS

The SS7 process of providing the caller ID or calling party number (CPN), is known as Calling Line Identification Presentation (CLIP). In CLIP, the CPN is sent along with a call request using the initial address message (IAM) to the destination exchange of the called party. The relevant details of CLIP are defined in ITU-T Recommendation Q.731.3 [15], Q.81.1 [16], Q.951.3 [17], and I.251.3 [18].

The CPN is either provided by the originating local exchange or by the calling party, where the CPN parameter is inserted in the initial address message, which is sent as part of the basic call procedures according to Recommendation Q.764 [19]. The IAM routes through transit exchange switches until it reaches the destination exchange of the called party, in which the called party's local exchange carrier would convert and retransmit the CPN to a specific caller ID format for the called party's user equipment during the incoming call setup process.

The parameter value of the CPN is placed within the optional part of the initial address message. The IAM follows the ISUP (ISDN User Part) message format as defined in Q.763 [20]. The CPN parameter follows a structured binary coding format as defined in Q.763.3.10 [20].

To spoof the caller ID, the caller's originating exchange or the calling party will declare the CPN parameter with false information. In the US and many other jurisdictions, the caller's telephone service provider does not have any legal

obligation to ensure that the caller ID number is genuine before it is transmitted. Even in jurisdictions that forbid telephone service providers from providing falsely declared caller ID information, with Internet access to an untrustworthy telephone service provider, it is easy for a malicious caller to start the call request from a different origin, and transmit the false caller ID to the destination exchange of the called party.

4. WHY SECURITY INDICATORS MATTER

In the internet ecosystem, the HTTP and email are arguably the most popular types of communication used today. In HTTP communication, the universally recognized padlock indicator displayed in the address bar of modern web browsers (such as the one shown in Fig. 1) provides users with immediate trust in the web site's domain name identity.

Figure 1: An example of HTTPS security indicator in Google Chrome with extended verification

In email communication, the key-shaped security indicator of the email sender (such as the one shown in Fig. 2) in email clients provides the users with immediate trust in the identity of the email sender.

Figure 2: An example of email security indicator in Gmail

Figure 3: An example of proposed caller ID security indicator for an incoming call

These security indicators are crucial to informing the user that the information is from a verified source. The distinctive appearance of the security indicator provides an immediate cue of the authenticity of the sender's identity. The universality of the security indicator symbol provides an immediate cue of the functionality of the indication. By examining the authenticity of certificates that underpin the security indicator, users are able to protect themselves from phishing and impersonation scams.

Figure 4: An example of proposed caller ID security indicator for an incoming SMS

This is why having a security indicator can be an effective solution against caller ID spoofing. Examples of possible caller ID security indicators for incoming call and SMS are shown in Fig. 3 and 4. By having assurance in the security indicator, users can quickly determine if the sender is authentic by recognizing an icon. Furthermore, the prevalence of security indicators promotes awareness that the user should only trust senders that are verified, which may inspire them to be more vigilant of calls and messages from unverified sources.

5. DESIGNING THE CALLER ID AUTHENTICATION SCHEME

Before we discuss the technical detail of designing the underlying caller ID authentication scheme behind the security indicator, we first present an overview of the parties involved in the transmission of a call request.

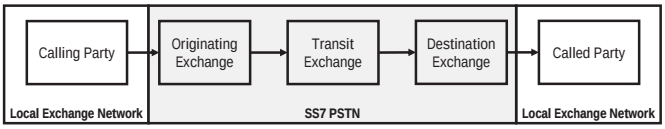


Figure 5: An overview of the parties involved in the transmission of a call request

Calling Party is the party initiating the call request with an user equipment (UE) or software client that connects with the originating exchange.

Originating Exchange is a switch in the PSTN that generates and transmits the IAM to the destination exchange pertaining to the call request from the calling party.

Transit Exchange is an interconnecting switch in the PSTN that helps to route the messages from the originating exchange to the destination exchange.

Destination Exchange is the terminating switch in the PSTN that receives the IAM and sets up the call with the called party.

Called Party is the party with an user equipment or software client of the intended called party for the call request.

In general, the sequences within a local exchange network define how user equipment interacts with the local exchange carrier during a call setup, and the sequences within the PSTN define how SS7 switches interact with each other during a call setup. More details of basic call control and signaling procedures can be found in Q.764.2 [19].

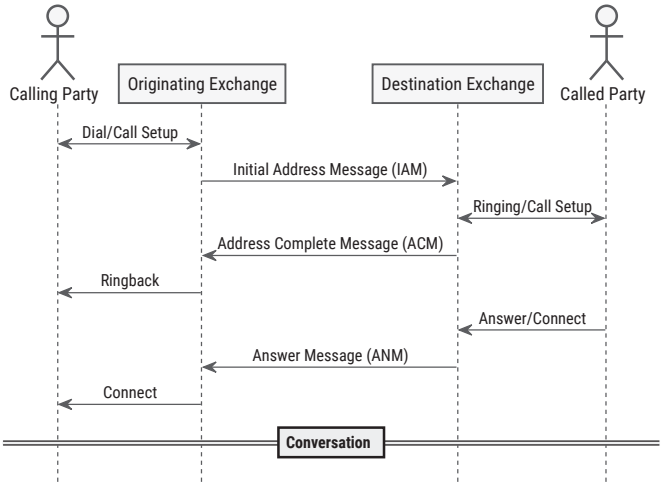


Figure 6: A simplified sequence of basic call control signaling

The current SS7 calling line identification presentation scheme has two fundamental insecurities: (1) a lack of verification of the declared caller ID and (2) a lack of integrity of the transmitted caller ID. The current calling line identification presentation scheme allows the CPN to be declared arbitrarily. Furthermore, there are currently no mechanisms to protect the CPN from unwanted modification during transmission. Even if the caller has proven that she indeed owns that phone number, an actor (perhaps in association with the caller) along the transit link may still intercept and alter the caller ID number.

Therefore, the design goal of a prospective caller ID authentication scheme must address the aforementioned fundamental security flaws, allowing the CPN to be verified before transmission, and making sure that the CPN is authenticated and can only be produced by the calling party or the originating exchange. The authentication scheme must provide a process that ensures the verified CPN can protect its integrity and guard against unwanted modification during transit. It must also be able to coexist with the existing call control signaling protocols for deployability.

When designing an authenticated caller ID scheme, an immediate idea is to model it after the SSL/TLS protocol of the Internet. However, this design, although can guarantee the authenticity and integrity of caller IDs, is ill-suited for the PSTN. After establishing an initial end-to-end connection with a TCP 3-way handshake, SSL/TLS authentication requires two additional round-trips (4-way handshake) to establish a secure connection. However, in a typical SS7 call request, this handshake is a one-way process, where the originating exchange sends an initial address message to the destination exchange to present the calling party's phone number identity. Implementing a naive SSL/TLS scheme would require SS7 call requests to support the multi-way handshake process, which is hard to adapt for the current SS7 scheme, and could potentially add delay to the call request process. In addition, SSL/TLS is designed for a client-server environment, which requires the server ("called party") to first acquire a certificate from a certificate authority (CA), whereas in the PSTN scenario, we are mainly concerned with authenticating the client ("calling party"). Finally, SSL/TLS is also designed to encrypt the data communication, which adds transport and processing overhead, whereas in our case, the primary goal is sender identity authentication instead of encrypting the conversation.

Therefore, we need to design an authentication scheme better suited for the PSTN. Designed as an initial reference, we propose a caller ID authentication scheme, which will guide and shape an authenticated calling line identification presentation process for the SS7 ecosystem.

The high-level idea of the protocol is that it takes advantage of the fact that *receiving* a message is proof of phone number ownership in the PSTN. The originating exchange or calling party first verifies with a certificate authority that the originating exchange or calling party actually owns the CPN by sending a message through the PSTN routing mechanism, and is issued a caller ID certificate. The originating exchange can then use this caller ID certificate to generate an authenticated call request by extending the parameters within the optional part of the IAM. The destination exchange or called party's user equipment then checks the validity of the authenticated call request, and presents the validated caller ID using a security indicator during the call setup to the called party.

The role of each actor with regards to the caller ID authentication scheme is as follows:

Certificate Authority is an entity in the PSTN that verifies phone number ownership and issues caller ID certificates to requesters that successfully provided proof of phone number ownership.

Calling Party sets up a call request with the originating exchange for the called party.

Originating Exchange obtains a caller ID certificate from the certificate authority for the calling party's phone number, if acquired, generates and transmits an authenticated IAM upon a call request from the calling party to the destination exchange.

Transit Exchange helps to route the IAM to the called party's destination exchange as usual.

Destination Exchange receives the authenticated IAM and checks the validity and authenticity of the call request, and it sets up the call with the called party with a security indicator showing the caller ID verification status.

Called Party receives the call request showing a security indicator.

The processes of the authentication scheme can be divided into 2 parts: Caller ID Verification and Authenticated Call Request.

In Caller ID Verification, the core process is sequenced as follows:

Prerequisites to the process: (1) the CA's public key P_S is publicly known, and (2) the CA has his private key Q_S.

1. Originating exchange or calling party generates a public-private key pair for the calling party's phone number, P_A and Q_A.

2. Originating exchange sends calling party's phone number $From_A$ and public key P_A to the CA.
3. CA creates an encrypted nonce $ENonce_S$ by first generating a random nonce $Nonce_S$ and then encrypting it with the calling party's public key. $ENonce_S$ = Encrypt(P_A){$Nonce_S$}. This ensures that only someone with the calling party's private key can decrypt $ENonce_S$.
4. CA signs the $ENonce_S$ to create a signature $ENonce\text{-}Sig_S$. This is to safeguard the authenticity of the nonce during transmission.
5. CA sends $ENonce_S$ and $ENonce\text{-}Sig_S$ to calling party's telephone number $From_A$. The phone number should route to the originating exchange or calling party.
6. Originating exchange verifies the signature $ENonce\text{-}Sig_S$ to ensure CA's identity.
7. If $ENonce\text{-}Sig_S$ is verified, the originating exchange decrypts $ENonce_S$ with private key Q_A to obtain $Nonce_S$.
8. Originating exchange sends decrypted $Nonce_S$ to CA, proving that the originating exchange/calling party is really the owner of the phone number and public key.
9. CA verifies $Nonce_S$ and, if valid, sets a short expiration time $Expiry_A$ and generates a caller ID certificate (CIC) for the calling party CIC_A by signing the calling party's phone number $From_A$, public key P_A, and $Expiry_A$ using the CA's private key.
10. CA sends CIC_A to originating exchange.

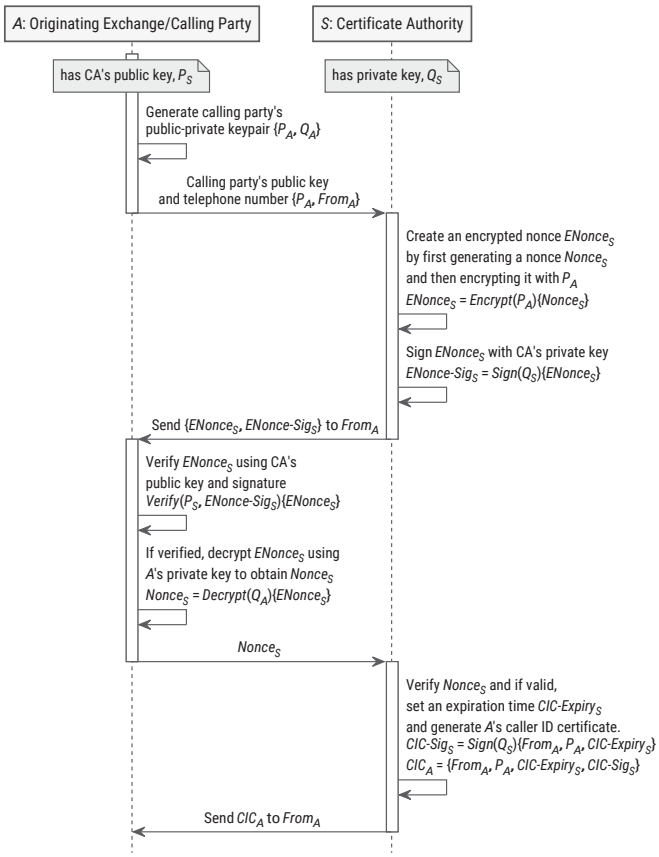


Figure 7: Sequence diagram showing the core steps to obtain a caller ID certificate

A sequence diagram of the Caller ID Verification process is shown in Figure 7.

In actual deployment, there can be several CAs, allowing different users, such as in different networks or regions, to verify with an appropriate CA.

With regards to the caller ID certificate format, the certificate could be based on ITU-T X.509 format [21], and the telephone number in the certificate could be based on international E.164 format [22]. The required critical extension field for the X.509 certificate could be as follows (in RFC5280 style [23]):

```
Extensions ::=
    SEQUENCE {intlPhoneNumber  E.164}
E.164 ::== PrintableString (SIZE (3..15))
```

In Authenticated Call Request, the core process is sequenced as follows:

Prerequisites: (1) the originating exchange has CA's public key P_S, and (2) the originating exchange has caller ID certificate CIC_A and his private key Q_A.

1. Originating exchange generates an IAM for the call request as usual.
2. Originating exchange generates an IAM Signature $IAM\text{-}Sig_A$ by signing all enclosed fields in the IAM along with current the current UTC timestamp $Time_A$. The inclusion of a UTC timestamp ensures that the call request is transient and unique with regards to time and destination, in order to guard against "cut and paste" and replay attacks.
3. Originating exchange attaches the UTC timestamp $Time_A$, IAM Signature $IAM\text{-}Sig_A$, and Caller ID Certificate CIC_A in the optional part of the IAM and sends the extended IAM to the destination exchange.
4. Destination exchange obtains the extended IAM and checks if CIC_A is valid, expired or revoked.
5. If the CIC_A is valid, verify IAM signature against all the enclosed fields.
6. If the IAM signature is valid, check if the UTC timestamp is valid (within a reasonable delay and clock drift), and check if the called party number is correct.
7. Setup the call request with the called party and present a security indicator for the verification result.
8. Destination exchange sends address complete message (ANM) with verification result back to the originating exchange.

A sequence diagram of the Authenticated Call Request process is shown in Figure 8.

Due to the one-way process of transmitting the authenticated call request in the IAM, the call verification process can be implemented adding negligible delay to the existing call

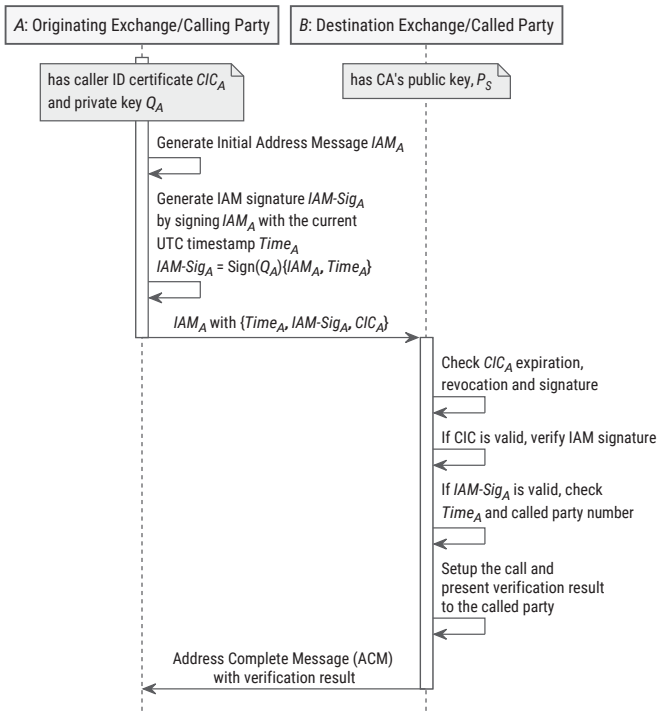


Figure 8: Sequence diagram showing the core steps to initiate an authenticated call request

setup process.

The existing parameters of the IAM is listed in Q.763 [20] Table 32. The proposed extended IAM parameters could be as follows in Table 1.

Table 1: Extended IAM parameters proposed

Parameter	Type	Length (octets)
UTC Timestamp	Optional Part	4-?
Signature Algorithm	Optional Part	1-?
Signature	Optional Part	16-?
Caller Identity Certificate	Optional Part	32-?

To ensure transit compatibility, the extended IAM would include a *Parameter Compatibility Information* parameter to instruct the existing transit exchanges to transfer the extended IAM parameters transparently to the destination exchange. The specifics of the Parameter Compatibility Information parameter can be found in Q.764.2.9.5.3.2 [19].

To inform the originating exchange that the authenticated call request has successfully pass verification at the destination exchange, we also recommend including a *Request Verification Status* parameter in the optional part of the address complete message to provide a feedback on the verification result. This would be useful for the originating exchange to determine if the extended IAM has been successfully verified by the destination exchange and make corrections if needed.

After the last step, the called party decides whether to answer the call request based on the caller ID and the verification result.

5.1. Security Considerations

Even as we outlined the reference scheme to authenticate the caller ID, we also need to assume that there is a constant threat of malicious actors stealing the caller's identity, such as by seizing control of the caller's phone number, or stealing the private key of the caller ID certificate. Cell phone theft is an ever-prevalent issue, and many users simply do not secure their cell phones. Furthermore, having a valid caller ID certificate does not imply that the caller should always be trusted. As a critical security measure, the certificate authority therefore must also be able to deal with revocations of a previously-issued certificate.

Learning from the pains of maintaining and distributing revoked certificates on the Internet, where Certificate Revocation Lists (CRLs) [24] have the disadvantage of being unboundedly bulky for a large number of revocations, and the alternative Online Certificate Status Protocol (OCSP) [25] has the disadvantage of requiring the receiving party to open a real-time connection with the issuer, potentially stalling the communication, therefore, we need to explore a different approach to handle certificate revocations in the PSTN. Furthermore, in many cases, we also need to assume that the victims of identity theft may not realize that they have been compromised, therefore revoking a certificate after an incident may not help.

With that in mind, we will provide some additional discussion on how certificate revocations should be handled in the PSTN. First, we recommend having the CA issue short-term caller ID certificates to limit the expiration period. We recommend limiting the certificate expiration to no more than 72 hours (inspired by the typical time needed to settle an automated clearing house transaction in the US). The certificate requester can request to have a certificate with an even shorter expiration. There are two notable benefits to having short-term certificates. First, it reduces the risk from a successful theft of the certificate private key or phone number by containing the impersonation threat within a bounded period. Second, it significantly reduces the size of revocation lists as the CAs would only need to assert or revoke unexpired certificates within a bounded period. Of course, the downside of having short-term certificates is that the caller ID certificates must be renewed frequently. However, unlike the Internet domain certificates where it can take hours due to a manual process and DNS propagation delay, caller ID certificate renewals would not have this problem because the process of verifying a telephone number can be fully automated and completed within seconds. Furthermore, because the amount of future certificate renewals is largely predictable, the CAs would be able to pre-adjust the quality of service to meet future demands, and perhaps even pre-generate some caller ID certificates to further improve service efficiency.

Second, we recommend having the CA issue caller ID certificates for conditional usage, such as by limiting the usage to a specific method of contact, or by excluding features such as call forwarding, SMS, MMS, etc. This further reduces the risk from a successful identity theft by containing the threat

to limited methods of contact. For instance, it is unlikely that a customer support department would need to contact individuals using SMS or MMS, hence, a successful theft of the bank's caller identity would force the attacker to use a live human or synthesized voice when contacting their victims, which could make the impersonation scam sound suspicious.

Finally, we recommend using CRL over OCSP when verifying revoked certificates. A phone call is more urgent compared to email and web communication, if a phone call is stalled by the certificate verification process, the calling party may assume that the called party cannot answer and hang up. It is important that the authentication scheme does not cause significant delays, otherwise some users may even choose to abandon security verification. CRL has an advantage over OCSP in this regard, because the revocation list can be cached at the destination exchange for immediate verification. Of course, the downside of CRL is that it does not receive real-time revocation updates, however, we believe that the risks can be mitigated by having the originating exchange or calling party choose to use even shorter-term certificates, and by having the destination exchange choose to update the revocation lists more frequently.

5.2. Local Deployment Considerations

As we outlined the process to verify the calling party number at the destination exchange, we also need to consider how the security indicator for the caller ID verification status would be transmitted and presented to the called party.

At the destination exchange, the local exchange carrier would present the caller ID verification status in a local call setup format (e.g., POTS, GSM, CDMA, UTMS, SIP, etc.). Each local exchange carrier would decide on the implementation of this presentation scheme, since they have full control over the vertical stack of network standards within their own network. An immediate thought is to simply implement the caller ID verification status as an indicator flag added to the existing caller ID format. However, this can be risky, we will provide some additional discussion on how it should be implemented.

In mobile telephone networks, popular technologies of which include the GSM (Global System for Mobile Communications), CDMA (Code Division Multiple Access), and LTE (Long Term Evolution). In these technologies, the caller ID is typically a parameter within the SETUP message transmitted to the called party's user equipment via an encrypted wireless signal. Assuming that the wireless transmission is well encrypted, a key consideration here is whether the identity of the base station is authenticated. In technologies that provide mutual authentication between the mobile phone and the base station, the presentation can be implemented as a flag indicator parameter, after performing the call verification at the destination exchange. However, in technologies where base station authentication is missing or flawed, the local exchange network should not use the flag indicator approach, because the verification status flag would be vulnerable to being spoofed by an attacker that could spoof a base station. If the call request can be spoofed by a fake base station (such as an IMSI-catcher), the verification status flag can also be spoofed by the fake base station. Instead, the presentation of caller ID verification status should be implemented as a full conversion of the extended IAM parameters, transmitted to the called party, to allow the called party's user equipment to perform verification of the authenticated call request.

In landline telephone services, the most popular technology of which is the POTS (Plain Old Telephone Service), the caller ID is a parameter within the header message encoded in SDMF (Single Data Message Format) or MDMF (Multiple Data Message Format), transmitted to the called party's telephone terminal in FSK (Frequency Shift Keying) signal between the first and second ring. Assuming that the connection to the central office exchange is secure (such as from physical protection), a key consideration here is whether the call request header is integrity protected. In POTS, the call request header is potentially vulnerable to "Orange box" attacks, where a malicious caller is able to alter the SDMF/MDMF header with spoofed FSK signals, as a result, the verification status flag would be vulnerable to being spoofed by the malicious caller. Hence, in such cases, the conversion should also be implemented as a full conversion of the extended IAM parameters to ensure that only the bona fide calling party can produce the authenticated call request.

Therefore, in summary, when implementing the presentation of the caller ID verification status at the local exchange network, only in scenarios where (1) the local exchange network connection is secured, (2) the identity of the local exchange carrier is authenticated, and (3) the call request header is integrity protected, should the local exchange carrier implement the presentation of verified caller ID as an indicator flag, otherwise, the conversion should be implemented as a full conversion of the extended IAM parameters to allow the called party's user equipment to perform verification of the call request.

6. RELATED WORKS

Peterson et al. [26] recently proposed an identity authentication mechanism for end users that originate SIP (Session Initiation Protocol) requests. The scheme proposes having the SIP proxies generating and inserting a PASSporT object [27] (a type of identity token) in the Identity header of every SIP request. Other than transport protocol and data format differences, the scheme uses a similar identity-token based mechanism in authenticating and verifying the caller identity. However, Peterson et al's proposal requires TLS connection for every communication, for reasons mentioned before, is difficult to adapt to the PSTN.

Reaves et al. [28] recently proposed an in-band modem for executing a TLS-inspired authentication protocol over the voice channel of the conversation. The modem is designed to overcome the challenges of low transmission bitrate due to voice codec and transmission losses. After the in-band modem established a data channel between the two parties

over the voice channel, the scheme uses a cryptographic challenge-response based scheme to verify the caller's identity. The scheme can provide strong security guarantees comparable to the TLS. However, the verification process require both parties' telephone terminals to support read-write access and live processing of the voice signals, which would require significant computation power on both parties' telephone terminals. It could also invoke privacy fears due to voice recording capability, and potentially add significant delay prior to the voice conversation.

7. CONCLUSION

With increasing abuse of PSTN's insecurities from untrusted parties, telephone spam, phone fraud and caller ID spoofing is poised to increase significantly. To ensure a sustainable future for the PSTN, the SS7 is in critical need of an upgrade of its core robustness. As a first step, we propose a caller ID authentication scheme for Q.731.3 calling line identification presentation. This work will serve as an inspiration for future standards to specify the verification processes and formats in authenticating SS7 identities.

REFERENCES

[1] Federal Trade Commission, "Consumer Sentinel Network Data Book for January - December 2015," 2016.

[2] Financial Fraud Action UK, "FFA - Fraud the Facts 2016," 2016.

[3] Federal Communications Commission, "Telephone Consumer Protection Act 47 U.S.C. 227," 1991.

[4] OFCOM: The Office of Communications, "Nuisance calls and messages," 2012.

[5] Federal Trade Commission, "National Do Not Call Registry," https://www.donotcall.gov/, 2016.

[6] TPS: Telephone Preference Service, "Register," http://www.tpsonline.org.uk/tps/number_type.html, 2016.

[7] TrueCaller, "Americans lost $8.6 billion in phone scams: Learn to protect yourself," http://www.truecaller.com/blog/americans-lost-86-billion-in-phone-scams-learn-to-protect-yourself, 2014.

[8] Federal Trade Commission, "National Do Not Call Registry Data Book for Fiscal Year 2015," 2016.

[9] Huahong Tu, Adam Doupé, Ziming Zhao, and Gail-Joon Ahn, "SoK: Everyone Hates Robocalls: A Survey of Techniques against Telephony Spam," in *Proceedings of the 37th IEEE Symposium on Security and Privacy*. IEEE, 2016.

[10] Jason Fagone, "The serial swatter," *The New York Times Magazine*, Nov 2015.

[11] NANPA : North American Numbering Plan Administration, "About the North American Numbering Plan," https://www.nationalnanpa.com/about_us/abt_nanp.html.

[12] Jonathan Rosenberg, Henning Schulzrinne, Gonzalo Camarillo, Alan Johnston, Jon Peterson, Robert Sparks, Mark Handley, and Eve Schooler, "Sip: session initiation protocol," Tech. Rep., 2002.

[13] K Morneault and J Pastor-Balbas, "Signaling System 7 (SS7) Message Transfer Part 3 (MTP3)-User Adaptation Layer (M3UA)," 2006.

[14] Dune Lawrence, "An Identity Thief Explains the Art of Emptying Your Bank Account," *Bloomberg Businessweek*, 2015.

[15] International Telecommunication Union, "Q.731.3 : Stage 3 description for number identification supplementary services using Signalling System No. 7 : Calling line identification presentation (CLIP)," 1993.

[16] International Telecommunication Union, "Q.81.1 : Stage 2 description for number identification supplementary services : Direct dialling-in," 1988.

[17] International Telecommunication Union, "Q.951.3 : Stage 3 description for number identification supplementary services using DSS 1 : Calling line identification presentation," 1993.

[18] International Telecommunication Union, "I.251.3 : Number identification supplementary services : Calling Line Identification Presentation," 1992.

[19] International Telecommunication Union, "Q.764 : Signalling System No. 7 - ISDN User Part signalling procedures," 1999.

[20] International Telecommunication Union, "Q.763 : Signalling System No. 7 - ISDN User Part formats and codes," 1999.

[21] International Telecommunication Union, "Information technology - Open Systems Interconnection - The Directory: Public-key and attribute certificate frameworks," 2012.

[22] International Telecommunication Union, "E.164 : The international public telecommunication numbering plan," 2010.

[23] D Cooper, S Santesson, S Farrell, S Boeyen, R Housley, and W Polk, "Internet X. 509 Public Key Infrastructure Certificate and CRL Profile," *RFC5280*, 2008.

[24] Russell Housley, W Polk, Warwick Ford, and David Solo, "Internet X. 509 public key infrastructure certificate and certificate revocation list (CRL) profile," Tech. Rep., 2002.

[25] Michael Myers, Rich Ankney, Ambarish Malpani, Slava Galperin, and Carlisle Adams, "X. 509 Internet public key infrastructure online certificate status protocol-OCSP," Tech. Rep., 1999.

[26] Cullen Jennings, Jon Peterson, and Eric Rescorla, "Authenticated Identity Management in the Session Initiation Protocol (SIP) draft-ietf-stir-rfc4474bis-09," *IETF*, 2016.

[27] Chris Wendt and Jon Peterson, "Persona Assertion Token draft-ietf-stir-passport-03," *IETF*, 2016.

[28] Bradley Reaves, Logan Blue, and Patrick Traynor, "Authloop: End-to-end cryptographic authentication for telephony over voice channels," in *25th USENIX Security Symposium*. 2016, USENIX Association.

SESSION 6

SUSTAINABILITY AND SMARTNESS

S6.1 Certified security systems for sustainable cities of the 21st century.

S6.2 WiFi networks on drones.

CERTIFIED SECURITY SYSTEMS FOR SUSTAINABLE CITIES OF THE 21ST CENTURY

Simone Wurster

Chair of Innovation Economics
Berlin University of Technology,
Berlin - Germany

Irene Kamara

Vrije Universiteit Brussel
Research Group on Law, Science,
Technology and Society (LSTS),
Brussels - Belgium

Thordis Sveinsdottir

Trilateral Research & Consulting,
London - United Kingdom

ABSTRACT

The United Nations formulated 17 sustainable development goals to "transform our world". Goal 11 aims to "make cities and human settlements inclusive, safe, resilient and sustainable". As such, security systems have become increasingly relevant, particularly in the past several years as significant and dangerous threats have emerged throughout the world. In addition to said risks, a number of security solutions, for example in the field of CCTV, are linked with significant privacy risks. Therefore, an appropriate certification scheme for security systems that not only considers security aspects but also additional issues, e.g. data protection and privacy, is needed in Europe. The EU Project CRISP (Evaluation and Certification Schemes for Security Products) aims to facilitate this process via the development of pan-European certification. This paper shows CRISP's solutions based on the current outcomes of the project and its specific contribution to research and practice.

Keywords— Security; conformity assessment; standards; certification; privacy

1. INTRODUCTION

The United Nations formulated 17 sustainable development goals to "transform our world". The 11th goal aims to "make cities and human settlements inclusive, safe, resilient and sustainable" and includes several sub-goals including the wish to provide access to safe and sustainable transport systems for all by 2030, and to substantially increase the number of cities "adopting and implementing integrated policies and plans towards (…) resource efficiency (and) resilience to disasters" by 2020 [1]. In this context, the present article refers to certification, certification schemes and seals for security systems. Certification is a "third-party attestation related to product, processes, systems or persons". A certification scheme is a "certification system related to specified products, to which the same specified requirements, specific rules and procedures apply" while a seal, mark or certification mark "indicates on a product, system or service, that it has been certified according to existing standards or manufactured according to specific, high standards" (see [2] and the sources listed there).

Effective and efficient high security systems that guarantee performance and respond to the emerging security threats in various areas, e.g. in the fields of critical infrastructure, citizen security and along borders, are required (see [3]). To provide assurance, a Pan-European certification scheme for security solutions is demanded ([4], [5]). Besides highlighting the need for a more harmonized regulatory framework in this area described by [5], [6] stresses that "(e)ffective civil security must embrace interoperability, standardization, certification (and) validation, (…) consultations on privacy issues and other factors that cut across public and private spheres (…)." Furthermore [4] highlights that "an improved infrastructure for validating and certifying security products and technologies would provide mechanisms that contribute to enhancing security within the EU and have the potential to enhance the competitiveness of the EU security industry (…)" (p. 17).

Beyond addressing core security issues appropriately, assurances to protect fundamental rights and freedoms are needed. As shown, privacy is an important concept in this regard. It is defined as "the condition of being protected from unwanted access by others – either physical access, personal information, or attention" ([7], p. 277). Considering privacy reflects that personal data protection constitute core values of democratic societies (see [8]) and fundamental rights ([9], art. 7 and 8).

As mentioned, security systems may have unintended consequences, some of which harms not society as a whole, but also the individuals comprising said society (see [10]). Moreover, there are risks of discrimination, mistrust and "social sorting" [11], resulting from the inappropriate and erroneous use of security technologies. Such solutions include, for example, advanced Closed Circuit Television (CCTV) also known as video-surveillance systems.

Video-surveillance systems are comprised of cameras, recorders, interconnections and displays that are used to monitor activities in a store, a company or more generally any specific infrastructure and/or a public place (definition for CCTV systems) [12]. Currently, "the CCTV surveillance market is undergoing a remarkable technology revolution" including "CCTV-Based Remote Biometric & Behavioural Suspect Detection" [13].

In addition to the challenge of providing security, CCTV systems also face significant privacy risks, see [14] for a

detailed overview. According to [15], 40% of European citizens think that CCTV invades privacy. Furthermore, the participants of a survey in [16] mentioned the particular risks associated with intelligent CCTV. Resulting problems are in particular risks of abuse, discrimination risks and possible intimidation effects (see [17]).

Concerns surrounding invasions of privacy can be concisely abridged as follows: Security systems collect a vast amount of data, more than necessary to achieve security goals, through constant operation in public and semi-public locations. The information is collected, stored and shared with both private and public bodies without citizen consent and quite often without even citizen awareness. This amounts to a serious threat to people's rights to respect for private life; protection of personal data, fair, non-discriminatory and equal treatment, presumption of innocence and due process. These trepidations are critical to the functionality of society and the well-being of individuals and must therefore be adequately addressed. The European General Data Protection Regulation 679/2016 (GDPR) [18] provides a recent response to these needs in articles 42 and 43, which detail a data protection certification mechanism with the objective to 'enhance transparency and compliance' with the Regulation in order to protect personal data.

In summary, there is not only a need to certify security aspects but also other issues of the relevant systems with privacy certification for CCTV systems being of particular importance, see [14], pp. 42-43. [14] describe the goals for such a solution in detail, e.g. to make the design and implementation of CCTV systems more transparent, to ensure effective control of CCTV systems and to boost privacy and data protection practice visibility for subjects of CCTV surveillance. In response to these needs, it is a priority of both the European Union and the EU Project CRISP[1] to cultivate a pan-European certificate for security solutions, which ensure protection and make considerations important societal needs. In response to the high number of identified privacy challenges, CCTV systems were chosen as CRISP's first application area.

2. PROTECTION IN THE SUSTAINABLE CITY – A MIX OF CHALLENGES TO BE ADDRESSED BY SECURITY SYSTEMS

Security solutions use is concomitant with numerous obstacles. As they relate to monitoring technologies, for instance, "Surveillance activities can be well-intentioned and bring benefits. [...] But unseen, uncontrolled or excessive surveillance activities also pose risks that go much further than affecting privacy. They can foster a climate of suspicion and undermine trust [19].

As a solution for privacy issues, [20] introduced a concept, defined as "Privacy by Design", demanding that privacy be taken into account throughout the entire engineering process of a technical solution. According to [20], seven core principles are included within this concept: 1. Proactive not Reactive; Preventative not Remedial, 2. Privacy as the Default Setting, 3. Privacy Embedded into Design, 4. Full Functionality - Positive-Sum, not Zero-Sum, 5. End-to-End Security - Full Lifecycle Protection, 6. Visibility and Transparency - Keep it Open and 7. Respect for User Privacy - Keep it User-Centric.

"Privacy by Design", or its variation "Data Protection by Design"[2], has become synonymous with a development method for privacy-friendly systems and services (see [8]), thereby going beyond mere technical solutions to address organizational procedures and business models. This approach requires the adoption and implementation of principles, such as proactive protection of privacy, full lifecycle protection, visibility and transparency and privacy embedded into the design of the product or system.

According to Article 25 of the GDPR, "Data Protection by Design and by Default", data protection must be included directly in processes, systems and products in the future [18]. [8] explains what may be understood as state of the art in this regard and analyses why privacy enhancing techniques have played a minimal role in practical applications.

Data Protection by Design suggests the protection of personal data throughout the whole lifecycle of the product, starting from the design process by building data protection safeguards in the product itself [18]. Another safeguard for these rights is the Privacy Impact Assessment (or Data Protection Impact Assessment)[3] [10], [21].

A need for trade-offs between privacy and security can be expected, i.e. compatibility of privacy and functionality. Nonetheless, [22] identified the solution of Positive-Sum Privacy, consisting of "a starting point (…), an evolutionary step and an assessment of (that) method: Starting point of a comparative evaluation is an outdated predecessor with less than full functionality and less than full privacy, both greater than zero. In an evolutionary step, a trade-off between privacy and functionality is acceptable if and only if it results in a positive-sum of functionality and privacy. The positive-sum has to be clear and not based on biased evaluation methods. If there is reasonable doubt, Positive-Sum Privacy is not fulfilled." [22], p. 267. The EU Charter of Fundamental Rights ([9]) in its art. 8 and 52, has its own balancing 'test' for trade-offs between rights, such as the right to protection of personal data and the right to security. The limitation of a fundamental right has to be provided by law, be necessary, proportionate and serve public interest. Methods and certificates, which build on an assessment of such dimensions, are missing to date.

Similar to challenges relating to privacy and trust, efficiency and security also represent competing goals. This may be due to the fact that, "Security is not usually an investment that provides profit" [23]. [24] go as far as referring to

[1] This research has been funded by the European Union's Seventh Framework Program for research, technological development and demonstration under grant agreement no 607941.

[2] "Data protection by design" is a legal obligation in art. 25 of the 679/2016 EU General Data Protection Regulation (GDPR) now.

[3] See article 35 of the GDPR.

security as "the enemy of efficiency". With the goals of these two concepts being so divergent, it is essential to develop a means for identifying the optimum allocation of funding for security, in addition to ratios, e.g. the return on security investment (see [23]). Furthermore, [15] assert that specification questions, for example, "How effective is CCTV?" and "What are the costs of CCTV?" should be used to analyze such systems. Moreover, matters of efficiency should also be examined as, e.g. "the mere presence of a (CCTV) system in a shopping mall does not mean that it is permanently used in a well-organised and efficient manner" (p. 38). Likewise, EU security research (e.g. the SIAM project, 2011-2014)[4] has shown that not only technical security requirements are significant for a security system, but also other issues, such as the trustworthiness of the system, the efficiency and to what extent it does not interfere with individual rights and freedoms. This indicates that trust, the protection of people's rights and efficiency are intrinsically linked. This aspect is an important factor in decisions to investment in security solutions and an actuality that the CRISP project, introduced in section 4, fully addresses.

3. STATE OF THE ART IN SECURITY AND PRIVACY CERTIFICATION IN EUROPE

As shown, security certification is faced with various challenges by not only assessing protection functionalities but also other factors, specifically privacy protection. Based on an evaluation of the current approaches and certificates to assess the protective nature of a security system, [5] stresses that none of them meets pan-European needs appropriately. Furthermore, summarizing the findings of [25] and [26] regarding privacy, [27] highlights that a related question is "(n)ot whether but how privacy should be protected." Flaws of current *privacy* seals are described in detail by [14]. They include, for example, a lack of transparency (including the criteria used to award seals and what is certified); close relationships between schemes and members; lack of regulatory oversight; lack of harmonization and common standards and conflicts of interest [28]. Furthermore, not surprising, no *security* seal considers privacy aspects appropriately so far although [29] highlights that, "It is not just isolated technological elements that must be labelled, but rather everything that depends on data." In addition, [14] analyzes the specific use of current privacy seals in the area of CCTV and stresses the need for alternative solutions. In summary, [28] formulated 14 requirements on best practice seals:

1. Valid and achievable scheme objectives, clear definition of scope
2. Harmonized rules for implementation of the scheme and use of the mark or logo
3. Additional legal rules to strengthen credibility of scheme and provide support
4. Robust (yet adaptable) certification criteria
5. Clear and uniform framework of standards and scheme criteria
6. Key stakeholder confidence and support
7. High quality, understandable, transparent, enforceable and globally accepted
8. Standards requirements must have a sound basis and meet needs
9. Rigorous application of standards
10. Dedicated body or bodies overseeing the scheme's design, implementation and innovation
11. Co-operation between a centralized entity controlling the scheme and relevant national bodies
12. Adequate and meaningful monitoring and enforcement
13. Measures for regular review, improvement and making innovations to the scheme
14. Sustainability.

Based on these criteria, the approach developed in the CRISP project will be discussed in section 6.

4. THE CRISP PROJECT, ITS GOALS AND METHODOLOGY

CRISP is a three-year security research project (April 2014 - March 2017), funded by the European Union. It aims to:

- Contribute to measures that increase citizen trust and confidence in security technologies through the evaluation and certification of social and legal impacts of security systems;
- Facilitate a more harmonized playing field for the European security industry by providing pan-European certification for security systems;
- Support the goal to provide protection in an efficient manner [30].

Security system end users, e.g. local authorities, emergency organizations, transport operators, law enforcement authorities, retail organizations, health organizations and educational organizations, represent CRISP's target group. To update the findings shown in the previous sections, a stakeholder needs analysis was performed during the research phase of the CRISP project, consisting of web surveys, semi-structured interviews[5] and the results were then validated in an expert workshop.

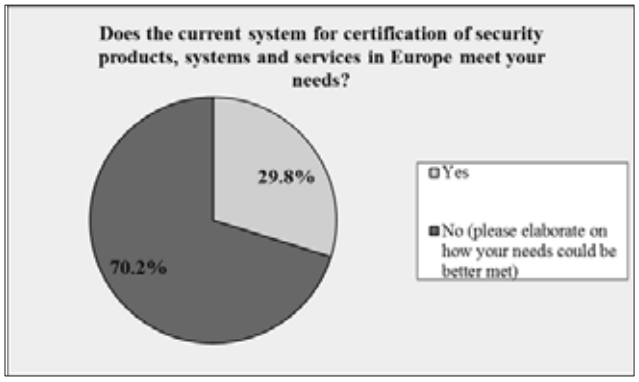
The web surveys were sent to an extensive contact list inclusive of security industry stakeholders, certification bodies, standardization bodies, security service industry bodies, accreditation bodies and consumer representative organizations across Europe. Low response rate should be taken into account when findings are read, but the findings were further supported by interviews and the experts present at the validation workshop. Key findings revealed that the most important benefit of certification identified by stakeholders, 85% of supply-side respondents and 82% of demand-side

[4] See website http://cordis.europa.eu/project/rcn/97990_en.html

[5] Full results of this work can be found in [31]

respondents, was that it offers a 'seal of quality' or assurance (see [31]). These findings resonated in the interviews, during which it was found that strong certification seals are perceived as a desirable supplement to branding. Furthermore, well-established schemes are believed to augment end user trust in products and services. From a survey, it was ascertained that the proposed CRISP scheme is viewed positively by supply and demand stakeholders. Moreover, an identified benefit would be a singular, recognizable European seal (71.4% and 54.3% respectively), a view which was further validated in interviews and the stakeholder workshop.

When asked whether the current security certification landscape met the needs of demand-side stakeholders, a majority stated that it does not (see Figure 1, which shows the view of the users of security systems) and this view was further supported by interview respondents and workshop participants. Key reasons stated were a lack of transparency for certification schemes and processes and too many competing schemes where differences were not clear. For manufacturers of security products, the need to re-certify for different national markets was seen as a costly and time consuming process that provides only unclear benefits. This extends the findings of [28] for privacy seals to the European *security* landscape.



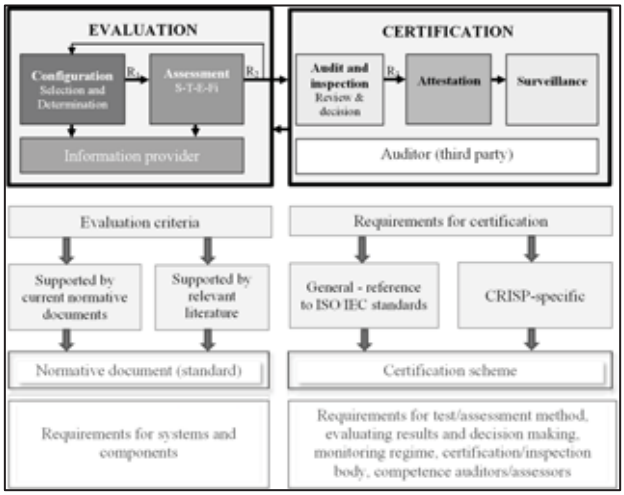
Source: [31], number of responses: 28

Figure 1: Stakeholders' answers on the question "Does certification meet the needs of demand-side stakeholders?"

Figure 2 provides a visualization of the means by which an innovative evaluation and certification methodology is be developed by the CRISP project.

Based on the challenges described earlier, the conclusion of the SIAM project and CRISP's stakeholder needs analysis, the CRISP project developed an innovative evaluation and certification methodology for security systems, that includes not only technical, but also socio-legal and economical requirements. When a security system is awarded certification by the CRISP scheme, evaluation in accordance with the S-T-E-Fi dimensions will be implicit:

- **Security**: the functionality of a security system in countering threats and reducing risks
- **Trust**: experiences and perceptions of the users of security systems, both employees and persons subject to scrutinizing, related to e.g. transparency, openness, fairness and accountability
- **Efficiency**: economical dimension of the related technology, related to e.g. the product life cycle costs, such as the purchasing costs, the implementation costs, interoperability and the operating costs
- **Freedom infringement**: impact of security systems on the freedoms and rights of persons, e.g. related to enhanced personal data collection and processing.



Source: modification of the CRISP project based on [32]

Figure 2: CRISP methodology

As shown in Figure 2, the CRISP methodology includes two phases: evaluation and certification. In the first phase, the configuration stage involves the collection of basic information by the 'information provider' from the part of the applicant, the determination of a relevant 'scenario' and the provision of evidence to the evaluation body to support the information on the system. After the first stage of the evaluation, information and evidence related to the security application area, the functionalities of the security system, the Technology Readiness Level and the technical specifications are (among others) known to the evaluation body. The second part of the evaluation stage includes the S-T-E-Fi assessment on the basis of predefined evaluation criteria based on legislation and technical standards. The criteria are related to the four S-T-E-Fi dimensions, but also interrelated to each other. One criterion for instance in the Trust dimension relates to the 'Information Signs' warning for the surveillance system. This criterion is also related to the Freedom Infringement dimension and the transparency of data processing criterion. The evaluation is performed by a group of experts in each of the four dimensions. This is therefore a participatory, bottom-up approach.

Following the S-T-E-Fi assessment stage, a report (R1) is issued with the results of the evaluation. The evaluation report, the provided evidence, and, if needed, on spot controls, are the basis of the audit and inspection part of the certification stage. Following the report of the audit and inspection component, there is an attestation stage and

finally the granting or rejection of the CRISP certificate. Quite important in terms of trustworthiness is the post-certification surveillance stage. The CRISP certification phase is based on ISO/IEC standards on conformity assessment, but also specific requirements tailored to the socio-legal dimensions of CRISP.

The CRISP methodology incorporates both technical, and non-technical parameters – something that has previously often been overlooked by other evaluation methods. Incorporation of social and legal components in the CRISP evaluation and certification will be initially piloted for use with video surveillance systems as a means of testing and refining the methodology. Upon the completion of the pilot phase, there is ample opportunity to expand the scheme to include additional types of security systems and products. It should be noted that CRISP builds on the accumulated experience and the existing security certification market. The technical specifications currently in place will not be redefined by the CRISP scheme, rather the fundamentals of S-T-E-Fi will be incorporated to provide the assessment for the incorporation of ethical and legal criteria for security system certification. The CRISP scheme will support the protection of fundamental rights, as well as encourage and stimulate compliance with pertinent EU laws, especially the GDPR 679/2016. Parties interested in procurement and those already using video surveillance may seek CRISP certification. In addition, CRISP evaluation may be implemented for systems still in development, albeit certification will not be granted until installation is completed as the context, use and environment in which the system is operated plays a substantial role in the assessment of social dimensions.

5. CRISP'S FURTHER STEPS TO CERTIFY EFFICIENT PRIVACY-AWARE SECURITY SYSTEMS

The current version of the S-T-E-Fi methodology builds on input from various stakeholders. The feedback loop, based on a workshop series, began in the second half of 2015 with a validation workshop that included a broad range of external participants who engaged in a series of scenario-based workshops. Four workshops with varying themes were scheduled and conducted in order to test the methodology. These activities resulted in the specification of CRISP's scope.

The key element of the CRISP implementation plan rests on the goal that an interested organization will take on the organization and the management of the future CRISP scheme. Using information garnered in the previously mentioned workshops, a certification manual for the impending CRISP organization was constructed, as well as a roadmap outlining the manner in which the CRISP scheme will be implemented. These activities were incorporated into a workshop on roadmap validation with actors from varying backgrounds, e.g. participants from the areas of certification and accreditation, regulatory and data protection authorities, consumer associations as well as standardization bodies, and a round table with certification bodies. Input from these participants allowed for further specification of CRISP's scope, upon which this article is based and a more solid alignment with the EU General Data Protection Regulation. The roadmap[6] for implementation is based on the aforementioned workshops.

The development of the CEN Workshop Agreement (CWA) "Guidelines for the evaluation of installed security systems, based on S-T-E-Fi criteria" is the next essential step in the methodology process. Its finalization is expected in March 2017. For more information on CRISP's roadmap, see [30]. Regarding the success of the future CRISP scheme, stakeholders identified the following challenges:

1. Lack of appeal to and/or added value for end users, resulting in poor market demand
2. Lack of acceptance from certification bodies
3. Lack of acceptance from regulators and policy makers, or conflicting with regulations and legislation
4. The fragmented nature of the security industry
5. Lack of regulatory oversight
6. Potential innovation gaps / Understanding difficulties

This also reflects the findings of [28] who formulated comparable challenges concerning privacy seals. To address these issues appropriately in the CRISP context, various action items shown in the Tables 1 and 2 (on the next page) were defined.

Table 1: Action items for the CRISP organization for the successful implementation of the CRISP scheme

CRISP organization (organization, which will manage the CRISP scheme in the future)
Ensure early engagement of all relevant stakeholders
Establish stable communication with experts on EU and Member States' law to create appropriate characteristics of the scheme and to update the project results
Complete the CRISP scheme appropriately by ensuring that it provides significant added value for the end users
Mobilize stakeholders for CRISP standardization activities
Ensure early engagement of regulators to ensure an appropriate development and refinement of the requirements of the scheme as well as greater promotion and acceptance
Negotiate the use of the CRISP scheme and mutual recognition agreements with European certification bodies
Conduct activities to sign licensing agreements
Conduct an appropriate marketing campaign to highlight the added value for end users in different markets
Create an appropriate training programme for evaluation bodies, certification bodies and other relevant stakeholders

The suggestions are supported by the relevant literature. Regarding policy makers, for example, ENISA stresses the need for their support by the development of new incentive mechanisms for privacy-friendly services and the need to promote them (see [8], iv).

[6] Available at: http://crispproject.eu/research-reports

Table 2: Action items for external stakeholders for the successful implementation of the CRISP scheme

Stakeholders
European Commission
Publicly advocate for and support the CRISP scheme in the EU Member States to facilitate positive framework conditions for the emergence of the CRISP organization
Develop a proposal for legislation regarding security certification with minimum requirements[7]
Industry/industry representatives and associations
Partake in CRISP's standards development
Support the formation of the scheme to ensure that it is fit-for-purpose
European standards bodies
Develop (a) standard(s) based on CRISP's CWA
National bodies: agree on starting such standardization, work on the common document, seek agreement on its content later
Members of relevant Technical Committees: check the need for additional standards in the CRISP context and initiate relevant activities where necessary
European certification bodies
Promote the CRISP scheme to the customers
Undertake trainings in CRISP-based assessment
Participate in the further development of the CRISP scheme
Partake in CRISP's standards development
Receive CRISP accreditation after the completion of the relevant standard(s)
European accreditation bodies
Monitor that there are no national differences in CRISP-related certification
Regulators
Oversee the CRISP certification so that it complies with legislation and regulation to the highest possible degree[8]
Participate in the development of the CWA and the CRISP standards
Member States
Conduct market surveillance, ensure conformity, and inform the public
Examine and identify whether the scheme posts any challenges to the national regulatory/legislative framework and care for appropriate solutions, if/where necessary

Source: own table based on [30]

6. CONTRIBUTION TO PRACTICE AND LITERATURE

Besides the existing need for a pan-European security certificate, [25] and [26], summarized by [27], stressed that it is not a question whether a European privacy seal should be created but how it will be created. CRISP provides an answer to this question by fulfilling a dual purpose: in the specific area of security systems it responds to the need for security certification and the need for privacy certification as well. Furthermore, it addresses additional needs for action in the areas of trust and efficiency. Regarding its contribution for practitioners, CRISP stands for the development of a pan-European certification scheme for security solutions, which will contribute to the harmonization of the European security market and the improvement of the framework conditions in the European market.

The CRISP scheme seeks to act as a passport for trade by eliminating the need for security systems to be recertified in each EU country in which they are sold. This means saving both costs and time. Providers of security systems who certify their solutions by the CRISP methodology would gain broader societal trust as the proposed scheme includes legal provisions and the enforcement of technical requirements and also the integration of social, consumer and human rights. Its benefits for end users include in particular: Integral protection through the use of the innovative S-T-E-Fi dimensions; Increased trust and confidence; Real gains in efficiency and cost-effectiveness; Additional safeguards exclusively designed for users; Prevention of breaches.[9] Additional benefits for this target group are described in detail by [33].

Section 3 of this paper showed 14 requirements of best practice schemes. CRISP's response is shown in Table 3:

Table 3: CRISP's response to good practice requirements

Requirement, CRISP's consideration		Comment
1	✓	Clear scope of pilot area and future activities defined
2	✓	Foreseen
3	✓	CRISP criteria build on EU regulation etc.
4	✓	Builds on the CWA and future standards
5	✓	Foreseen in CRISP's roadmap
6	✓	Builds on specific activities in the CRISP project and action items for the CRISP organization
7	✓	Builds on specific activities in the CRISP project and action items for the CRISP organization, currently with an EU focus
8	✓	Builds on various activities with different stakeholder groups in the CRISP project
9	✓	The CRISP concept builds on the rigorous application of standards
10	✓	Based on CRISP's roadmap
11	✓	Foreseen
12	✓	Foreseen
13	✓	Foreseen
14	n.a.	Has to be analyzed in the future

Furthermore, the CRISP concept addresses specific challenges: conflicts of interest between the requirement of providing operative security services and the preservation of fundamental rights, especially the rights to privacy and personal data protection, which are at risk as a result of data

[7] See [10], p. 23. on the lack of existing legislation in the field

[8] E.g., to ensure compliance with data protection requirements

[9] This description of CRISP's benefits is taken from several briefing papers created in the project's work package 7 for different stakeholder groups to which the authors contributed. The documents are available at: http://crispproject.eu/crisp-briefing-papers/.

collection by security and ICT systems, CCTV systems included. In addition, as stated by [34], security methods, which encroach upon basic human rights, the right to respect for private life, for example, require adequate safeguards in order to balance the encroachments (see [34]). However, despite the fact that the Privacy by Design concept has found its way into European legislation via the GDPR, requirements for its concrete implementation are still lacking (see [8], iii). Moreover, it is essential that security solutions are capable of responding to economic pressures and assessment solutions to ensure the fulfilment of this requirement are needed. All these apprehensions will be respected by the CRISP certification as a means of risk prevention through the development of a demanding certification process, which will work to protect the fundamental rights of non-discrimination, presumption of innocence, bodily integrity privacy and personal data. The evaluation component of the future CRISP scheme allows for an assessment of both security functions and privacy issues, as well as trust and efficiency aspects in the design stage of a security system. Accordingly, CRISP certification will show how an installed system meets the requirements in these four areas.

The CRISP approach also demonstrates how evaluation concepts can be exploited by standardization and how their quality can be raised by open development processes which meet the quality requirements of CEN. It also provides a new example on how the development of CWAs can be used for the exploitation of EU projects. The results of the future CEN workshops will also give opportunities to derive lessons learned, which can be used for the application of CRISP in future areas and its further standardization efforts. To ensure a permanent high level of quality of the scheme, regular reviews of the compliance and updating auditing processes will be important (‘surveillance’ phase).

7. CONCLUSION AND OUTLOOK

As shown, CRISP provides an opportunity to certify security solutions regarding security, privacy protection and the two other important aspects trust and efficiency simultaneously, with the long-term vision to contribute to more sustainable and resilient societies. Its successful implementation also requires several specific external measures. In addition, [35] highlights the particular importance of European data protection authorities reaching a consensus on specific requirements and recommended practices. This reflects the importance of various action items, which were specified for the implementation of the future CRISP scheme. The CRISP project ends in March 2017. Currently the CRISP partners are preparing to further develop the methodology along with a panel of interested parties at a CEN workshop with the aim of issuing a CWA, which is a best practice document for further standardization efforts and essential input for the development of a future certification scheme. The concrete development of the certification scheme lies outside the remit of the CRISP project, which ends in March 2017. The CRISP consortium will, on completion of the project, seek interested parties for developing the CRISP methodology and CWA into a fully-fledged certification scheme.

As shown, CRISP will first be piloted in the area of CCTV systems with other areas to follow. According to the results of a recent CRISP workshop with European Data Protection Authorities, CRISP’s concept of a holistic approach might be a pertinent challenge in this regard. Further research is recommended as to how this goal could be reached. Linked with the response to the various security and privacy-related requirements, additional issues emerge. [36] notes the slow implementation of privacy techniques as a significant worldwide problem. Therefore, policy makers should actively support incentive measures, such as audits and seals of approval. This goes much further than CRISP’s application field in the security area requires. Many data protection techniques are still not known by developers and engineers. Therefore, research results have to be communicated better in the future. Development tools that enable intuitive and simple implementation of data protection requirements should be provided (see [35]).

REFERENCES

[1] United Nations, “Sustainable Development Goals,” 2015. [Online]. Available: https://sustainabledevelopment.un.org/topics/sustainabledevelopmentgoals. [Accessed 04 07 2015].

[2] Fritz, F., Kreissl, R., von Laufenberg, et al., “Glossary of security products and systems. Deliverable D 1.1 of the CRISP project V2,” http://crispproject.eu/wp-content/uploads/2016/05/CRISP-Deliverable-1.1_Glossary-of-security-products-and-systems_REVISED.pdf.

[3] European Commission, “Programming Mandate Addressed to CEN, CENELEC and ETSI to Establish Security Standards, M/487,” 2011.

[4] European Security Research & Innovation Forum, “ESRIF Final Report,” 2009.

[5] ECORYS, “Security Regulation, Conformity Assessment & Certification, Final Report-Volume I: Main Report, Brussels,” 2011.

[6] Commission of the European Communities, “Communication from the Commission - "A European security research and innovation agenda - Commission's initial position on ESRIF's key findings and recommendations", COM/2009/0691 final,” 2009. [Online]. Available: http://eur-lex.europa.eu/legal-content/ EN/ TXT/HTML/?uri=CELEX:52009DC0691&from=EN.

[7] L. Wilkins and C. G. Christians, The Handbook of Mass Media Ethics, New York: Routledge, 2008.

[8] ENISA, “Privacy and Data Protection by Design,” 2014. [Online]. Available: https://www.enisa.europa.eu/publications/privacy-and-data-protection-by-design. [Accessed 04 07 2016].

[9] European Union, *Charter of Fundamental Rights of the European Union*, Official Journal of the European Communities, OJ C 364/01 (18.12.2001), 2000.

[10] Kamara, I.; De Hert, P.; Tanas, A.; et al., *Legal Analy-*

sis of existing schemes, Deliverable D4.1 of the CRISP project, 15 February 2016.

[11] D. Lyon, Surveillance as social sorting. Privacy, risk, and digital discrimination, New York: Routledge, 2003.

[12] ISO, *ISO 22311:2012 Societal security -- Video-surveillance -- Export interoperability,* http://www.iso.org/iso/catalogue_detail.htm?csnumber=53467, 2012.

[13] Homeland Security Research Corporation, "CCTV Based Biometric Standoff Suspect Detection Markets," 2013. [Online]. Available: http://marketing.homelandsecurityresearch.com/HSRC/CCTV/. [Accessed 15 04 2016].

[14] European Commission, Joint Research Centre, "EU privacy seals project. Comparison with other EU certification schemes: final report study deliverable 3.4," 2013.

[15] L. Hempel and E. Toepfer, "CCTV in Europe. Final Report of the EC-funded project Urban Eye," 2004. [Online]. Available: http://www.urbaneye.net/results/ue_wp15.pdf. [Accessed 15 06 2016].

[16] S. Wurster, "Security technologies for the protection of critical infrastructures – ethical risks and solutions offered by standardization," in *ITU Kaleidoscope Proceedings*, 2013, pp. 1-8.

[17] T. Wuerttemberger, "Rechtswissenschaftliche Begleitforschung zur intelligenten Videoüberwachung. BMBF-Innovationsforum "Zivile Sicherheit"," 2012. [Online]. Available: http://www.bmbf.de/pubRD/B1-I_Wuerttemberger_Redemanuskript.pdf. [Accessed 27 08 2012].

[18] European Commission, *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 L 119/1,* 2016.

[19] WP29, "Opinion 1/2007 on the Green Paper on Detection Technologies in the Work of Law Enforcement, Customs and other Security Authorities.," 2007. [Online]. Available: http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2007/wp129_en.pdf. [Accessed 04 07 2016].

[20] A. Cavoukian, *Privacy by Design – The 7 Foundational Principles,* https://www.ipc.on.ca/images/resources/7foundationalprinciples.pdf, 2011.

[21] D. Wright and P. De Hert, Privacy Impact Assessment (Law, Governance and Technology Series), Dordrecht, Heidelberg, London, New York: Springer, 2012.

[22] Bier, C.; Birnstill, P., P.; Krempel, E.; et al., "How Is Positive-Sum Privacy Feasible?," in *Future Security. Volume 318 of Communications in Computer and Information Science*, 2012, pp. 265-268.

[23] ENISA, "Introduction to return on security investment," 2012. [Online]. Available: https://www.enisa.europa.eu/publications/introduction-to-return-on-security-investment. [Accessed 01 07 2016].

[24] W. L. Garrison and D. M. Levinson, The Transportation Experience: Policy, Planning, and Deployment, Oxford, New York, div.: Oxford University Press, 2014.

[25] D. Solove, Nothing to Hide: The False Tradeoff between Privacy and Security, New Haven and London: Yale University Press, 2007.

[26] V. Reding, *Women and the Web – Why Data Protection and Diversity belong together,* Speech at the DLD Women Conference 2013, 2013.

[27] C. Pauner, *Certification, standardisation and data protection, CRISP- Setting the context,* Presentation at "Towards embedding data protection and other socio-legal requirements in the certification of security systems", Workshop for DPAs and data protection experts, AEPD, Madrid (Spain), 2016.

[28] D. Barnard-Wills, *Privacy Seals Research,* Presentation at CRISP workshop for DPAs, Madrid 30 Sept 2016, 2016.

[29] CNIL, "Privacy Towards 2020: Expert Views, IP Reports, Innovation and Foresight. 1," 2012. [Online]. Available: https://www.cnil.fr/sites/default/ files/ typo/document/CAHIER_IP_EN2.pdf.

[30] Wurster, S.; Burnik, J.; Tomšič, A.; et al., "Final Roadmap and Implementation Plan. Deliverable D 6.1 of the CRISP project," 30 June 2016.

[31] Sveinsdottir, T.; Finn, R.; Wadhwa, K.; et al., *D3.1 Stakeholder Analysis, Deliverable 3.1 V2, CRISP Project, 15 February 2016,* http://crispproject.eu/wp-content/uploads/2015/03/CRISP-D3.1-Stakeholder-Analysis-FINAL.pdf, 2016.

[32] Hempel, L.; Hirschmann, N.; Haponava, T.; et al., "Report on the scenario-based workshops and the refinement of the CRISP Methodology, Deliverable 5.2 of the CRISP Project," 1 February 2016.

[33] S. Wurster, N. Hirschmann, I. Kamara, et al., "Certified Systems to Reduce Security Risks in Modern Societies and the Contribution of the CRISP Approach," in *Proceedings of the IDRC conference 2016*, Davos, Switzerland, Global Risk Forum GRF, 2016, pp. 641-644.

[34] M. Hildebrandt, "Balance or Trade-off? Online Security Technologies and Fundamental Rights," *Philosophy & Technology,* vol. 25, no. 4, pp. 357-379, 2013.

[35] Heise, ""Privacy by Design": EU-Sicherheitsbehörde legt Empfehlungen vor," 2015. [Online]. Available: http://www.heise.de/newsticker/meldung/Privacy-by-Design-EU-Sicherheitsbehoerde-legt-Empfehlungen-vor-2517870.html. [Accessed 08 07 2016].

[36] ENISA, *Privacy and Data Protection by Design,* https://www.enisa.europa.eu/publications/privacy-and-data-protection-by-design, 2014.

WiFi NETWORKS ON DRONES

Antonio Guillen-Perez[1], Ramon Sanchez-Iborra[1, 2], Maria-Dolores Cano[1], Juan Carlos Sanchez-Aarnoutse[1] and Joan Garcia-Haro[1]

[1] Dept. of Information Technologies and Communications, Universidad Politécnica de Cartagena, Cartagena, Spain

[2] Dept. of Informatics, Universidad Técnica Federico Santa María, Santiago, Chile

ABSTRACT

The huge growth in the number of connected wireless devices leads to an increasing demand for network connectivity. In this context, aerial networks may play an important role by widening the concept of access networks. This paper describes and analyzes one of the most promising applications of Unmanned Aerial Vehicles, commonly known as drones, in the field of communications: Extending the capacity or coverage of wireless systems through the deployment of aerial communication networks. We present a comprehensive characterization study of an experimental system to deploy an aerial WiFi network. To do so, an Intel Galileo development board is appropriately configured and equipped as a WiFi node playing either the role of an access point in the infrastructure mode or of an intermediate hop in the ad-hoc operational mode. This device is then integrated onboard a drone. We compare both WiFi modes in terms of coverage area, throughput, and energy efficiency. Preliminary results reveal that there is a trade-off between coverage and data rates, for which the infrastructure mode performs better, and energy efficiency, where the ad-hoc mode is more responsive.

Keywords— Aerial network, UAV, drones, WiFi, low altitude platform, network access.

1. INTRODUCTION

Initially seen as an optional commodity, network connectivity has come to be considered an essential utility. There is not only a growing number of devices (e.g., smart phones, bracelets, wearables, different-nature sensors, etc.) that need to be almost continuously connected because of the advent of the Internet of Things (IoT) paradigm, but societal changes are also posing new challenges into the telecommunications arena. As a surprising example, it has been recently published in the news the need of Internet access that refugees request when they get to the refugees' camps. Among their first questions asked when they arrive was "when will we get WiFi?", and far from frivolous, this is a need for them as important as being fed [1]. Therefore, we face a world scenario with a tremendous growth of connected devices and redoubling demand of network access or connectivity.

This work was supported by the MINECO/FEDER project grant TEC2013-47016-C2-2-R (COINS).

On the other hand, UAV (Unmanned Aerial Vehicles) have recently drawn the interest of the research community. Due to the strong efforts made to improve their performance, e.g., miniaturization, energy efficiency, etc., UAV have become a useful tool widespread in different disciplines. These devices, also known as drones, provide disruptive applications in areas such as military, logistics, environmental monitoring [2], or rescue activities. For instance, drones eliminate the need of human presence to accomplish dangerous tasks and hence the risks that some of these activities entail are dramatically reduced. In the telecommunications field, one of the most promising applications of UAV is to use them as support equipment, aimed at extending the capacity or coverage of wireless systems through the deployment of an aerial communication network [3]. An interesting strategy used to achieve this goal is to assemble a light development board, e.g., Raspberry Pi [4], Intel Galileo [5], etc., onboard a UAV. Their low power consumption and the high number of connectivity elements that they support (communication chips, antennas, etc.) are among the most valuable features that these cheap single-board computers offer to perform this task. Thus, these devices are able to deploy a wireless network acting as network nodes within the system architecture, and allowing the end-users/things to gain connectivity through them. WiFi is so far the most widespread access network for providing connectivity to end-users' wireless devices. As specified by the standard (IEEE 802.11), it provides two different modes of operation: the infrastructure mode and the ad-hoc mode, with their corresponding pros and cons depending on the application scenario. Both modes are provided by all versions of the WiFi standard family (IEEE 802.11a/b/g/n/ac), which has increased its maximum transmission rate almost exponentially, using the 2.4 GHz (802.11b/g/n) as well as the 5 GHz (802.11a/ac) Industrial, Scientific, and Medical (ISM) radio bands.

In this paper, we present a complete characterization study of an experimental system for deploying an aerial WiFi network. To this purpose, we use the Intel Galileo development board integrated into a UAV. First, we conduct a theoretical coverage study by means of two suitable radio propagation models for the scenario under consideration. Second, we carry out an experimental deployment of the aerial network and explore both, the infrastructure or access point-based (AP) and the ad-hoc

modes. Then, we compare the maximum transmission rates and the coverage achieved for each operational mode and discuss which one better fits considering the system requirements. Results attained from the working system in a real flying experiment are included and compared with those obtained theoretically and from a controlled lab test-bench. Finally, we present an experimental study regarding the energy efficiency of the development board. To this end, the instantaneous current consumption obtained for the different modes of operation, in both idle state and transmitting under different-rate data flows is measured.

The rest of the paper is organized as follows. In Section 2, we review the most prominent works regarding the deployment of aerial networks using UAV. The methodology and equipment used in this work is described in Section 3. Section 4 presents and comparatively discusses the results obtained from both theoretical and experimental studies. Finally, Section 5 concludes.

2. RELATED WORK

According to recent estimates of EUROCONTROL (the European Organization for the safety of air navigation) [6], we are already living with some two million UAV around the world considering all their types, from toys to large military vehicles. Taking into account that there are around 200,000 manned aircraft, and the fact that the safety of these last aircrafts is a must, the European Aviation Safety Agency (EASA) is still trying to reach the objective established in December 2013 by the European Council in order to develop a strategy to support the progressive development of the UAV market in Europe. This regulatory framework stated a common regulation within all the EU countries for UAV with an operating mass of 150 Kg or more. However, the use of civil UAV below this weight should be regulated by individual Member States of the EU. In this last scenario, the Spanish Government has been doing its homework and can be considered as a pioneer: AESA (the Spanish National Agency for Air Security) published in April 2014 a regulation draft for the use of UAV [7] and it is expected to be updated shortly. The reasons that explain this leading position are varied: UAV technology is available in Spain, the aeronautic industry is well positioned, and there are appropriate weather and population conditions (low density, large extensions, etc.), among other factors.

Focusing on communications, UAV have been used to deploy air networks in different fields of action [8, 9], and research in this area is increasing. Drones usually operate on the ISM, IEEE-S, and IEEE-L bands. Due to the proliferation of wireless devices that also work in these bands, Saleem *et al.* suggested in [10] the possibility of using Cognitive Radio technology in UAV communications as a way to solve the spectrum scarcity problem. Nevertheless, using cognitive radio results in additional challenges to be solved at upper layers of the communication architecture, as indicated by the authors. Routing in aerial ad-hoc networks is addressed in [11]. Routing is challenging in this scenario mainly because of UAV high mobility and the fact that most ad-hoc routing protocols have been studied following planar graph-based techniques not suitable in 3D (as it would be the case of UAV networks). Their routing proposal seems effective in this environment, although as the authors stated, it still needs some refinement to improve its performance. UAV networks have been also tackled from the perspective of Delay Tolerant Networks. The work in [12] introduced a so-called Autonomous Flight Wireless Node to deliver data under poor network conditions from one location to another using an epidemic delay-tolerant routing approach. Other research focus has been flying planning [13, 14]. For instance, in [14], authors proposed to leverage public wireless communication infrastructures to connect UAV. They included a planning method so that UAV only fly within a 3G coverage area, i.e., avoiding non-signal zones, with interesting results.

Similarly, several initiatives have been announced using UAV to provide broadband connectivity. For instance, the European projects ABSOLUTE, ANCHORS, and AVIGLE. When a UAV is exploited with this aim, the term Low Altitude Platform (LAP) can be also employed. The work done by Säe *et al.* [15] studied the coverage of temporary WiFi networks built using drones as access points. The authors only used computer simulation to theoretically evaluate the coverage area by applying a deterministic radio propagation model called the Dominant Path Prediction Model. As we will discuss in next sections, the lack of experimental results in Säe *et al.*'s work is a notable shortcoming since their results will likely not match with real measurements. Likewise, the work in [16] presented results about the use of the IEEE 802.11n and 802.11ac standards in aerial WiFi networks. These protocols were used for communication among UAV, between a UAV and a ground station, and to provide terrestrial coverage. They operated drones flying at an altitude of 50 m and the ground station was located at a height of 2 m. Their experimental results revealed that greater throughputs can be achieved by using 802.11ac when compared to IEEE802.11n, although different performances are achieved for 802.11 depending on the used driver, so further experimental research is needed to better understand this behavior. Please note that in addition to use different versions of the WiFi standard, in our work we will use the UAV as an intermediate node between two ground communication endpoints, whereas in [16] the communication endpoints were the UAV and the ground station. As we will see in the results section, this detail has an important impact on the measured performance metrics.

The trend of mixing cellular communications with UAV networks to extend connectivity has been also recently addressed. For example, authors proposed in [17] to transform the UAV into a 5G base station, so that a heterogeneous network (known as HetNet in 5G

terminology) can be formed. By means of computer simulation authors showed that it is possible to improve coverage and capacity by deploying this system. Nevertheless, further experimental work is needed to corroborate these promising results. Finally, the energy efficiency of these systems was evaluated in [18]. With an extensive analytical study, authors demonstrated that the use of passive scanning for the mobiles and the periodic beaconing for UAV is possible to optimize the drones' energy consumption maintaining high connectivity.

3. METHODOLOGY AND EQUIPMENT

In this section, we describe the methodology and the main features of the devices employed to perform this study. The Intel Galileo board was used as the central element of the system. This development board is based on the Quark SoC X1000 Intel 32-bit processor at 400MHz. Specifically, the Intel Galileo Generation 1 was used in the experiments conducted. This board ran a well-suited version of Linux (Linux quark 3.19.8 yocto-standard). This Linux image is based on BSP 1.2.0 [19] (iot-devkit, Intel IoT Development Kit) with Kernel 3.19.8 of Linux and several patches that built a consistent platform to provide the different communication configurations as described in this work. Regarding the Galileo board power supply, an external battery of 10400 mAh was used. This battery allows up to 15 hours of Galileo's working time, depending on the operational mode and the traffic load. More information about the lifetime will be provided in the results section. The Intel Dual Band Wireless-AC 7260 wireless card was connected to the Galileo board through its PCI Express port. This network card allows for connections up to 867 Mbps, supporting several WiFi standard versions, namely, IEEE 802.11a/b/g/n/ac. This wireless card can operate as an access point, or as an ad-hoc node within a mesh network. In addition, we also used two 5 dBi omnidirectional gain external antennas providing large coverage range. The total weight of the board, the battery, and the antennas is approximately 340 g. Then, we illustrate the methodology followed. Our work was divided into three phases. First, we developed a theoretical study of a UAV coverage area equipped with the Intel Galileo board acting as a WiFi node. We assumed an open-air scenario and selected two radio propagation models: Free Space (or Friis) model [20] and the Wireless World Initiative New Radio (WINNER) D1 model [21]. Based on these models, we calculated the maximum expected uplink and downlink radio coverage for several versions of the WiFi standard.

Second, we experimentally tested the performance of the Galileo board as an intermediate node within a WiFi network. These tests were performed first in a controlled lab test-bench (static on-the-ground) and then, in a real aerial deployment. In both cases, we studied the two main operational modes of the IEEE 802.11. The usual mode of operation is the infrastructure mode, which is implemented in most commercial WiFi routers. This mode deploys a network within the coverage area of a central device (the access point, AP), which interconnects all nodes composing the system and acts as a gateway towards an external network (e.g., Internet). Thus, the AP assumes all management tasks of the WiFi network. On the other hand, the ad-hoc mode is a less extended operational mode for WiFi networks. This mode of operation lacks of a central point, so all nodes composing the network connect each-other forming a mesh network. In this case, each node simultaneously assumes both client/host and router tasks, hence network nodes periodically scan for retrieving information about their neighboring nodes. This ad-hoc mode allows the dynamic routing of data-traffic, eliminating the coverage constraint imposed by the access point in the infrastructure mode. In addition, if the ad-hoc network nodes can move, then these networks are called Mobile Ad-hoc NETworks (MANET), which means that links between nodes may be dynamically dropped or created due to nodes movement. An efficient operation of ad-hoc networks strongly relies on the routing algorithm, which is responsible for an effective information delivery towards its destination. Due to the dynamic nature of the network topology, the cost of the links should be periodically evaluated to search new routes and hence choosing the best paths depending on the provided service.

Therefore, in order to test these two WiFi modes, two additional end-devices were included in the system. These nodes were two Linux Ubuntu laptops with wireless cards compatible with the IEEE 802.11 a/b/g/n standard. At the receiving endpoint of the communication, a laptop with the Intel Centrino Advance-N 6230 wireless card was located, and another laptop with the Intel Dual Band 3160 wireless card was employed in the other communication endpoint (transmitter). The Intel Galileo board was always the central point of the system, either working as the AP (infrastructure) or as the intermediate node (ad-hoc) between transmitter and receiver. In the latter, the well-known ad-hoc routing protocol BATMAN [22] was employed. This protocol has already shown a superior performance than other important ad-hoc routing protocols under heavy traffic loads [23]. We used the iPerf3 tool to obtain network metrics. iPerf3 performs active measurements for determining the maximum achievable bandwidth on IP networks, among other metrics such as packet loss or delay. It supports the set-up of several transmission parameters and network protocols. For this study, we used Constant Bit Rate (CBR) transmissions between the communication endpoints with a duration of 30 s, under different bit rates (1, 3, 5, 7, 9, and 11 Mbps), and two different packet sizes (512 and 1024 Bytes).

Regarding the real aerial network, measurements were acquired from the architecture shown in Fig. 1. The two endpoints were progressively separated from each other along the X axis. For the different measurement points (represented as marks in the X axis of Fig. 1), we obtained the maximum available bandwidth using the iPerf3 network tool and the signal level using the Rohde &

Schwarz's FSH3 spectrum analyzer [24]. The UAV was an Idea-Fly IFLY-4S drone, able to carry up to 700 g of payload, positioned at two different heights (10 and 20 m, respectively). It was equipped with the Galileo board, its corresponding battery, and the couple of antennas described above (see Fig. 2).

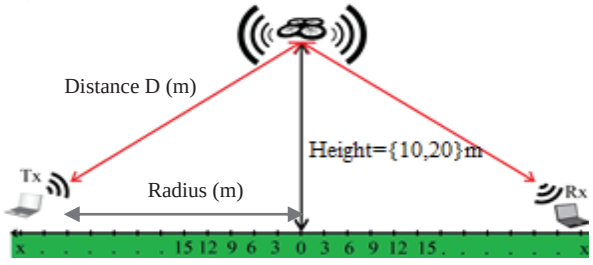
Third (and last), we also conducted a study of the Galileo board's energy consumption. To this end, the instantaneous board's current consumption was measured for the different operating WiFi modes under consideration. The testbed employed for obtaining these measurements is illustrated in Fig. 3. Note that the demanded current is checked at a resistance in series with the power supply feeding the board. By using a data-acquisition card, the board consumption was monitored by means of the Matlab software.

4. RESULTS

Prior to the effective deployment of any network, it is necessary to conduct a performance evaluation of the proposed system. In the following, we specifically explored three key metrics for this type of system: (i) coverage area, (ii) transmission rate, and (iii) energy efficiency.

4.1 Coverage range

In this section, we present both the theoretical study of the coverage area that a UAV can reach with an Intel Galileo onboard and the experimental results obtained. For the theoretical study, it was assumed that the deployed scenario was an open area, without obstacles or any other source of interference, hence existing permanent line-of-sight (LOS) between the intermediate node (UAV) and the communication endpoints. Under these conditions, the analysis of the signal propagation loss was done by means



(a) Diagram

(b) Real snapshot

Figure 1. Experimental test-bench.

Figure 2. Drone used in the experimental aerial WiFi network.

Figure 3. Experimental setup for current consumption measurements.

of two well-known radio propagation models, namely, the Free Space (or Friis) model [20] and the WINNER D1 model [21]. The Free Space model is usually employed to predict the signal strength when there is LOS between transmitter and receiver and there are no nearby objects that might obstruct communication. This model is defined by expression (1). In order to use rigorously this model, it must be applied just under far field conditions. This term is met in the considered environment due to the small size of the used antennas and the long distances covered by the proposed system (please refer to [20] for a more thorough analysis of this concept).

In turn, the WINNER channel model is a stochastic model that predicts channel-introduced losses for different types of environments. The WINNER D1 variation was the specific model chosen for this work due to the aforementioned characteristics of the scenario under study, namely, an open outdoor environment with no obstacles and with the "base station" located at an elevated position; thus, with existence of LOS between the UAV and the other nodes composing the network. The WINNER D1 model is defined by (2), where A = 21.5, B = 44.2, and C = 20.

$$L_{PROP} = 10\log_{10}(4\pi d/\lambda)^2 = $$
$$= 20\log_{10}\big(d(m)\big) + 20\log_{10}\big(f(Hz)\big) + 20\log_{10}\left(\frac{4\pi}{c}\right)$$
$$= 20\log_{10}\big(d(m)\big) + 20\log_{10}\big(f(MHz)\big) - 27.55 \quad (1)$$

$$L_{PROP} = A \cdot \log_{10}\big(d(m)\big) + B + C \cdot \log_{10}\left(\frac{f_c[GHz]}{5.0}\right) \quad (2)$$

Therefore, taking into account a transmission power for both the uplink and downlink of 20 dBm (100 mW) and the characteristic gain and sensitivity values of each wireless card as specified in the datasheets, we calculated the maximum coverage distances for the different versions and operating bitrates of the IEEE 802.11 standard. These results are written in Table 1 (downlink) and Table 2 (uplink). Please note that the scenario for these

Table 1. Comparison of radio coverage. Downlink.

Standard		Rate (Mbps)	Prop. Mod	Radius (m)	Max D (m)
2.4 GHz	802.11b	11	Friis	6719	6718,9
			Winner D1	4613	4612,9
	802.11g	54	Friis	950	949,95
			Winner D1	747	746,93
	802.11n	144	Friis	754	753,93
			Winner D1	603	602,92
		300	Friis	534	533,91
			Winner D1	438	437,89
5GHz	802.11a	6	Friis	2562	2561,9
			Winner D1	1882	1881,9
		54	Friis	362	361,86
			Winner D1	305	304,84
	802.11ac	78	Friis	182	181,73
			Winner D1	161	160,69
	802.11n	144	Friis	145	144,65
			Winner D1	130	129,61
		300	Friis	115	114,56
			Winner D1	105	104,52
	802.11ac	200	Friis	102	101,51
			Winner D1	94	93,47
		866	Friis	65	64,23
			Winner D1	62	61,19

calculations is shown in Fig. 1 and we assumed that the drone is located at a height of 10 m. After this theoretical study, the signal strength was also evaluated in the real experiment as described before. Fig. 4 compares the level of received signal in both modes of operation (infrastructure, Fig. 4(a) and Fig. 4(c), and ad-hoc, Fig. 4(b) and Fig. 4(d)) at two drone-flying heights, namely, 10 m and 20 m. Observe the notable difference between the expected theoretical values and the actual attained figures. This can be explained by many elements influencing the real experiment: drone instability, interference caused by the drone's chassis, atmospheric conditions, etc. Comparing both modes of operation, we detect that the level of received signal in the infrastructure mode (Fig. 4(a) and Fig. 4(c)) was significantly higher than in the ad-hoc mode (Fig. 4(b) and Fig. 4(d)). This increase in the signal level for the infrastructure mode was reflected also in a better throughput, as it will be discussed in the next subsection. Please, note that the reduced number of measurement points for the ad-hoc mode (Fig. 4(b) and Fig. 4(d)) can be explained by the fact that only the measurements obtained when the drone acted as an intermediate point between the two communication endpoints were included, ignoring the results gathered when the transmitter and the receiver laptops were directly connected. In other words, while the distance between the communication endpoints was smaller than 60 m (drone flying at a height of 10 m) or 80 m (drone flying at a height of 20 m), the UAV did not act as an intermediate node in the ad-hoc mode.

Table 2. Comparison of radio coverage. Uplink.

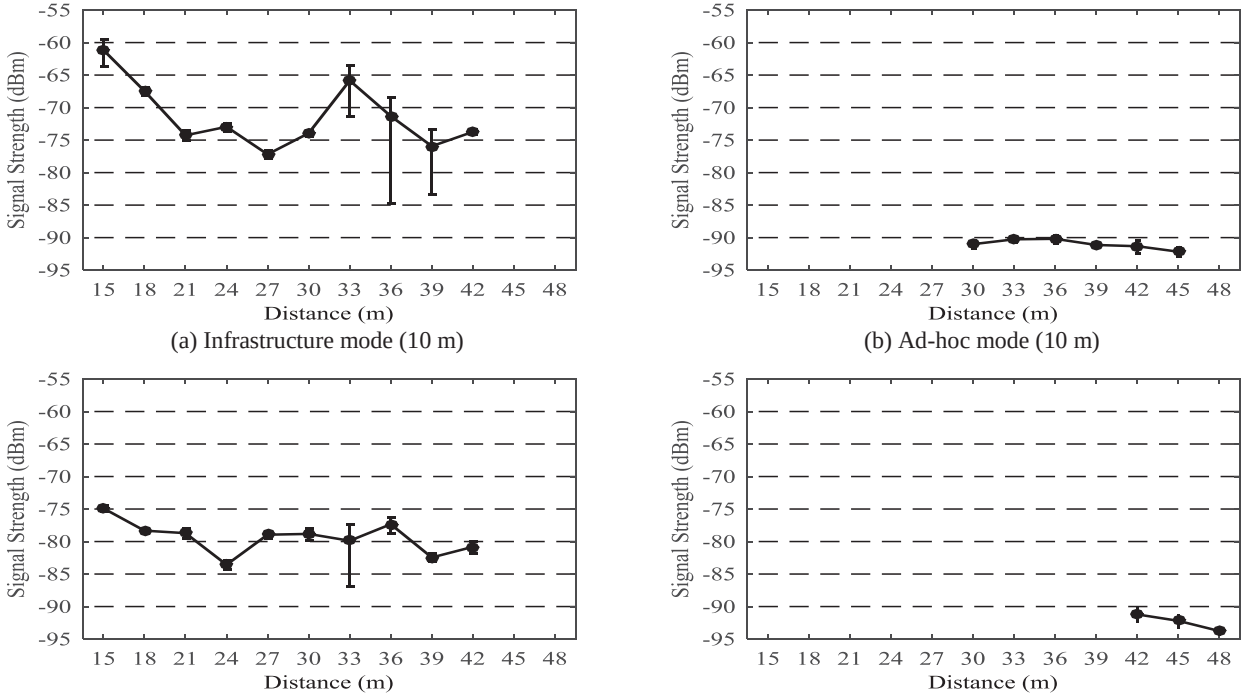
Standard		Rate (Mbps)	Prop. Mod	Radius (m)	Max D (m)
2.4 GHz	802.11b	11	Friis	2440	2439,9
			Winner D1	2227	2226,9
	802.11g	54	Friis	487	486,90
			Winner D1	402	401,88
	802.11n	72.2	Friis	274	273,82
			Winner D1	235	234,79
		300	Friis	194	193,74
			Winner D1	171	170,71
5GHz	802.11a	6	Friis	1475	1474,9
			Winner D1	1126	1125,9
		54	Friis	234	233,79
			Winner D1	203	202,75

4.2 Transmission Rates

In order to evaluate the performance of the infrastructure and the ad-hoc modes in terms of throughput, we used iPerf3 for several CBR transmissions lasting 30 s each, between both communication endpoints. We accomplished two experiments in which the Galileo board acted as an intermediate hop between the two communication endpoints (laptops). In the first experiment, which was carried out under lab controlled conditions, four different configurations were explored, namely, infrastructure and ad-hoc with two different packet sizes (512 B and 1024 B). The obtained results are included in Table 3. Observe that the performance in terms of throughput of the infrastructure mode was notably better than that attained by the ad-hoc mode. Furthermore, it was also advantageous to use a larger packet size (1024 B vs. 512 B) in both modes of operation. This behavior was confirmed by a second experiment, which was performed in a real deployment. The results obtained in the experimental aerial WiFi network are depicted in Fig. 5. We observe that the transmission rates achieved for the infrastructure mode were always higher than those obtained using the ad-hoc mode. For these measurements, the experiment was limited to a packet size of 512 B. Please also note that the attained values are far from those measured under controlled conditions, so the drone motion, a possible radio interference caused by the drone's chassis, or the environmental changing conditions seem to have a remarkable impact on the system performance. In the same way, these results are much lower than those obtained in [16], likely because in our experiments the UAV is an intermediate node (either an AP or an intermediate ad-hoc hop), whereas in their work the UAV was a communication endpoint, and because they set the 802.11n version and we allowed the wireless card to auto-select the best version of the 802.11 standard (b/g/n) depending on the link conditions.

Finally, we also investigated the maximum theoretical data rate that the wireless cards could reach, depending on the link conditions during the aerial tests. This maximum

data rate is defined by the tabulated values set by the IEEE 802.11 standard. Accordingly, the cards are self-configured to work in a different version of the standard as a function of the link quality, thus employing different modulations, which in turn determines the maximum link transmission speed. In our real experiment, the cards were always alternating the 802.11g and 802.11n versions. The attained results are specified in Table 4. We can observe the expected decrease on the link quality with the distance, which forces the wireless cards to employ more



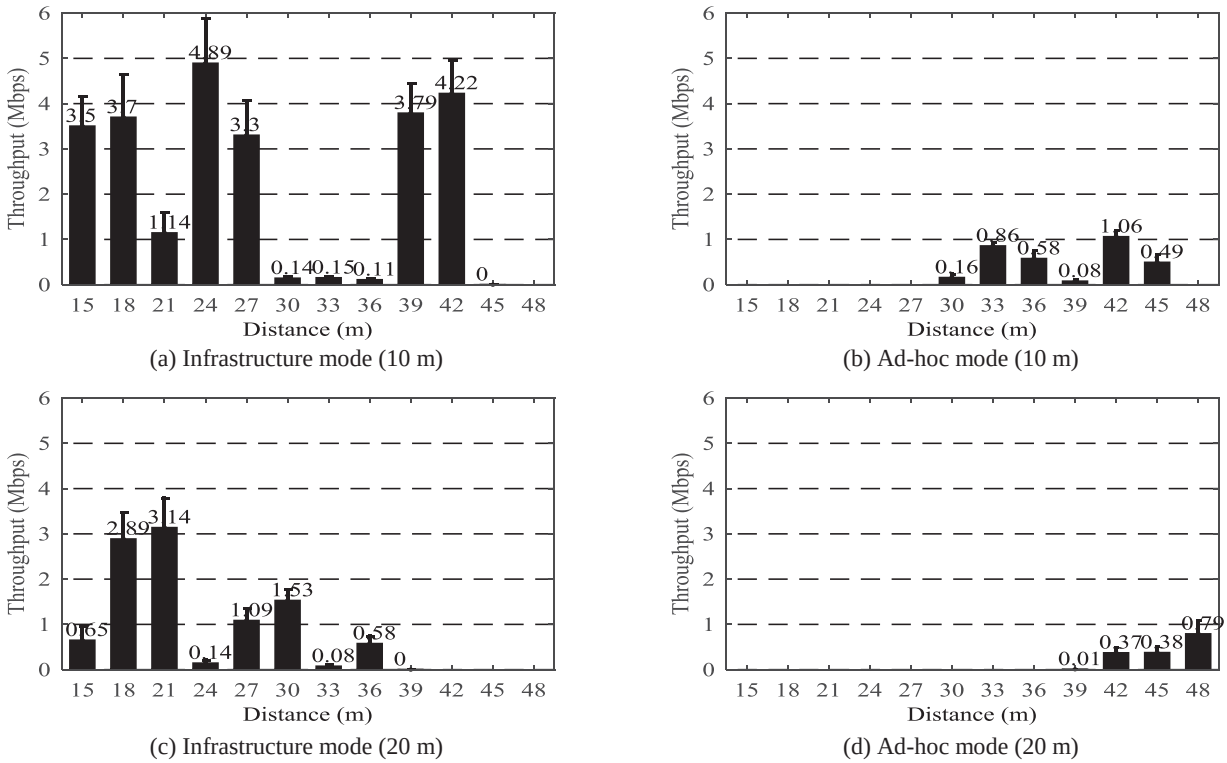
(a) Infrastructure mode (10 m)

(b) Ad-hoc mode (10 m)

(c) Infrastructure mode (20 m)

(d) Ad-hoc mode (20 m)

Figure 4. Average signal strength and confidence intervals at the receiver side.



(a) Infrastructure mode (10 m)

(b) Ad-hoc mode (10 m)

(c) Infrastructure mode (20 m)

(d) Ad-hoc mode (20 m)

Figure 5. Average throughput and confidence intervals achieved between TX and RX nodes.

Table 3. Comparison of Maximum Transmission Rates.

Operation Mode	Packet Size (Bytes)	Max. Throughput (Mbps)
AP	512	10.5
	1024	11
Ad-hoc	512	4.5
	1024	7.5

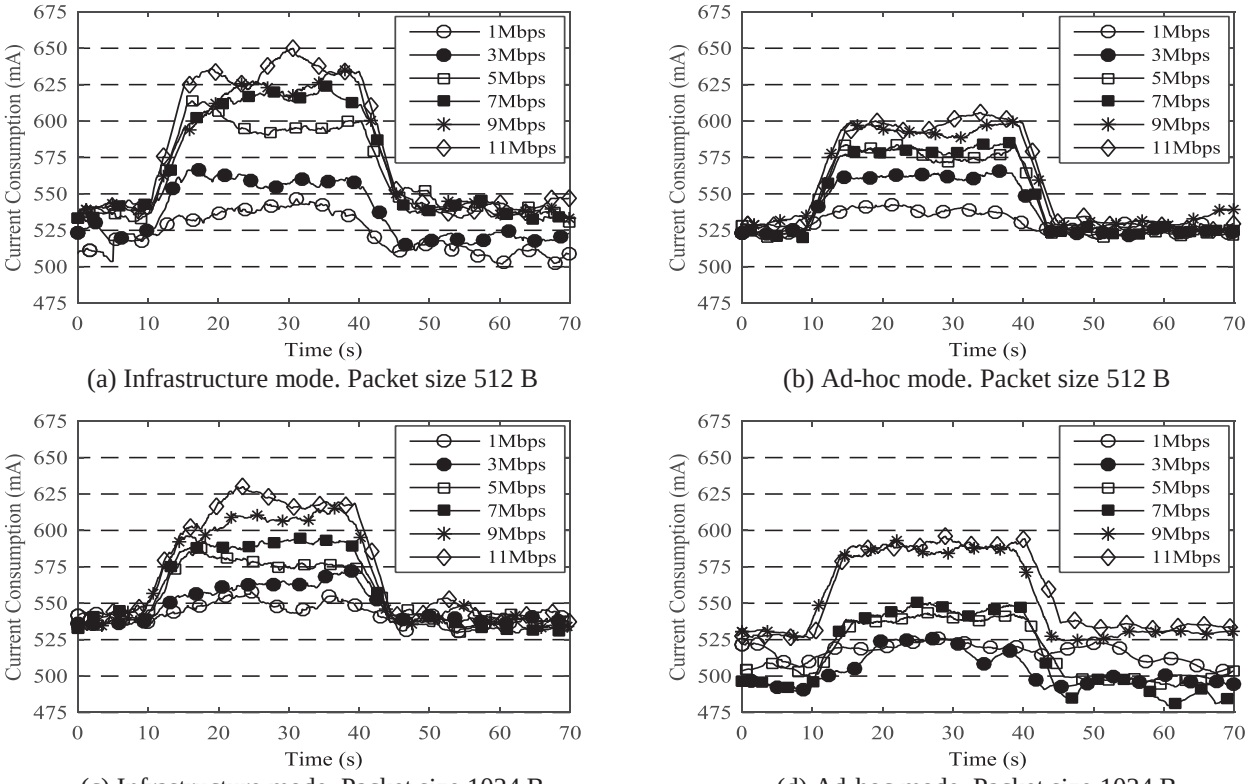
Table 4. 802.11 Operational Bandwidth.

	Drone at 10 m high		Drone at 20 m high	
Distance(m)	Infrastructure	Ad-hoc	Infrastructure	Ad-hoc
15	54 Mbps	--	45 Mbps	--
18	60 Mbps	--	135 Mbps	--
21	54 Mbps	--	90 Mbps	--
24	90 Mbps	--	30 Mbps	--
27	81 Mbps	--	54 Mbps	--
30	5 Mbps	6 Mbps	30 Mbps	--
33	5 Mbps	6 Mbps	40 Mbps	--
36	5 Mbps	11Mbps	30 Mbps	--
39	81 Mbps	5.5 Mbps	30 Mbps	5.5 Mbps
42	81 Mbps	11 Mbps	81 Mbps	5.5 Mbps
45	5 Mbps	1 Mbps	5 Mbps	1 Mbps

conservative modulations and hence reducing their maximum data rate. Comparing both modes of operation, please note that, in general, the infrastructure mode allows the use of higher transmission rates than the ad-hoc mode. This behavior matches the results discussed above regarding the level of received signal and the real transmission rate, in which both of them were higher for the infrastructure mode.

4.3 Energy Efficiency

Another important characteristic of the onboard system is its current consumption. By determining the Intel Galileo's current demand under its different operational modes and supporting several traffic loads, its battery can be accurately dimensioned. This is a crucial factor due to the strict load's weight restrictions usually imposed by drones. Consequently, we carried out a study of the instantaneous current consumption of the Intel Galileo board for both the infrastructure and the ad-hoc modes. These measurements were taken at both, idle state and transmitting traffic at a constant bit rate (CBR) using the same controlled lab test-bench as in the previous subsection (i.e., the two laptops as communication endpoints and the on-the-ground board as intermediate device). Each CBR transmission lasted 30 s and we tested several bit rates (1, 3, 5, 7, 9, and 11 Mbps).



(a) Infrastructure mode. Packet size 512 B

(b) Ad-hoc mode. Packet size 512 B

(c) Infrastructure mode. Packet size 1024 B

(d) Ad-hoc mode. Packet size 1024 B

Figure 6. Comparison of WiFi modes in terms of current consumption using 2 different packet sizes (512 B and 1024 B).

Fig. 6 depicts the current-consumption evolution for the different conditions under consideration. The x-axis represents the test duration, starting in an idle state, then transmitting traffic (30 s), and coming back to the idle state. Observe the clear growth in the Intel Galileo board's current demand when the traffic load increases. This behavior was maintained for all the evaluated conditions. In addition, the ad-hoc mode consumes less current than the infrastructure mode (please, compare Fig. 6(a) and Fig. 6(c) with Fig. 6(b) and Fig. 6(d)). Focusing on the ad-hoc mode, observe that it is advantageous to employ a packet size of 1024 B because it allows a notable reduction of the board's current demand. This can be explained as follows. The larger the packet size (1024 B), the lower the number of packet operations (reception and forwarding) are needed to reach a given bandwidth. Finally, we observe the existing tradeoff between both operational modes. Whereas the infrastructure mode offered a superior performance in terms of data rate and coverage, the ad-hoc mode showed better energy efficiency that is also a critical factor for this type of systems.

5. CONCLUSION

This paper explored one of the most promising applications of UAV in the field of communications: The use of aerial networks to increase network connectivity. To this end, we used the Intel Galileo development board, appropriately configured and equipped to work as a WiFi node (either as an AP in the infrastructure mode or as an intermediate hop in the ad-hoc mode) onboard a UAV. We first carried out a theoretical coverage study of the flying WiFi node using the Free Space and the WINNER D1 propagation-loss models, being the latter the most restrictive. Afterwards, we compared the two WiFi operational modes in experimental scenarios in terms of coverage, throughput, and energy efficiency. Results revealed a better performance of the infrastructure mode regarding received signal strength and bandwidth, but a worse behavior in terms of current consumption compared with the ad-hoc mode. As future work, we plan to carry out an extensive performance evaluation of an aerial WiFi network consisting of several drones. To sum up, drones add a new dimension that assists to extend telecommunications beyond their current conventional limits, but the new potential benefits and risks are still an open research area that should be further investigated.

REFERENCES

[1] M. Kane, "Refugees in Greece need internet so badly that they'll stop a riot to let the wifi guys work," *Quartz, <http://qz.com/711529>*, 2016.

[2] M. Rumpler *et al.*, "Evaluations on multi-scale camera networks for precise and geo-accurate reconstructions from aerial and terrestrial images with user guidance," *Comput. Vis. Image Underst.*, pp. 1–16, 2016.

[3] S. Hayat *et al.*,"Survey on unmanned aerial vehicle networks for civil applications: a communications viewpoint," *IEEE Commun. Surv. Tutorials*. In press.

[4] "Raspberry Pi," <http://www.raspberrypi.org>, 2016.

[5] Intel, "Intel Galileo." [Online]. Available: http://download.intel.com/support/galileo/sb/galileoprodbrief_329680_003.pdf. [Accessed: 22-Apr-2016].

[6] "Eurocontrol.", <http://www.eurocontrol.int> .

[7] "Ley 18/2014, de 15 de octubre, de aprobación de medidas urgentes para el crecimiento, la competitividad y la eficiencia," *Spanish Off. Bull.*, vol. 2014, no. 252, pp. 83921–84082, 2014.

[8] L. Song and T. Huang, "A summary of key technologies of ad hoc networks with UAV node," in *International Conference on Intelligent Computing and Integrated Systems*, 2010, pp. 944–949.

[9] L. Gupta, R. Jain, and G. Vaszkun, "Survey of important issues in UAV communication networks," *IEEE Commun. Surv. Tutorials*, vol. 18, no. 2, pp. 1123-1152, 2016.

[10] Y. Saleem *et al.*, "Intregation of Cognitive Radio Technology with unmanned aerial vehicles: issues, opportunities, and future research challenges," *J. Netw. Comput. Appl.*, vol. 50, pp. 15–31, 2015.

[11] J.D.M.M. Biomo, *et al.*, "Routing in Unmanned Aerial Ad Hoc Networks: A Recovery Strategy for Greedy Geographic Forwarding Failure," *Proc. IEEE WCNC Mob. Wirel. Networks*, pp. 2236–2241, 2014.

[12] N. Uchida *et al.*, "Proposal of Seeking Wireless Station by Flight Drones base don Delay Tolerant Networks," *Proc. 9th Int. Confenrece Broadband Wirel. Comput. Commun. Appl.*, pp. 401–405, 2014.

[13] M. Bekhti *et al.*, "Path Planning of Unmanned Aerial Vehicles with Terrestrial Wireless Network Planning," *Proc. Wirel. Days*, pp. 1–6, 2016.

[14] C. Ting-Yun and E. Al., "Civil UAV Path Planning Algorithm for Considering Connection with Cellular Data Network," *Proc. IEEE 12th Int. Conf. Comput. Inf. Technol.*, pp. 327–331, 2012.

[15] J. Sae *et al.*, "Coverage Aspects of Temporary LAP Network," *Proc. 12th Annu. Conf. Wirel. On-demand Netw. Syst. Serv.*, pp. 100–103, 2016.

[16] S. Hayat *et al.*, "Experimental Analysis of Multipoint-to-Point UAV Communications with IEEE 802.11n and 802.11ac," *Proc. IEEE 26th Int. Symp. Pers. Indoor Mob. Radio Commun.*, pp. 1991–1996, 2015.

[17] M. Arvind *et al.*, "UAV Assisted Heterogeneous Networks for Public Safety Communications," *Proc. IEEE Wirel. Commun. Netw. Conf.*, 329–334, 2015.

[18] S. Koulali *et al.*, "A Green Strategic Activity Scheduling for UAV Networks: A Sub-Modular Game Perspective," *IEEE Commun. Mag.*, pp. 58–64, 2016.

[19] Intel, "BSP." [Online]. Available: https://downloadcenter.intel.com/download/23197/Intel-Quark-BSP. [Accessed: 06-Jul-2016].

[20] H. T. Friis, "A note on a simple transmission formula," in *IRE'46*, 1946, vol. 34 (5), pp. 254–256.

[21] Winner and I. S. Technologies, "IST-4-027756 WINNER II. D1.1.2 V1.2. WINNER II Channel Models," 2008.

[22] A. Neumann *et al.*, "Better Approach To Mobile Ad-hoc Networking (B.A.T.M.A.N.)," *IETF Draft*, 2008.

[23] R. Sanchez-Iborra *et al.*, "Performance evaluation of BATMAN routing protocol for VoIP services: a QoE perspective," *IEEE Trans. Wirel. Commun.*, vol. 13, no. 9, pp. 4947 – 4958, 2014.

[24] "Rohde & Schwarz - FSH3." [Online]. Available: https://www.rohde-schwarz.com/product/fsh3-6-18-productstartpage_63493-7578.html. [Accessed: 20-May-2016].

POSTER SESSION

P.1 Adaptive video streaming over HTTP using stochastic bitrate prediction in 4G wireless networks.

P.2 Assessing Internet performance over mobile networks: from theory to practice.

P.3 Systematic analysis of geo-location and spectrum sensing as access methods to TV white space.

P.4 Task-based process modeling for policy making in smart cities.

P.5 CleanWiFi: the wireless network for air quality monitoring, community Internet access and environmental education in smart cities.

P.6 Cloud based patient prioritization as service in public health care.

ADAPTIVE VIDEO STREAMING OVER HTTP USING STOCHASTIC BITRATE PREDICTION IN 4G WIRELESS NETWORKS

Dhananjay Kumar, S. Aishwarya, and A. Srinivasan, and L. Arun Raj

Department of Information Technology, Anna University, MIT Campus, Chennai
dhananjay@annauniv.edu, aishwarya.gsv@gmail.com, asriniit@gmail.com, arun4u85mit@gmail.com

ABSTRACT

Video streaming over Hypertext Transfer Protocol (HTTP) used in multifarious applications creating a multimedia environment faces a challenge in 4G wireless network due to the fluctuating nature of internet traffic and variable capacity of wireless channel over time. The existing Dynamic Adaptive Streaming over HTTP (DASH), though works well for stored video up to some extent, poses a complication in live transmission thereby depreciating the streaming quality due to high link bit rate fluctuation. In this paper, we have proposed an efficient ARIMA Based Bit Rate Adaptation (ABBA) model in the receiver/client side that estimate the link traffic based on the incoming packet bit rate to predict the subsequent future link capacity in order to notify the sender/server. Based on the response from the receiver the server adapt its outgoing stream as per forecasted link data rate, and hence eliminate the degradation of video due to channel throughput variations. The proposed ABBA algorithm was implemented on IP over 4G wireless network and the streaming quality was evaluated on several full reference metrics of video quality. The test result outperformed an existing buffer based approach and also a fuzzy based adaptation algorithm. For example, the ABBA algorithm exhibited an average increase of 22 % in PSNR and 9% in SSIM than a buffer based method.

Keywords: 4G Wireless, HTTP, ARIMA, Client-server, Adaptation, Video quality evaluation metrics

1. INTRODUCTION

The user data traffic in wireless mobile network has been increasing rapidly across the globe. As per the Cisco Visual Networking Index [1], the 4G wireless network will have the highest stake (40.5 %) of total mobile connections worldwide, and 75% of the global mobile data traffic will be video by 2020. Such remarkable progress is fueled by the video streaming service over internet by YouTube, Netflix, etc. The ever increasing number of smart phones with internet access over 4G wireless network is another reason for the tremendous increase in streaming video traffic.

The ultimate objective of all streaming services is to deliver seamless content to the end user in real-time, though it poses a huge challenge due to large fluctuating bandwidth in the network. To provide user the seamless multimedia service with maximum achievable Quality of Experience (QoE), the media content in particular video need to be adaptive to match the available bit rate in the network. The traditional streaming method based on progressive download fails to cope up with dynamic network traffic [2] thereby degrading the media quality.

The streaming techniques [3] are classified into three major classes: i. traditional (Real-time Transport Protocol (RTP) / RTP Control Protocol (RTCP)), ii. progressive streaming (HTTP/TCP), and iii. adaptive streaming (HTTP/TCP, UDP). The HTTP based Adaptive Streaming (HAS) has exhibited resilience to the internet traffic and hence widely used as DASH [4] in the present systems. The use of DASH in entertainment based utilities, where the stored videos are being streamed to the client, requires segment based information and pre-defined streaming parameters to facilitate ease in deciding the upcoming bandwidth changes. However in live streaming where the video content is created and encoded only when the systems connect in real time over the network, the adaptability of DASH to intimate the sender about the link bandwidth becomes an encumbrance for targeting an improvement in the perceived quality of video by the user [5].

For a case study on existing 4G wireless network, the uplink and downlink data rate on the Airtel 4G LTE-TD Hotspot [6] system was monitored in laboratory environment (Fig. 1). Although these wireless systems are designed to support up to 100 Mbps in high mobility access, the actual capacity at user premises not only fall much below the specified values, but also fluctuate over time. Clearly there is a high incentive in developing a video streaming system which can adapt to this network operating environment.

Figure 1. Bitrate observed during streaming of live videos using Airtel 4G dongle [6]

The variations of incoming bit rates while the video is being streamed can be analogized to a time series to rigorously analyze the past observations to make forecasts about

network conditions. The limitations of traditional linear regressive models need to be rectified where the seasonality of a regularly repeated pattern could be eliminated with increased focus towards accurate predictions. In this paper, a new algorithm based on a non-linear stochastic model called, Auto Regressive Integrated Moving Average (ARIMA) Based Bit Rate Adaptation (ABBA) is proposed. The ABBA algorithm is modeled on a time series consisting of sequence of sampled bit rate over a continuous time interval to analyze a successive statistical measurement that has no natural ordering for the observations. This stochastic model is used for trend analysis of the forthcoming bit rates, which decides the resolution of video to be sent from the server. The strategic decisions are based on the successive observation of sampled bit rate in a regular time interval to understand the nature of series.

To analyze the performance of the proposed ABBA algorithm, two existing approaches: i. Heuristic Decision Rate Adaptation (HDR), and ii. Buffer Switching Rate (BSR) algorithms has been formulated and developed here. The HDR [7] employs the difference in arrival time of packets and buffering time as inputs for predicting the near future using a set of decision rules whereas the BSR [8] monitors the buffer occupancy level dynamically and chooses the mode of operation based on its fill percentile using harmonic mean to effectively identify the nature of network for streaming the videos.

The ABBA, HDR, and BSR algorithm were implemented using VLC Framework in Java (VLCJ) that is completely open source and can easily be plugged into to the existing systems. The system level implementation was in adherence to the ITU-T J.247 recommendation (Table 1) that describes about the 'objective perceptual multimedia video quality measurement'. The developed system were tested for delay variability ITU-Y.1540 [9] and quality of video were observed using PSNR ITU-R J.340 [10] and VQM ITU-J.149 [11] along with other standard popular video quality evaluation metrics.

Table 1. Test factors as per ITU guidelines

Parameter	Standard	Metrics
Frame Rate	ITU –T J.247	5 to 30 fps
Codec		H.264
Resolution		QCIF,CIF,VGA
Temporal errors		<=2 sec
Min bandwidth Required		QCIF 16 kbps to 32kbps CIF:64 kbps to 2Mbps VGA 128 kbps to 4 Mbps
Performance Metrics	ITU-R J.340	PSNR >=25
	ITU-T Y.1540	Delay Variation (Quantile and min delay difference should not be >50 ms)
	ITU-T J. 149	VQM [0-1]

The rest of the paper is organized as follows. The client-server model of the system architecture is presented in Section 2. Section 3 describe about the mathematical model and formulate of the proposed method. Section 4 deals with the algorithm development procedures. The implementation environment along with different metric of performance is described in Section 5. The result and discussion is presented in Section 6. Section 7 concludes with scope for future work.

2. SYSTEM ARCHITECTURE

2.1 The Client Server Model

The proposed system architecture emulates the client server model where server's job gets simplified on the expense of client's increased monitoring and analysis process. The server consists of three sub modules: i. frame capture, ii. streaming the video, and iii. receiving feedback. On the other hand, the client consists of three modules: i. decoder / player, ii. stream flow analysis, and iii. receiver's feedback. The video being streamed is encoded dynamically using ITU-T H.264 [12] video codec. The live (or stored) video is streamed from the server to client through 4G wireless networks and the client scrutinizes the link bandwidth and analyses its trend to make an intelligent decision based on prevailing scenario. This decision is sent as a feedback to server which tries to match channel capacity and the sends the video at corresponding resolution and frames per second.

The client samples the incoming bit rate and monitors the pattern with an aim to analyze and predicts the near future bandwidth. The server is notified with the predicted link capacity which in turn responds with content adaptation process. The bit rates of packets are related to a time series model where a set of data points denotes the bit rates over successive time. The sampled data are arranged in a proper chronological order continuously and the past observations are analyzed to develop a mathematical model (ABBA algorithm) that captures the underlying data to make strategic decisions. This parametric approach considers that the underlying stochastic process has a certain structure which can be described using two parameters: auto-correlations and auto-covariance to forecast the future bit rates using regression.

2.1.1 Sender Sub-modules

The server's main job is to acquire the media content live from the camera or fetch from a memory location in case of a stored video. The following modules represent the workflow of streaming at the server side (Fig. 2).

a) Init: This module initializes the VLCJ player and identifies the media locator required for transmission.

b) Stream: It is used to establish connection with the requesting client using sockets while creating instance of player to stream at required quality.

c) Adapt: This module receives message from the client and uses this feedback to adapt to the network prevailing conditions by interpreting the data from client to make a strategic decision.

2.1.2 Receiver Sub-modules

The following modules illustrate how the client analyses (Fig. 3) the link instability based on ABBA Algorithm.

a) Playback: The client requests the sender to start streaming and initiates connection with the sender in an appropriate port using HTTP.

b) Analyze: The client analyzes the incoming bit rate of packets to ascertain how the link will support stream in near future using ABBA algorithm.

c) Send Feedback: The client creates a message based on defined format of feedback and makes an intelligent decision to alert the sender.

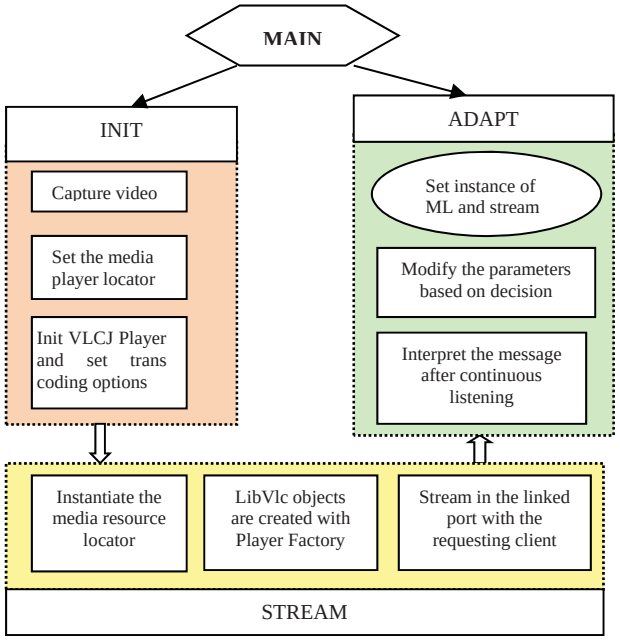


Figure 2. Server side modules

3. SYSTEM MODEL

The client side of the proposed system has higher complexity than the server, and it applies stream analysis algorithm to handle the non-linearity in the incoming traffic as they may vary rapidly over time that is too complicated to fit into any specific predefined classes. A heuristic based stochastic algorithm is formulated to overcome the existing problems and ensure delivery of higher quality videos in the prevailing circumstances. Based on the predicted link behavior the client identifies the trend of the series that is labeled as i. advancing, ii. degrading, iii. oscillating, and iv. stable. The traffic load in the network could momentary increase/decrease or prolong to increase/decrease. The projected incoming flow rate is then sent to the server as a feedback for it to adapt effectively and modify its parameters instantaneously.

3.1 Stochastic Prediction Model

There is need to explore a suitable statistical model which captures the dynamics of incoming bit rate and maps to a time series model that can be used to predict the future trend considering the current network conditions.

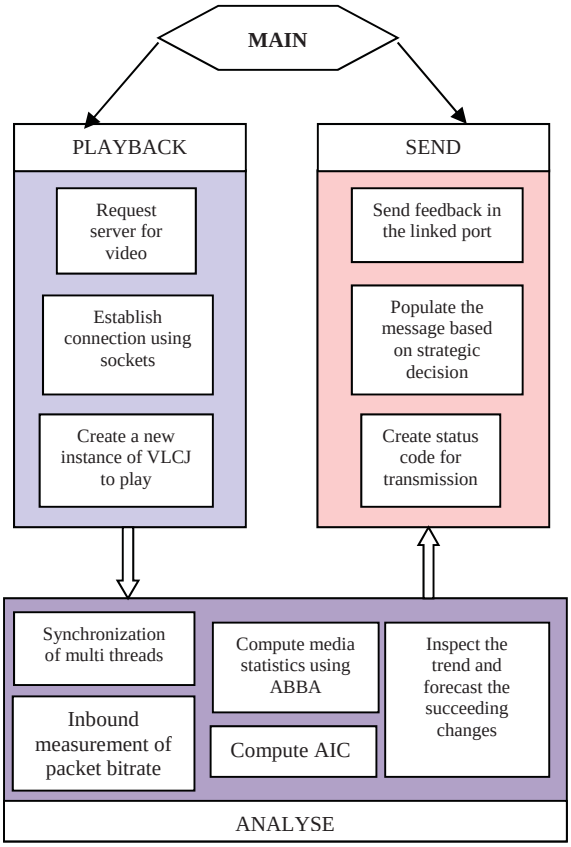


Figure 3. Client side modules

3.1.1 Auto-regressive (AR) Component

The Auto-regressive (AR) part is used to establish the covariance between the bit rates fluctuating over time [13] that can be used to foresee how the variations would take place in the future.

$$Auto\_Reg = 1 - \sum_{i=1}^{p} \phi_i L_i \quad (1)$$

where ϕ_i represent covariance and L_i lag operator for i^{th} packet, and p denotes the number of bit rate samples taken over time.

The covariance ϕ_i signifies a statistical relationship between bit rate and time that is used for trend analysis and foresees the upcoming bit rates, which is expressed as:

$$\phi = E\left[X_t, X_s\right] - \mu_t \times \mu_s \quad (2)$$

where μ_t and μ_s represents the mean associated with the random variables X_t and X_s.

3.1.2 Auto Regressive Integrated Moving Average (ARIMA) Model

Considering X_{tp} to be predicted bit rate where '*tp*' denotes the index number, *ARMA(p,q)* model with the integration of correlation factors is defined as:

$$\left(1-\sum_{i=1}^{p'}\alpha_i L^i\right)X_{tp}=\left(1+\sum_{i=1}^{q}\theta_i L^i\right)\varepsilon_t \quad (3)$$

where L^i is the lag operator, α_i is the autoregressive part (Auto_Reg) and θ_i is the moving average part (*Mov_Avg*) of the model linking the correlation between the successive time windows under evaluation for the i^{th} packet.

Now, the *Auto_Reg* polynomial would have a unitary root of multiplicity *d* when applied to a non-linear stochastic process where the first order differencing of a characteristic equation is non-stationary as

$$\left(1-\sum_{i=1}^{p'}\alpha_i L^i\right)=\left(1-\sum_{i=1}^{p'-d}\phi_i L^i\right)(1-L)^d \quad (4)$$

The stationarity here refers to the time series bit rate based model whose variance and auto correlation structures do not vary over time.

Integrating polynomial (4) with the Moving Average component (Mov_Avg) [14] to determine the future sequence by factorization of $p = p' - d$ would give rise to ARIMA model as

$$\left(1-\sum_{i=1}^{p}\phi_i L^i\right)(1-L)^d X_{tp}=\left(1+\sum_{i=1}^{q}\theta_i L^i\right) \quad (5)$$

The ARIMA in (5) can be generalized by adding a stochastic drift constant δ that denotes the change of average of bit rates in a continuously changing process that is modeled as a regression drift constant given by

$$\left(1-\sum_{i=1}^{p}\phi_i L^i\right)(1-L)^d X_{tp}=\delta+\left(1+\sum_{i=1}^{q}\theta_i L^i\right) \quad (6)$$

Now, the near future bit rates X_{tp} and trend of bandwidth fluctuations can be predicted to notify the server so that it can adapt accordingly as in (7) that integrate the correlation and variance into a form of regression.

$$X_{tp}=\delta'+\left(1+\sum_{i=1}^{q}\theta_i L^i\right)/\left(1-\sum_{i=1}^{p}\phi_i L^i\right)(1-L)^d \quad (7)$$

$$\text{where } \delta'=\delta/\left(1-\sum_{i=1}^{p}\phi_i L^i\right)(1-L)^d \quad (8)$$

4. ALGORITHM DEVELOPMENT

There is a need for a standard benchmark criterion to ascertain if the model predicts relatively accurate value and in this context the Akaike Information Criteria (AIC) [15-16], is embedded in the proposed ABBA algorithm which helps in fine-tuning the quality of prediction model. The AIC act as a quality gauge in mathematical models using statistical parameters that computes the quality of a single model which is used as one of decision parameter.

To compare the performance of the ABBA algorithm, two existing approaches: (i) Buffer Switching Rate (BSR) adaptation, and (ii) Heuristic Decision based Rate Adaptation (HDR) method is formulated and presented here.

4.1 ARIMA Bitrate Based Adaptation (ABBA) Algorithm

The sliding window size used for analysis can be incremented using a constant factor *inc_fac* that could be predefined to overcome the stationarity issues based on the sample data points taken into consideration.

ABBA Algorithm:

Input: Bit Rates of packets for a Period of Time

Output: Forecasted Bit Rate for the next sequence of packets

1. Start streaming of video content
2. Sample the data rate by choosing a sequence of packets
3. Initialize *N*, *inc_factor*, *p*, *q*, *d*, *k* for *AIC*.
4. Compute μ_r, μ_s, μ_t. // Calculate mean
5. While (i<*N*) // Take *N* samples
6. If ($\mu_r == \mu_s$) //Test for stationarity
7. If (C_{xx}(r,s) == C_{xx}(r-1, s-1)) // C_{xx} is the covariance
8. *N* = *N* + *inc_factor* // Increment window size
9. for i = 1 to *m*
10. Evaluate the Lag Operator

$$a(L)X_t = X_t - \sum_{i=1}^{d} a_i X_{t-i}$$

11. Evaluate the variance σ_s and σ_r

$$\sigma = \frac{\sum_{i=1}^{n}(X_i - X_{avg})^2}{n-1}$$

12. Compute θ as stated in (1) // Moving Average
13. for *j*=1 to *q* // *q* is no of sample points chosen
14. Evaluate *Auto_Corr* = θ * *L* // Correlation
14. if (*Auto_Corr*!=0)
15. Compute Φ as shown in (2) // Auto Regression
16. for *k*=1 to *p*
17. Evaluate *Auto_Var* = Φ * *L* // Covariance
18. if (*Auto_Var*>0)
19. Compute the forecasted X_{tp} using (7)
20. Compute variance as shown

$$\sigma_{(AIC)} = \frac{\sum_{i=1}^{n}(X_i - X_{tp})^2}{n-1}$$

21. Calculate the Akaike Information Criteria

$$AIC[i] = \log\sigma + \frac{2K}{N}$$

22. else
23. *N*= *N* + *inc_factor* //Increment Window size
24. end for
25. Find the *p,q,d* that corresponds to the maximum $\{AIC\ [i]\}_{max}$ in the array
26. Designate the optimal values for *p, q, d* and repeat from Step 9.
27. Repeat from step 5 until streaming occurs.

4.2 Buffer Switching Rate (BSR) Adaptation Algorithm

Buffer based switching algorithm [8] considers the current buffer occupancy level and number of segments of video to provide the best possible quality of streamed video to the client. The harmonic mean is computed to measure the throughput for the entire video session and to avoid instantaneous variation of throughputs. Weights ($s_1, s_2, s_3..$) are assigned proportional to their segment sizes ($t_1, t_2, t3..$) to compute the weighted harmonic mean, which is expressed as,

$$H_n = \frac{\sum_{i=1}^{n} s_i}{\sum_{i=1}^{n} \frac{s_i}{t_i}} \quad (9)$$

The different types of resolutions depicting various qualities of video $r_1, r_2, r_3 \ldots r_{cur, \ldots}$ where r_{cur} denotes the current resolution, are the input to the playout buffer. The buffer is associated with three thresholds E_1, E_2 and E_3 as shown in Fig. 4.

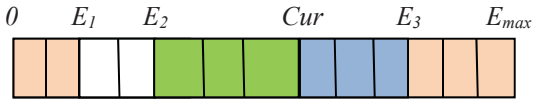


Figure 4. Buffer occupancy level

BSR Algorithm

Input: Segment information, Current buffer occupancy level *Cur*, Weighted Harmonic mean download rate H_n

Output: The next predicted bit rate l_{n+1}

1. Read the current buffer occupancy level *Cur* for n^{th} segment while streaming
2. If ($Cur \leq E_1$) //Speedy start phase
3. l_{n+1} = r1 //Choose lowest quality
4. else
5. if $\left(\frac{s_{n+1}^{cur}}{H_n}\right) > cur - E_1$
6. $l_{n+1} = \max\left(r_{cur}>0 \ , \ \frac{S_{n+1}}{H_n} \leq cur - E_1\right)$
7. Set $d = 0$ //immediate download
8. else if ($Cur \leq E_2$) // progressive increase phase
9. if $\frac{S_{n+1}}{H_n} < cur - E_2$
10. l_{n+1} = r_{cur} +1 //Increment resolution in steps
11. else
12. l_{n+1} = r_{cur} // Maintain current level
13. Set $d=0$
14. else if ($Cur \leq E_3$) // Rapid Shift Phase
15. $l_{n+1} = \max\left(r_{cur}>E_2 \ , \ \frac{S_{n+1}}{H_n} \leq cur - E_2\right)$
16. Set $d=cur-E_2$
17. elseif ($Cur>E_3$) // Delayed download
18. $l_{n+1} = \max\left(r_{cur}>E_3 \ , \ \frac{S_{+1}}{H_n} \leq cur - E_2\right)$
19. Set $d = cur - E_2$
20. else
21. l_{n+1} = r_{cur} // Maintain current level
22. Set $d=0$
23. Repeat the above steps until streaming terminates

4.3 Heuristic Decision based Rate Adaptation (HDR) Algorithm

Sample data set contains input variables being mapped to the set of membership functions using a fuzzy heuristic logic. The fuzzification technique involves the conversion of crisp value to continuous fuzzy value. The HDR algorithm is based on the decision of the heuristic control system which in turn is based on the if-then rules [7].

The difference in arrival time of the packets is classified as short (S), near (N) and extended (E) that describes the remoteness of the current buffering time from target buffering time *T*. This heuristic based rate adaptive decisions is used to avoid buffer underflow and also retain the difference between current and previous resolution quality to zero in order to avoid frequent variations .The buffering time is the difference between the time at which the packet is received and time at which it is played. The difference between the subsequent buffering times is classified as degrading (F), stable (S) and increasing (I). The rules corresponding to Decrease (D), Short Decrease (SD), Steady (S), Sort Advance (SA), and Advance (A) are as follows:

(r_1): if (small) and (decreasing) then D
(r_2): if (near) and (decreasing) then SD
(r_3): if (extended) and (decreasing) then S
(r_4): if (small) and (stable) then SD
(r_5): if (near) and (stable) then S
(r_6): if (extended) and (stable) then SA
(r_7): if (small) and (increasing) then S
(r_8): if (near) and (increasing) then SA
(r_9): if (extended) and (increasing) then A

These are five decision conditions which are sent to the server to make appropriate adjustment in parameters of the video being streamed. Finally, the centroid method is used for de-fuzzification i.e., to map the arrival and buffering times in terms of a single parameter *h* given by [7]

$$h = \frac{N_2*D+N_1*SD+Z*S+P_1*SA+P_2*A}{SD+D+S+A+SA} \quad (10)$$

where N_2 , N_1, Z, P_1, P_2 are the membership values defined as in Table 2 [7] and *A*, *SA*, *S*, *SD*, and *D* are formulated in terms of rules (r_1 - r_9) given by

$$A = \sqrt{r_9^2} \quad (11)$$
$$SA = \sqrt{r_6^2 + r_8^2} \quad (12)$$
$$S = \sqrt{r_3^2 + r_5^2 + r_7^2} \quad (13)$$
$$SD = \sqrt{r_2^2 + r_4^2} \quad (14)$$
$$D = \sqrt{r_1^2} \quad (15)$$

Table 2. List of parameters in HDR [7]

Parameters	Value	Definition
T	35 sec	Target Buffering Time
d	60 sec	Time Period estimating connection throughput
N_2, N_1, Z, P_1, P_2	0.25, 0.5, 1, 1.5, 2	Factors of membership functions

5. IMPLEMENTATION ENVIRONMENT

Different video resolutions namely SQCIF, QCIF, CIF, QVGA, and VGA were used for the transcoding of the input video at the server side for every streaming instance. The frames per second (fps) designated for the streaming are 10, 15, 25, 30 and 35 respectively with the default value set to 25 fps. Java programming environment was used as it is platform independent and supports VLCJ (VLC for Java). The VLCJ is an open source framework that is used for video streaming that enables the media content to get embedded in a Java Swing. Since this platform is completely open source there are many customizable options available that are deployed for obtaining the media statistics. Synchronization of multithreads is carried out for communication between the client and server for transfer of data. Transcoding of input stream i.e., the process of converting media object from one configuration to another, allows to switch between various resolutions at the server side. Dshow [17] is the API that is used to capture the video and process it for streaming in the appropriate format.

The wireless network for experimental set up was established through Airtel 4G LTE-TD Hotspot [6] that demonstrated an average of 3.7 Mbps in downlink during real time testing although it is intended to support more than 8 Mbps as specified by the service provider. For case study the end-to-end link bandwidth was estimated with the help of an online tool Speedof.Me [18]. The client and server were implemented in Lenovo idea pad laptop which has Intel Core i5 processor having 4 GB RAM and Windows 7 Professional 64 bit operating system. The streaming of video was implemented on top of the HTTP with UDP as its underlying transport protocol.

5.1 Performance Evaluation Parameters

Since in our experimental set up original video sequence was readily available, the Full Reference (FR) metrics were employed to evaluate the system performance. Further, FR metrics provide the most accurate result as it is computed with direct reference to the original sequence. The commonly used FR parameters are: Peak Signal to Noise Ratio (PSNR), Structural Similarity (SSIM) index, Video Quality metric (VQM), Aligned PSNR (A-PSNR) etc. Although the conventional PSNR is relatively simple to compute it exhibits imprecise measurement when used to measure the quality of streamed video over wireless network. Since the packet loss in the wireless mobile network cannot be neglected, more complex metrics like A-PSNR, VQM, and different types of SSIM are employed for the evaluation of system.

6. RESULTS AND DISCUSSIONS

6.1 Peak Signal to Noise Ratio (PSNR) Measurement

The PSNR metric was evaluated offline on the data generated by a live stream during the experimentation for the three implemented algorithms. The proposed ABBA performs better as it well predicts the future trend and allows the video content as per the network conditions thereby minimizing the mean square error in the decoded frame at the receiver. Table 3 lists the PSNR values corresponding to the ABBA (proposed), BSR, and HDR algorithms. The ABBA algorithm exhibits a higher average PSNR, which is 21% and 12% higher than the buffer based (BSR) algorithm and Heuristic based (HDR) algorithm respectively.

Table 3. Comparison of PSNR values

#	ABBA (dB)	HDR (dB)	BSR (dB)	#	ABBA (dB)	HDR (dB)	BSR (dB)
1	36.01	24.26	24.50	11	35.03	27.86	31.54
2	36.35	32.26	24.25	12	35.03	27.69	31.07
3	35.11	32.25	36.10	13	32.04	35.69	31.21
4	34.11	35.10	29.28	14	36.35	35.97	31.92
5	35.11	29.28	28.32	15	36.35	25.21	25.73
6	33.11	28.32	31.22	16	37.35	35.54	27.69
7	35.125	24.50	31.38	17	34.37	27.69	27.75
8	35.116	34.38	27.54	18	36.53	35.02	29.85
9	35.32	34.22	27.46	19	37.37	35.62	31.98
10	35.066	27.29	31.97	20	31.24	36.19	32.04
Average(dB)					35.10	31.37	29.13

6.2 Structural Similarity (SSIM) Measurement

The SSIM index was computed for the three algorithms (ABBA, BSR, and HDR), and the proposed ABBA algorithm exhibited a higher value with an average of 0.9541 which is 9% higher than the buffer based algorithm 7% better than heuristic based algorithm.

The SSIM index on few decoded consecutive frames at receiver corresponding to ABBA, BSR, and HDR algorithms is plotted in Figure 5. The higher SSIM index for ABBA algorithm is a reward for the perceived video quality as there is need for the perseverance of luminance and contrast factors that are influenced by the distortions.

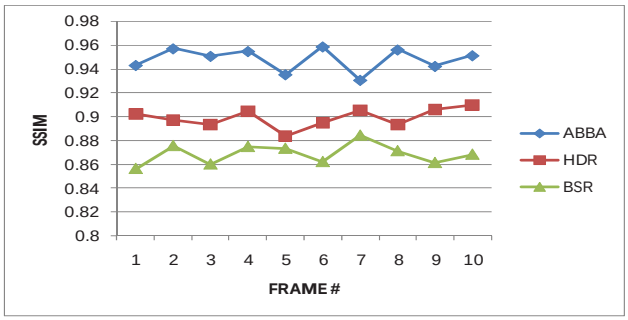


Figure 5. The SSIM index

6.3 Video Quality metric (VQM)

The VQM metric was employed to measure subjective quality assessment of the streamed video at receiver. Since the VQM score is the sum of many weighted parameters and its higher value represents the maximum loss of quality in the video, the lower values observed by the ABBA, HDR, and BSR is desirable. The ABBA based method shows 13% lower than the HDR and 17% lesser than the BSR scheme (Fig. 6) for the VQM measurement.

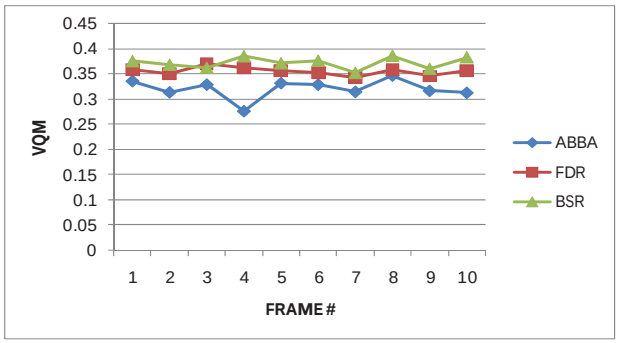


Figure 6. The VQM index

6.4 Aligned-Peak Signal to Noise Ratio (A-PSNR)

If there are multiple consecutive (say 5 and more) loss of frame in the wireless network, the conventional techniques like PSNR fails miserably due to application of fixed window model. This necessitates the adoption of new metric based on the dynamic window size. The Aligned-PSNR (A-PSNR) of ABBA algorithm produces 11.26% higher than the HDR, and 22.26% more than BSR algorithm (Fig. 7). The ABBA algorithm achieves the A-PSNR more than 30 dB which shows its usefulness in delivering high quality videos.

6.5 Multi Scale- Structural Similarity (MS-SSIM) Index

A multi scale SSIM being more flexible than single scale metric provides better result with respect to correlations to human perceptions. On an average ABBA algorithm produces 2% higher quality on MS-SSIM scale than HDR and 0.7% higher than BSR method. Figure 8 shows the variation of MS-SSIM values on different consecutive frames. Although the improvement in quality by ABBA algorithm over other two is marginal, it still leads the pack.

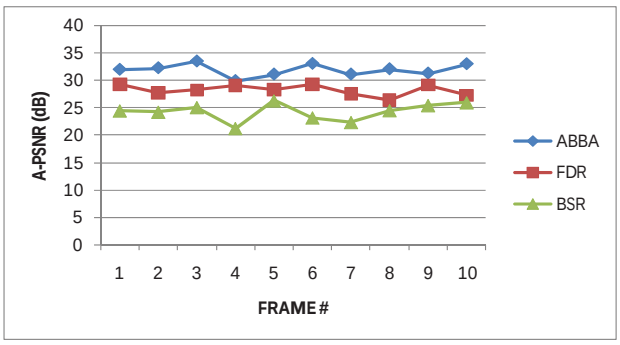


Figure 7. Aligned-PSNR values

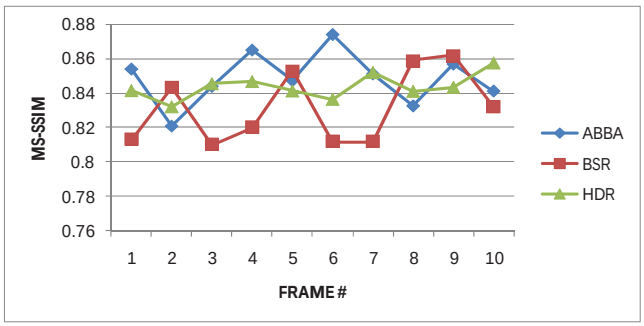


Figure 8. The Multi Scale SSIM observation

6.6 Inter Arrival Packet Delay

The inter arrival packet delay was observed on the Airtel 4G LTE-TD network during live video streaming. As shown in Figure 9, an average delay of 45.2 milliseconds was observed during experimentation. Though the proposed algorithm does not directly deal with delay profile of the packet stream, it indicates the characteristics of the underlying network.

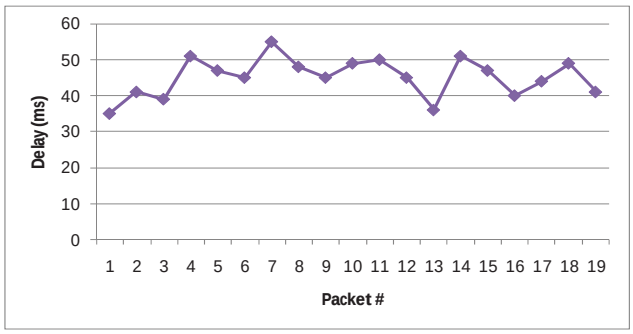


Figure 9. Inter Packet Delay for 20 packets.

6.9 Visual Frames

The Figure 10 and 11, shows the few frames of the original and decoded sequences of video during experimentation of the live streaming and stored video streaming in the laboratory environment.

7. CONCLUSION AND FUTURE WORK

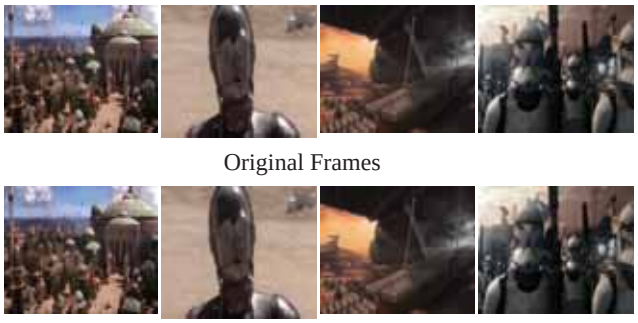
A dynamic adaptive streaming over HTTP in a wireless paradigm was implemented for live as well as stored video considering the vulnerabilities of the medium, which

withstands the fluctuations of widely varying link bit rates. Strategic decisions in real-time adaptation of video stream were based on the time series analysis of the streamed data, with an objective to achieve maximum obtainable quality. The proposed ABBA algorithm was implemented and tested on different standard video quality metrics and result showed an improvement over other two existing approach (HDR and BSR algorithm) in all quality index presented here. Although the system is targeted towards cinematic video, it can also support tele-medicine applications where doctors can communicate to their peers. The proposed work was implemented and tested in one way communication, but can be extended to two-way video streaming.

Original Frames

Decoded Frames

Figure 10. Live streaming from server to client



Original Frames

Decoded Frames

Figure 11. Stored Video Streaming for Entertainment

ACKNOWLEDGEMENT

The research work presented here is a part of the University Grant Commission (UGC), Government of India, New Delhi funded project. We would like to thank the UGC for the financial support and permission to publish the outcomes.

REFERENCES

[1] Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2015–2020, February 3, 2016.

[2] Luca De Cicco, "An Adaptive Video Streaming Control System: Modeling, Validation, and Performance Evaluation", IEEE/ACM Transactions on Networking, Volume 22, Issue 2, pp.526-539, April 2014 .

[3] Jan Willem Kleinrouweler, “Modeling Stability and Bitrate of Network-Assisted HTTP Adaptive Streaming Players”, International Tele traffic Congress (ITC 27), Ghent, pp-177–184, 2015.

[4] Junghwan Lee, Nakjung Choi, ChuckYoo, “HAVS: hybrid adaptive video streaming for mobile devices”, IEEE Transactions on Consumer Electronics , Volume 60 , Issue 2, pp.210-216, May 2014.

[5] Dhananjay Kumar, Nandha Kishore, Easwaran, A. Srinivasan, A. J. Manoj Shankar L., Arun Raj, "Adaptive Video Streaming Over Http Through 3g/4g Wireless Networks Employing Dynamic On The Fly Bitrate Analysis", ITU Kaleidoscope: Trust in the Information Society, Barcelona, December 2015.

[6] “Airtel 4G Wi-Fi Hotspot”. [Online]. Available at: http://www.airtel.in/4g/index. [Accessed 30 May 2016].

[7] Angelos Michalas, Aggeliki Sgora, Dimitrios D. Vergados, ”FDASH: A Fuzzy-Based MPEG/DASH Adaptation Algorithm”, IEEE Systems Journal, Volume 10, Number 2, pp.859-868, June 2016.

[8] Parikshit Juluri, Venkatesh Tamarapalli, “SARA: Segment aware rate adaptation algorithm for dynamic adaptive streaming over HTTP”, IEEE International Conference on Communication Workshop (ICCW), pp 1765-1770, June 2015.

[9] ITU-T Recommendation Series Y, “Global Information Infrastructure, Internet Protocol Aspects and Next generation Networks”, document # Y.1540, March 2011.

[10] ITU-T Recommendation Series J, “Reference algorithm for computing peak signal to noise ratio of a processed video sequence with compensation for constant spatial shifts, constant temporal shift, and constant luminance gain and offset”, document # J.340, June 2010.

[11] ITU-T Recommendation Series J,” Method for specifying accuracy and cross-calibration of Video Quality Metrics (VQM) ”, document # J.149, March 2004.

[12] ISO/IEC JTC1/SC29/WG11 and ITU-T SG16 Q.6, “New Video Quality Metrics in the H.264 Reference Software”, document # JVT-AB031, July, 2008.

[13] Huda M. A. El Hag, Sami M. Sharif, "An adjusted ARIMA model for internet traffic", IEEE Africon, Windhoek, pp.1-6, September 2007.

[14] Tano, Subanar, "New procedure for determining order of subset autoregressive integrated moving average (ARIMA) based on over-fitting concept", International Conference on Statistics in Science, Business, and Engineering (ICSSBE), Langkawi, pp 1-5, Sep 2012.

[15] Liang Wang and Gaetan A. Libert, "Combining Pattern Recognition Techniques with Akaike’s Information Criteria for Identifying ARMA Models", IEEE Transactions On Signal Processing, Vol. 42, No. 6, June 1994.

[16] A. Seghouane, Canberra Res, “The Akaike Information Criterion With Parameter Uncertainty”, Fourth IEEE Workshop on Sensor Array and Multichannel Processing, Waltham, pp.430-434, July 2006.

[17] “DirectShow (Windows)”. [Online]. Available at: https://msdn.microsoft.com/en-us/library/windows/desktop/dd375454(v=vs.85).aspx [Accessed 30 May 2016].

[18] “HTML5 Speed Test”. [Online]. Available at: http://speedof.me/ [Accessed 30 May 2016].

ASSESSING INTERNET PERFORMANCE OVER MOBILE NETWORKS: FROM THEORY TO PRACTICE

Eneko Atxutegi, Jose Oscar Fajardo, Eva Ibarrola and Fidel Liberal

Faculty of Engineering of Bilbao, University of the Basque Country UPV/EHU, Spain
{eneko.atxutegi, joseoscar.fajardo, eva.ibarrola, fidel.liberal}@ehu.eus

ABSTRACT

The proper execution of performance tests is of utmost importance for the analysis and evaluation of Internet-related technologies, protocols and deployment strategies. There are plenty of tools available for experimenters ranging from simulation tools, emulation equipment, to small and large scale experimentation testbeds. Each of these performance evaluation frameworks introduces a series of capabilities and drawbacks. Additionally, there is a need for using a common methodology to perform different performance tests and to create comparable outcomes. Seeking for the standardization of reliable and comparable Internet speed measurements and trying to provide end-users with trustworthy measurement tools, different SDOs have standardized different measurement methodologies. This paper focuses on the applicability of the distinct measurement frameworks and methodologies in the field of mobile Internet, with special focus on the impact of using TCP over mobile broadband connections. The paper describes a series of experiments over different frameworks and, based on the obtained results, identifies the implications of the different types of performance tests into the evaluation outcomes and states the need for large scale measurements.

Keywords— mobile networks, performance measurements, simulation, experimentation.

1. INTRODUCTION

Mobile data traffic has grown 400-million-fold over the past 15 years and it is expected to grow nearly eightfold between 2015 and 2020 [1]. All these predictions highlight the importance of mobile networks. In this respect, end-users' experienced service (QoE) and fair and contract-compliant Quality of Service (QoS) are two key factors to consider when analyzing the suitability and adaptability of Mobile Operator service towards end-users.

TCP-based and model-based quality of service (QoS) measurement methodologies has congregated a fair amount of researchers, developers and different stakeholders (i.e. ITU [2-5], ETSI [6] and IETF's [7]) reaching no agreement yet on the best feasibility vs. reliability trade-off to assess the key performance indicators (KPIs), leading to a research and development dichotomy with measurement studies regarding either transport layer [8] or IP layer [7].

Mobile networks, due to their intrinsic variability, propose a more challenging field to perform QoS and QoE measurements. Such variability, inflicted by propagation, fading and interference effects impact on the channel quality and hence its transmission rate. In this regard and related to auto-inflicted interference, user equipment's (UE's) mobility pattern and the associated speed severely impact on Signal to Interference plus Noise Ratio (SINR) and therefore on reported channel quality indicator (CQI), the available capacity variability and eventually the application-level performance [9].

Our work presents an experimentation methodology as a research guidance that could be applied to QoS measurements or other studies regarding mobile networks, transport protocols and so on. The paper is not intended to provide the reader with a novel approach but it will define and classify the current measurement approaches, presenting them together with their strengths and weaknesses, involving distinct environments, testbeds and deployments.

The highly variant environment of radio networks is a handicap for the performance of measurements in its goal to achieve the maximum available bandwidth, especially TCP-based ones. In addition to this, the existence of different TCP implementations (congestion control algorithms – CCAs-) implies different behaviors that produce different outcomes that can be seen as QoS/QoE degradation at higher layers [10].

This paper also analyzes the impact factor that LTE actors such as UEs and eNodeBs have on the fulfillment of the proposed measurement approaches. In the same way, it also covers the study of precise TCP-based performance features amongst the steps that the experimentation stair offers so as to analyze the impact that may well have.

The topic of TCP over LTE has been deeply studied with findings regarding the relation between retransmissions and throughput decay [11], undesired Slow-Start impact [12], spectral efficiency and resource assignment under different speeds [9] and promising study and definition of the adaptability and suitability of congestion controls to mobile networks [13] among other. However, none of the abovementioned works have studied various CCAs and the impact that the mobility pattern and related speed have at the same time. Our work not only covers such missing research areas but also extends some of the mentioned works. As a validation step for the accuracy of different Internet quality measurement methodologies over LTE and future TCP enhancements, a large-scale measurement campaign proposal is explained with different contexts of use, experimentation plan and considerations wrap-up.

Summarizing, the main contributions of this paper are:

- Experimentation methodology for QoS measurements, study of transport protocols and mobile networks.
- Study of LTE actors' impact regarding the use of real UEs, usage of realistic environments in terms of fading and propagation, distinct devices differences and so on.
- An analysis regarding TCP performance features.
- A proposal of large-scale measurement campaign considering previous conclusions and findings.

The paper is organized as follows: Section 2 is responsible for the explanation of the experimentation methodology with general related pros and cons and the standardization efforts regarding QoS measurements and our linkage with them. Section 3 focuses on the analysis of TCP performance tests over mobile networks among different steps of the experimentation stair. Section 4 explains the abovementioned large-scale measurement campaign proposal. Finally, Section 5 covers the conclusions and considerations report.

2. EXPERIMENTATION METHODOLOGIES

This section covers the explanation of the experimentation methodology divided in two major subsections. Firstly, the experimentation stair will be described (illustrated in Figure 1). In this regard, four different experimentation approaches are cut off highlighting the strengths and weaknesses of them in terms of scalability and suitability to model real deployments. Secondly, we review the efforts performed in different standardization bodies concerning mobile Internet performance measurement methodologies and frameworks and the linkage with our work.

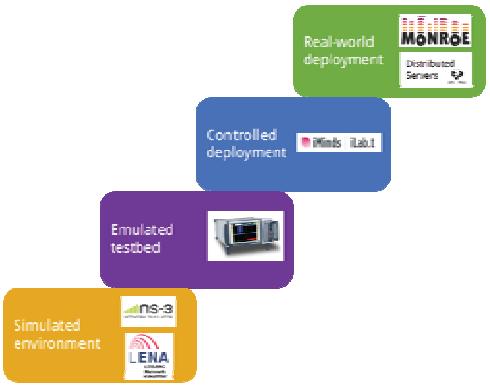


Figure 1. Experimentation stair

2.1. Experimentation stair

The description will follow the stepped structure shown in Figure 1 by the explanation of each experimentation level and the deployment associated to them. Table 1 shows a wrap-up regarding the explained pros and cons of each experimentation level.

2.1.1. Simulated environment

One of the most used deployments in research is the simulated environment. The resources that have to be spent are very little (for instance a single computer) and therefore it is an available experimentation way for almost everyone.

In the framework of this work, we deployed a hybrid simulation/emulation setup based on the ns-3 network simulator. Even though ns-3 allows performing TCP-based measurements with its natively modeled TCP flavors, the behavior of these TCP servers differ from real-world servers. Thus, we used Direct Code Execution (DCE) instead, allowing the execution of a full Linux network stack [14] and enabling the tweak of nowadays TCP/IP stack. The target was the understanding of TCP's behavior over LTE under different mobility patterns and speeds. For the correct usage and dissemination of LTE, we employed the LTE/ Evolved Packer Core (EPC) module developed in the project called LENA.

Regarding the strengths, two are the most important, the repeatability among different simulations and therefore the ability to fully compare them, and the huge possibility to gather parameters and have them under control. In contrast, applied fadings are synthetic and far from realism, the UEs are not real and so is the reporting mechanism and finally, trying to model a real deployment becomes a very tough task with many variables involved.

2.1.2. Emulated testbed

The next approach in experimentation stair comprises emulated environments. Precisely in LTE, the so-called digital radio tests or LTE emulators. These kinds of emulators play the role of the eNodeB, because they create the LTE radio signal and all the necessary LTE protocol events to support the attachment and registration of any LTE device through a radiofrequency cable or over the air.

In our work, we setup an emulated testbed in our lab with the following machines:

- UEs: Samsung Galaxy S III mobile phone and Samsung GT-B3730 LTE dongle. Both devices were connected to the emulator through a radiofrequency cable in order to avoid undesired signal behavior and for the govern prohibition in terms of non-registered radiations.
- Server: Linux server with http daemon running to serve the requested files to the UEs during the experiments.
- Digital radio test set: Aeroflex 7100 is a full Evolved UTRAN (EUTRAN)/EPC testbed with plenty of tests, configurable parameters and logging capabilities.
- Controller: Automation machine for the correct synchronization and commanding of the involved parts.

The emulated testbed aimed at involving both confirmations of previous findings and the possibility of accomplishing studies that were not suitable to achieve through raw simulated schemes. In this regard, the emulated testbed enables using real UEs and therefore managing a representative reporting to the eNodeB. Besides, it gives the opportunity to configure LTE parameters straightforward comparing with simulated environment. Finally, it has the ability to collect meaningful traces such as CQI, block error rate (BLER) and so on, making the research task easier to this respect. On the other hand, the utilized fadings have a random feature to make them closer to realism, but required

a high number of experiments to have statistically representative samples to model the behavior of both radio part and TCP. Apart from that, the emulated environments do not allow UEs' real movement, cutting down the possibility of fulfilling many studies in this regard.

2.1.3. Controlled deployment

Controlled deployments are a very good approach to experiment with real equipment and ad-hoc experiments out of the wild. Nowadays, many facilities are available due to the efforts made by federations such as Fed4Fire [15].

Our steps in relation with controlled deployments were devoted to study the behavior of TCP under different mobility patterns and speeds in a deployment with real equipment and therefore, having a real-world-alike testbed under controlled circumstances. Following cost-effectiveness, we were able to launch and study such experiments in iMinds' LTE facilities (LTE w-iLab.t [16]) in Zwijnaarde, Ghent. We performed TCP-based measurements with different mobility pattern helped by Roomba robots with LTE dongles on top of them, enabling to design, assign and follow a mobility path during the experiment in an almost fully-real environment.

Apart from all the positive features that have been highlighted regarding this kind of facilities, it has to be said that LTE transmission are done by air in a well prepared environment allowing a proper study of TCP and LTE events. Even though the movement is real, the space limitation of these small-scale deployments may well provoke to perform under limited speeds for feasibility/suitability/security reasons.

2.1.4. Real-world deployments

There is nothing more representative than performing experiments in real world deployments, world-wide if possible. However, the number of available ones is very little and the possibility to have access to one of them is very unlikely. Even though there exist works [17] that involve real nodes, they are capable of measuring simplified static scenarios and the impact that background traffic could have on foreground flows, but they lack the involvement of every phenomenon (i.e. self-induced delay due to the mobility, position according to the eNodeB and so forth) . So solve this equation MONROE [18] presents pure-realist deployment approach. It allows performing experiments in static and mobile nodes (buses, trains and trucks) with up-to-date equipment.

Apart from the aforementioned strengths, real-world deployments give trustworthy insights of movement impact, cross-traffic loads impact and they open the gate to study traffic patterns and detect deficiencies, either due to coverage, equipments or resource lack. As for the weaknesses, the possibility to gather information with high granularity is limited due to the impact of greedy background processes on overall performance and so is the ability to infer inter-layer performance. Moreover, since the utilized amount of data is under contract, there could be a limitation of data quota depending on the operator.

Table 1. Pros and cons of experimentation stair' steps

Experimentation level	Pros	Cons
Simulated environment	-Repeatability -Cheapest option -Parameters gathering	-Synthetic fadings -Faked UEs -Hard modeling
Emulated testbed	-Real UEs -Easy LTE configuration -Radio info. collection	-Statistics required -No real movement
Controlled deployment	-Real movement -Ad-hoc patterns -Air transmission	-Speed/space limitation
Real-world deployment	-Real speeds -Real patterns -Ability to study realism	-Limited parameters study -Data quota

2.2. Standardization bodies and efforts

Regarding the definition of QoS/QoE measurements and methods, the standardization efforts have been manifold [19]. The IP Performance Metrics (IPPM) WG is working on methods to measure the bandwidth. In [20] Mathis explains how TCP causes self-inflicted congestion as a result of numerous factors including TCP's circular dependencies between data rate, loss rate and Round-Trip Time (RTT) that inflict zero predictive value for any TCP-based speed test. Due to this, there is currently an ongoing remarkable effort towards the definition of a measurement mechanism that could overcome these and other TCP constraints, also being carried out by IPPM WG: Model Based Metrics for Bulk Transport Capacity [7].

The work regarding model-based metrics at IP level was extended and it was given a deeper insight of the statistical procedure in this regard [21]. The goal is to provide end-users with an 'as accurate and Congestion Control Algorithm (CCA) independent as possible' estimation. So far this alternative seems so promising that IETF IPPM WG and IETF Large-Scale Measurement of Broadband Performance (LMAP) WG [22] have standardized a framework for large-scale measurement platforms. However, doubts persist around the measurement proposal [7] regarding its accuracy, technical implementations feasibility, statistical meaningfulness and the extent to which it will resemble the actual performance end-users perceive. Moreover, the method requires the input of parameters such as Target Data Rate, Target RTT and Target Maximum Transmission Unit (MTU) to start analyzing a certain path.

As for ITU-T, it has been working in the last couple of years in the definition of new frameworks [3-4] and testing methodology [5] for Internet speed measurement in both fixed and mobile networks. Concretely, the last framework is also aligned to the draft specification TS 103 222-1 [6] from European Telecommunications Standards Institute (ETSI) and the output RFC 7594 from IETF LMAP [22] where the main entities of the framework are described. What this work amongst standardization bodies evidences is that there is an actual need for collaborating towards the definition of a common methodological technique to evaluate QoS/QoE performance in both fixed and mobile networks.

In relation to QoS measurement methodologies, we made a simulation and real-world distributed servers' study

[23] to define TCP-based measurements' constraints and ways to overcome them. The role of WS negotiation as a static constraint that may establish a ceiling in the maximum achievable effective window size (and therefore goodput) was shown. In some cases slow-start phase may be enough to fill large enough bottleneck link queue, resulting in stable goodput from that point on. However, in high bandwidth-delay-product (BDP) cases, it is no longer true. Therefore, the necessity to consider congestion epoch times as a stop criteria to ensure the achievement of the maximum available capacity (and accuracy of the estimation) has been identified. In order to shorten the test duration, the use of multi-thread was suggested based on the different tests and measurements carried out.

All the aforementioned advances propose QoS measurements that are highly focused on fixed networks features. Nowadays, QoS measurements should be adapted to mobile networks and to achieve that goal the understanding of TCP under different cellular network circumstances is essential. Our work has involved simulation environment, emulation testbed and controlled deployment and it has been devoted to this end with a triple purpose: a) Study TCP over LTE so as to provide real-world deployments with the knowledge to overcome constraints. b) Give insights to understand cross-layer behavior and be able to reason either TCP-based or model-based outcome. c) Detect extended deficiencies and verify them in different experimentation stair' layers (Figure 1).

3. APPLICABILITY TO PERFORMANCE TESTING OVER MOBILE NETWORKS

This section covers the applicability analysis of the described experimentation frameworks and methodologies for carrying out performance tests in the specific field of TCP over mobile broadband networks. Specifically, the general challenges are:

- The impact of using real-world equipment (clients, servers and network nodes) for performance tests.
- The impact of including different models to characterize the variable behavior of delay and bitrate in short and long term.
- The impact in the different phases of TCP operation.

In this regard, we analyze the importance on using real UEs, the impact of different devices, the relevance of real fading and propagation and the impact of small-scale cells. Following the different steps of the experimentation stair, some TCP features are analyzed in order to make a comparative study. Even though the study is purely TCP-based, the understanding of such behaviors would be able to provide model-based measurements with more accurate input parameters.

3.1. Impact of real-world equipment

According to the analysis in [24], in which several channel quality measurement methodologies are evaluated, none of the studies provided adequate channel quality information to the timescales required by a LTE scheduler. Two are the major issues: 1) None of the radio parameters (Received signal strength indication -RSSI- , Reference Signal Received Power –RSRP- , Reference Signal Received Quality -RSRQ- or Signal to noise ratio –SNR-) provide accurate mappings to CQI, neither for mobility nor static patterns, resulting in up to 10 units of CQI variability for a single value. 2) Even if the CQI value could be inferred precisely from physical measurements, several effects in commercial implementations prevent a direct mapping to upper layer performance. For instance, commercial UEs usually implement techniques such as Outer Loop Link Adaptation (OLLA) in order to gain or ensure robustness in its communication. In the same way, under low traffic demands, commercial eNodeB's, taking into account reported CQI, usually assign lower level of modulation and coding scheme (MCS) for robustness purposes.

These findings address the need for real UEs study so as to involve all aforementioned constraints, features and inaccuracies in the process and understand the effects that end-users may well suffer. Considering the emulated testbed as the first step in the experimentation stair that allows performing with real UEs, the usage of such testbed for a proper analysis is likely to be minimal.

Concerning the substantial importance of UE's and eNodeB's, we have measured the reported quality in many different conditions divided in two testing frameworks. In the emulated testbed we have modeled the channel conditions of two UEs (dongle and mobile phone) from their reported CQI in six of the LTE testing points that the 3GPP suggests. Table 2 gathers such testing points and their network conditions (considering different situation amongst the fading profiles of Extended Pedestrian A model –EPA-, Extended Vehicular A model –EVA- and Extended Typical Urban model (ETU)). All testing-points are static and adequate for emulated environments to avoid mobility constraint in such testbeds (as stated in section *2.1.2*).

Although there is a possibility of emulating UE's movement through the introduction of variable SINR traces, it is not a real movement and neither the fading and nor the propagation represent a purely realistic behavior so as to achieve a real variable bitrate and delay. This issue could be solved by comparing with realistic performance and finding common patterns, but cannot be studied in an isolated way.

Table 2. LTE testing points (3GPP)

No.	Propagation Condition	SNR
1	EPA5	20dB
2	EVA5	10dB
3	EVA70	20dB
4	EVA200	20dB
5	ETU70	10dB *
6	ETU300	10dB *

*Original testing point were placed in 0dB (edge of the cell). Adaptation to the middle of the cell with 10dB.

Having in mind the abovementioned testing points, Figure 2 shows as empirical cumulative distribution function (ECDF) the differences that the reporting of selected UEs have between each other in exactly same conditions with sufficient repetitions (5) and logging time (120 seconds). It is clear that for exactly the same network

conditions, different UEs could report a different quality and therefore the eNodeB will assign a different MCS for them, enabling in this case higher throughputs with the smartphone. Considering the maximum CQI value being 15, a 3-4 CQI levels difference in average values is a resounding gap between each other.

We conclude that, even using real UEs huge differences could appear. For that reason, it is important to perform experiments with up-to-date UEs and carry out comparative studies performing with similar or even same equipment, but also considering that it is difficult to infer from a certain outcome a general-purpose behavior.

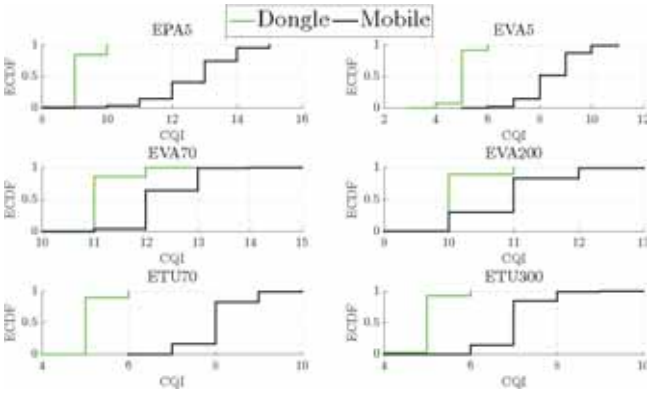


Figure 2. Reported channel quality by UEs in different conditions: Dongle and Mobile phone in emulated testbed

Secondly, we gathered the channel quality during the fulfillment of 3 different paths by the robots in w.iLab-t by means of RSSI values. The biggest difference between the designed paths is the distance to the femtocell. First path comprises the medium distance mobility pattern, whereas second path consist of the closest path. Finally, the third path forces the robot to perform in the furthest position. Figure 3 depicts the RSSI distribution of the before-mentioned mobility patterns/paths.

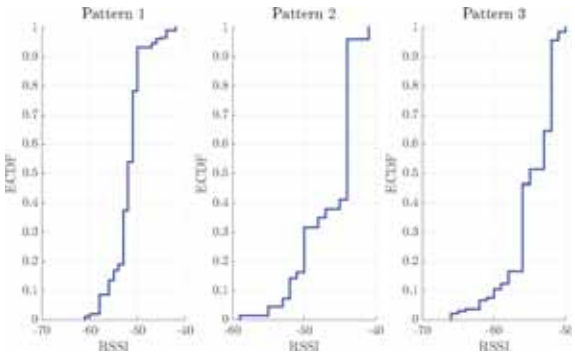


Figure 3. Reported channel quality by UEs in different conditions: Robots in controlled deployment

As a conclusion of Figure 3, we state that in this case, even though mean values are directly related to the distance from the femtocell, all patterns are in very high quality network circumstances and the possible impact of the channel quality is very little. This constraint together with the space/speed limitation of the deployment suggests a demand for experimenting over a real and large-scale deployment.

It has been demonstrated that there is a need for controlling the quality reported by the UE so as to be able to assess the impact of it. In the same way, these findings highlight the need for measuring in large-scale deployments in which the received quality by the UE is more changeable in accordance with the variability of realistic mobile networks.

3.2. Study among the steps of the experimentation stair

This subsection covers the study of precise performance features among the steps that the experimentation stair provides, analyzing them under different experimentation conditions in order to confirm or deny the resultant conclusions. The subsection is mainly divided in two TCP-based research analysis: short and long-term performance.

Firstly, this subsection covers the brief analysis of TCP's standard Slow-Start over LTE with an increase of one packet per acknowledge (ACK) and Hybrid Slow-Start [25] with the delay-awareness ability to skip earlier in order to avoid massive packet losses. The following 3 examples of both methods illustrate their performance in a short-term with special impact on short-lived flows, thus, on QoS measurements with a short execution period.

Figure 4 shows in a simulated multi-UE scenario, the ability to inject packets that both mechanism had during the time standard Slow-Start takes to converge. That is, the injection capabilities that both mechanisms show whilst standard Slow-Start converges. Standard Slow-Start has a gradual and continuous line shape, whereas Hybrid has two stages formed by the period of time in which CUBIC has ramped up as standard Slow-Start and the period after detecting a delay variation and exiting standard Slow-Start growth tendency.

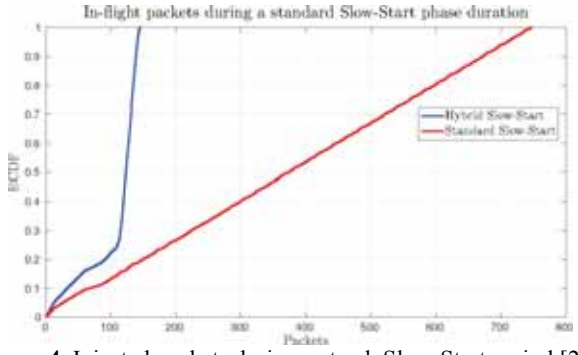
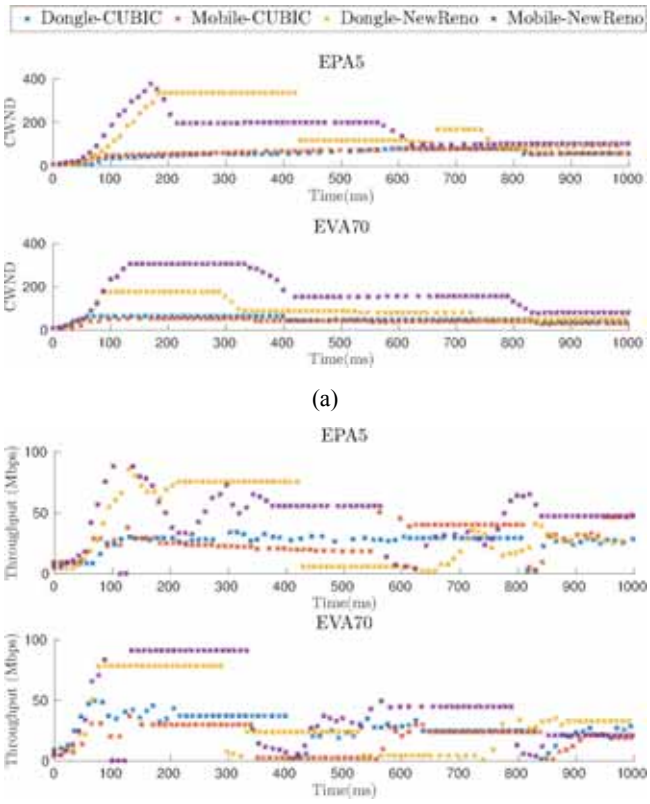


Figure 4. Injected packets during a stand. Slow-Start period [26].

Considering abovementioned behavior and willing to involve real UEs in the transmission and check out whether the effect has presence or not, we performed measurements in the emulated testbed with CUBIC (widespread CCA with Hybrid Slow-Start) and NewReno (classic CCA with standard Slow-Start) with both aforementioned UEs in the selected testing points. Figure 5 shows only 2 out of 6 cases because they are representative enough. All cases reported a quick skip of the fast ramp up for CUBIC due to the delay detection being triggered in Hybrid Slow-Start. Figure 5b depicts the impact on throughput and how Hybrid Slow-Start performs poorer.



(a)

(b)

Figure 5. Slow-Start algorithms comparison (emulated testbed): a) CWND evolution; b) Impact on throughput

The last comparison and verification example of the impact of Hybrid Slow-Start was carried out in a LTE controlled deployment. To the correct dissemination of the results, 3 tests were launched in each mobility pattern. In this case the representation of final throughput is illustrative enough to understand the effect.

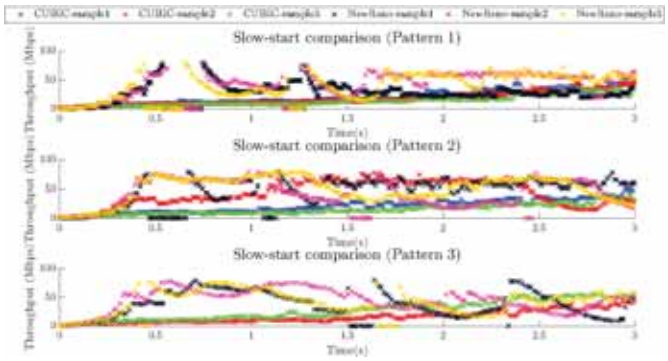


Figure 6. Slow-Start algorithms comparison (controlled deployment): Impact on throughput

The outcome of Figure 6 represents a huge impact of Hybrid Slow-Start on final achieved throughput regardless the selected mobility pattern. The whole deployment described in Section *2.1.3* is composed of real equipment and the queuing and processing delay and interaction constraints are the closest to real-world deployments ones.

The main conclusion, comparing with the results obtained with the emulated testbed or simulation environment, is that the impact is even harder for the strong implication of Hybrid Slow-Start's poor performance with delay variability. The more realistic the delay over the system, the more challenging conditions of mobile network and therefore, more complex to adapt to. Different steps in the experimentation stair confirm the detected performance of Hybrid Slow-Start but denote a very distinct impact.

Secondly, related to longer TCP transmissions in which congestion avoidance phase has a crucial role to play in the adaptability to the available capacity, different CCAs propose distinct ways to increase and decrease the congestion window (CWND) in order to achieve maximum throughput. In this regard, it is not straightforward to propose QoS measurements that are able to consider the different procedural way each CCA has.

As explained in the Section 1, many are the factors that can affect the available capacity. In ns-3 we studied the impact that the mobility inertia (positions from and to) and speed could have on CCA's ability to take advantage of all radio resources. We deployed a scenario in which the UE moved from the eNodeB to further positions and the other way around. During the transmission we studied for each MCS range the number of times that each CCA wasted transmission opportunities due to the impossibility to constantly feed the network (Figure 7). The same tests were performed with different speeds (60 kilometers per hour - kmph- and 200 kmph). Figure 7 aims to show differences among CCAs under different speeds at a glance without addressing very precise effects. For that reason, the two movement inertias are not necessary and only the backward movement is depicted. The work [26] concluded that indeed the speed had an impact, the need for studying similar effects in real deployments was highlighted due to the necessity to measure all the involved effects such as propagation delay, processing, queuing, variable fading model and delay and so forth.

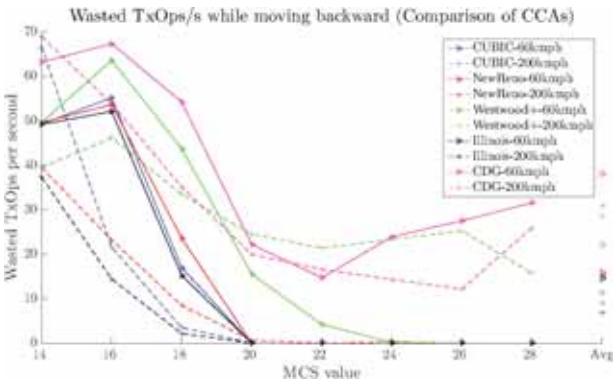


Figure 7. Different CCAs' performance at different speeds [26].

Even though all micro-effect were not considered, some important and macroscopic features were addressed regarding CCAs behavior over LTE. Firstly, Westwood+ showed a painful deficiency in cellular networks regarding its incapability to estimate the available bandwidth properly while applying the back-off, causing a huge constraint to be

overcome with a long congestion avoidance phase. NewReno and Illinois behaved very well among different mobility contexts. However, Illinois due to its delay-awareness achieved the best performance in regard to the use of available radio resources.

Considering the necessity of studying similar effects over real equipment, comparable measurements were carried out in the controlled deployment and the performance of CCAs (See differences in Figure 8) has confirmed previous macroscopic behavior findings (both the impact of Hybrid Slow-Start and CCAs' singularities).

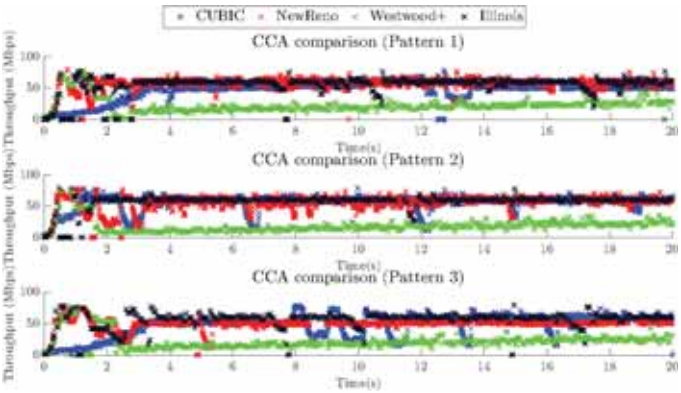


Figure 8. Different CCAs' performance (Controlled deployment)

All the results involved in this subsection have shown that performing measurements at different stair's levels help confirm findings and in the same way, detect impact differences between the measurement platforms themselves could provoke.

4. MISSING PIECE PROPOSAL: LARGE-SCALE MEASUREMENT CAMPAIGN

Our work misses only one piece and the following proposal will establish the basis for the definition of it. The work has to cope with QoS measurements dichotomy to push the methods work under real and challenging mobility conditions so as to confirm or deny related TCP-related and radio findings. The only way to fulfill all the needs addressed in the analysis of Section 3 in a deployment as realistic as possible (as remarked in Section 2) is the experimentation over real-world deployments.

As stated before, there is a clear interest in standardization and research community in general over the strengths and weaknesses of each QoS measurement method. The study should be dedicated towards the clarification of those uncertainties.

Since there is no development in relation to model-based measurement code, it is mandatory to create such required methodology. Both measurement approaches' development would be stardard-alike. Afterwards, the developments will be released as an open source contribution to the community.

The assessment and validation process will be held according to a specific plan. In the first stage no mobile node will be used, executing measurements in static nodes and representing a less disturbed figure of both methods. The tests will be only affected by LTE intrinsic variable resource assignment. Following a fair comparison between both QoS measurement philosophies and trying to mask cross-traffic effects, multiple measurements will be held. The detection of main features is very important in order to extrapolate those results and analysis to mobile schemes.

Second stage will be dedicated to constraint QoS measurement by testing them under real mobility patterns. To that purpose three main cases will be analyzed:

1) Predictable, unstable and low speed mobility pattern (PULP) – A mobility with multiple stops low average speed and quite defined traffic pattern (for instance buses). This pattern will be very challenging for both measurement methodologies for its variability regarding the channel conditions and the key feature to accomplish the hard work of measuring will be the responsiveness.

2) Predictable, constant and high speed mobility pattern (PCHP) - Mobility with straight patterns, higher speeds and less stops or at least longer gaps between stops (for example trains). This pattern is very convenient for the analysis of speed's impact and how to overcome it in order to get a proper measurement.

3) Completely unpredictable mobility pattern (CUP) - Very scattered conditions regarding stops, time of the day and speed; traffic-dependent (for instance trucks, cars and so on). This pattern is not a pattern itself since the unpredictability and uncertainty is the major characteristic. However, it is very important to have such a challenging conditions to be capable of having every imaginable condition and pattern variability wrapped up in a single day. It will be a fantastic playground for the differentiation of both measurement methods in which all possible conditions are folded.

The experimentation process should follow our recommendation regarding the experimental repetition with the same UE to be fully sure about its reported quality for a certain network condition. The measurements will also consider our findings regarding Hybrid Slow-Start so as to detect possible deficiencies while using TCP CUBIC implementation. It is also highly recommendable to define standard-compliant QoS assessments.

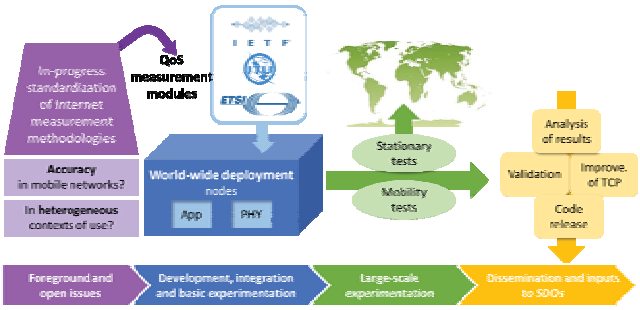


Figure 9. Conceptual proposal

The research and experimentation planning will allow us gathering broader-scale measurements to validate/extend our previous research outcomes and the standardized measurement methodologies from a world-wide mobile network perspective (conceptual proposal in Figure 9).

5. CONCLUSIONS

This paper has evaluated multiple experimentation proposals, findings and conclusions as well as the explanation of the precise research momentum of mobile Internet performance measurements. We analyze the experimentation stair with four different approaches and present our experiences in relation to QoS measurements and TCP study over LTE. Due to the intrinsic variability of mobile networks, it is been addressed the need for proper TCP understanding before the implementation of appropriate performance measurements. Four main conclusions are inferred from the described experiments: a) The use of real UEs is mandatory. b) The device could have a great impact for quality reporting reasons. c) The impossibility to perform realistic measurements in emulated testbeds in an isolated way. d) Large-scale deployments are needed for a full study and understanding.

The analysis of short transmissions, to help identify TCP performance and measurement constraints in the Slow-Start phase, has shown that Hybrid Slow-Start has a huge barrier related to its deceleration when a delay peak is detected. The sensitivity seems to be not adapted to mobile networks variability and it has been concluded that the closest to real-world deployments, the harder impact on performance. It denotes a strong need for verification and analysis over a real-world deployment so as to be able to propose enhancement to adapt CCAs to mobile networks and overcome such constraints.

Finally, through the analysis of TCP's long-term performance over LTE, it has been proven that the experimentation stair is capable of confirming macroscopic findings and behaviors amongst different steps. Considering all the aforementioned information, a large-scale mobile deployment campaign has been proposed to dispel doubts regarding QoS measurements dichotomy, TCP-based findings and so on.

For future work, we will apply the suggested proposal in the MONROE [18] open testbed to validate the framework with sufficient resources in terms of static and mobile nodes and to verify the suitability and feasibility of QoS measurement methodologies. It is also important to clarify that the validation framework has the capability to back-up explained findings related to TCP performance over LTE and give a deep input to further developments.

ACKNOWLEDGEMENTS

This work has been funded by the Spanish Ministerio de Economía y Competitividad (MINECO) under grant TEC2013-46766-R: "QoEverage" and by the European Union H2020-ICT grant agreement no. 644399 (MONROE).

REFERENCES

[1] Cisco, "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2015 – 2020," Cisco, Tech. Rep., 2016.

[2] ITU-T, Y.1543: Measurements in IP networks for inter-domain performance assessment, 2007.

[3] Y. FMIPQoS: Framework for Monitoring the QoS of IP network services, May 2016.

[4] ITU-T, Q.3960: "Framework of Internet speed measurements for the fixed and mobile networks". To be approved in 2016.

[5] ITU-T, SG-11. TD 857: "Draft new Recommendation ITU-T Q.TM_Int_sp_test: Testing methodologies of internet speed measurement system to be used on the fixed and mobile networks", September 2015.

[6] ETSI TS 103 222-1, Part 1, Section 6: "Broadband Speed Test (BST) – Measurement method of data transmission speed and transit time". November 2015.

[7] Matt Mathis and Al Morton, "Model Based Metrics for Bulk Transport Capacity," Internet-Draft draft-ietf-ippm-model-based-metrics-08,Internet Engineering Task Force, Jul. 2016.

[8] S. Shashikant et al., "*Survey of Bandwidth Estimation Techniques in Communication Networks," Wireless Personal Communications*, vol. 83, no. 2, pp. 1425–1476, 2015.

[9] R. Merz et al., "*Performance of LTE in a High-velocity Environment: A Measurement Study" AllThingsCellular Workshop* (SIGCOMM), pp. 47–52, 2014.

[10] H. Jiang et al., "*Tackling bufferbloat in 3G/4G networks*," IMC '12 Proceedings of the 2012 ACM conference on Internet measurement conference, Pages 329-342, 2012.

[11] S. Alfredsson et al.,"Cross-layer analysis of TCP performance in a 4G system," in SoftCOM, Sep. 2007, pp. 1–6.

[12] J. Huang et al., "*An in-depth study of LTE: effect of network protocol and application behavior on performance*", *SIGCOMM Comput. Commun. Rev. 43, 4*, pp. 363-374,2013.

[13] I. Johansson, "*Congestion control for 4G and 5G access*," Internet Engineering Task Force, Internet-Draft draft-johansson-cc-for-4g-5g-02, Jul. 2016, Work in Progress.

[14] H. Tazaki et al., "*Direct Code Execution: Revisiting Library OS Architecture for Reproducible Network Experiments*," in CoNEXT '13. New York, NY, USA: ACM, 2013.

[15] http://www.fed4fire.eu/testbeds/

[16] S. Bouckaert et al., "*The w-iLab.t Testbed*", Testbeds and Research Infrastructures. Development of Networks and Communities: 6th International ICST Conference, TridentCom 2010, Berlin, Germany, May 18-20, 2010.

[17] J. Garcia, S. Alfredsson & A. Brunstrom: Delay metrics and delay characteristics: A study of four Swedish HSDPA+ and LTE networks. In EuCNC, 2015.

[18] Anna Brunstrom, "*Mobile Broadband Measurement Platform - MONROE*", *IETF-93 Proceedings*. Prague, 19 July 2015.

[19] V. Bajpai and J Schönwälder, "A Survey on Internet Performance Measurement Platforms and Related Standardization Efforts", IEEE Communication Surveys & Tutorials, vol. 17, no. 3, pp. 1313-1341, Third quarter 2015.

[20] Matt Mathis, "*Why is BTC so hard?*", IP Performance Metrics Workgroup, *IETF-86 Proceedings*, March 2013.

[21] Al Morton, "*Improved Internet speed tests can enhance QoS and QoE*", *4th International Workshop on Perceptual Quality of Systems (PQS 2013)*, 2013.

[22] IETF RFC 7594 "*A framework for Large-Scale Measurement of Broadband Performance (LMAP)*".

[23] E. Atxutegi et al., "*Why we still need standardized Internet speed measurement mechanisms for end users*", ITU Kaleidoscope 2015, Barcelona, Spain, 9-11 December 2015.

[24] Jose Oscar Fajardo, "*Analysis of channel quality reporting mechanisms and their impact into 4G and beyond technologies*", PhD Thesis, December 2015.

[25] S. Ha and I. Rhee, "*Taming the elephants: New TCP slow start*," *Computer Networks*, vol. 55, no. 9, pp. 2092–2110, Jun. 2011.

[26] E. Atxutegi et al., "*TCP behavior in LTE: impact of flow start-up and mobility*", IFIP Wireless and Mobile Networking Conference - WMNC, 2016.

SYSTEMATIC ANALYSIS OF GEO-LOCATION AND SPECTRUM SENSING AS ACCESS METHODS TO TV WHITE SPACE

H. Mauwa[1], A. Bagula[1], M. Zennaro[2], E. Pietrosemoli[2], A. Lysko[3], T.X. Brown[4]

[1]ISAT Laboratory, Computer Science Department, University of The Western Cape (UWC),
Private Bag X17, Bellville, 7535, South Africa.
Email: {mhope, bbagula}@uwc.ac.za
[2]T/ICT4D Laboratory, The Abdus Salam International Centre for Theoretical Physics,
Via Beirut 7, 34151 Trieste, Italy.
Email: {mzennaro, ermanno}@ictp.it.
[3]Meraka Institute, Council for Scientific and Industrial Research,
Pretoria, South Africa
Email: alysko@csir.co.za.
[4]Carnegie Mellon University, Department of ICT,
Kigali, Rwanda
Email: timxb@andrew.cmu.edu

ABSTRACT

Access to the television white space by white space devices comes with a major technical challenge: white space devices can potentially interfere with existing television signals. Two methods have been suggested in the literature to help white space devices identify unused channels in the TV frequency band so that they can avoid causing harmful interference to primary services legally protected to run on the band. These methods are geo-location spectrum database and spectrum sensing. Discussions in the literature have placed much emphasis on the limitations of the spectrum sensing approach and mainly based on the developed world environment ignoring the performance requirements of the geo-location database approach and how the absence of these requirements in a developing region could affect its performance. This paper considers a broader analysis of the approaches by looking at factors that can affect the performance of each approach and how the presence or absence of these factors in a developed region or developing region can affect their performance. In so doing, the paper highlights the need to conduct more research on the performance of spectrum sensing in developing regions where there are plenty of white spaces to ascertain its use in these regions.

Keywords— Geo-location database, spectrum sensing, performance factors, best approach

Paper accepted for presentation at "ICTs for a Sustainable World" ITU Kaleidoscope Conference, Bangkok, Thailand, 14-16 November 2016, http://itu.int/go/K-2016.

1. INTRODUCTION

In the current static spectrum-allocation policies followed by governmental agencies, licensed holders are assigned wireless spectra on a long-term basis. The static spectrum-allocation policies have shown to be inefficient [1]. The inefficiency is more apparent in the TV broadcasting frequency band as several studies have shown that a huge portion of the assigned spectra is unused [1][2][3][4] most of the time. The unused TV channels (so called TV white spaces) have been hyped as the solution to meet the growing demand for the wireless data transmission. At the moment, governments are seeking better and innovative techniques that will offer new ways of exploiting the existing spectrum [5]. An efficient long-term solution that has been proposed is dynamic spectrum access (DSA) [6].

Detection of vacant channels by secondary devices called white space devices (WSDs) is difficult as the vacant channels vary according to location and time. Any secondary access can potentially cause harmful interference to the primary TV broadcasting services already running in the band if the channel is mistakenly assessed as vacant. This places a mandatory requirement on any white space device (WSD) to check if a primary user signal is present or absent in a channel before it goes ahead using it. Two techniques have been suggested to help WSDs do this: geo-location spectrum database and spectrum sensing.

Discussions in the literature have placed much emphasis on the limitations of the spectrum-sensing approach, which are based mainly on the developed world environment. This may have been the case because the idea to use TV white space (TVWS) originated from the developed world and the initial experiments were conducted there. Critical performance re-

quirements that each of the techniques requires to perform well have not been discussed in the literature clearly. The presence or absence of these requirements in a region could potentially affect their performance. Therefore, this paper considers a broader analysis of each approach by looking at factors that can affect their performance and also by looking at the impact of these factors in a developed region and/or developing region.

The paper is structured as follows; Section 2 gives a general discussion of the two approaches. This is followed by a discussion of the relevant performance factors of the approaches, provided in Section 3, and a discussion of what is considered to be the best approach, in Section 4. Section 5 provides ground truth evaluation of the approaches by comparing paths losses derived from measurement data versus predicted path-loss values of some propagation models being suggested for used in geo-location databases. Section 6 concludes the paper.

2. WHITE SPACE SPECTRUM DETECTION TECHNIQUES

Two approaches have been proposed to help WSDs measure the available TVWS in the TV frequency band. The first approach is to use a database processing information about known primary transmitters (geo-location spectrum database approach). An alternative is to use one device or a network of devices to physically scan the radio waves to detect the presence of TV signals (spectrum sensing approach).

2.1. Geo-location spectrum database

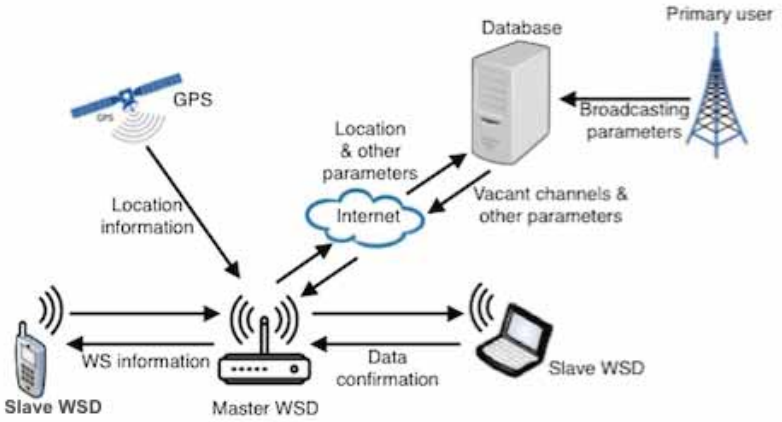
This approach consists of a WSD accessing a database of known transmitters and their primary operational characteristics such as location, antenna parameters (radiation pattern, height above the ground), transmit power, times of operation, protection requirements, allowed WSD transmitter power, and other related parameters [6][7]. The database information is used to predict which frequencies are available at different locations using the primary transmitter's details and radio propagation models. These models can be statistical, based only on distance and frequency, or they can based on ray tracing techniques that require detailed information about the terrain elevation in the area of interest. Therefore, correct identification of the presence of a TV signal at a given location depends on the fidelity of the database information and quality of the propagation model used to predict signal coverage [6]. A simple implementation example of this approach is shown in Figure 1a.

2.2. Spectrum sensing

In the spectrum sensing approach, the spectrum is analyzed by a secondary user to decide whether the spectrum is occupied or not by a primary user. This analysis is based on two general categories: energy detection and feature extraction. The energy detection can be performed with a spectrum analyzer like the Radio Frequency (RF) Explorer (*see Figure 1b*), while feature extraction is based on specific characteristics of the type of signal to be detected and is therefore more sensitive but also more complex [6]. Energy detection is the commonly proposed method due to its simplicity. It works by measuring the energy contained in a spectrum band and then comparing that with a set threshold value. If the energy level is above the threshold value, then the primary user signal is considered present otherwise the spectrum band is considered vacant.

3. RELEVANT PERFORMANCE FACTORS

Figure 2a and Figure 2b provide a general overview of the geo-location database approach and the spectrum sensing approach respectively. Some factors contribute to the optimal performance of each approach. For example, geo-location database approach requires the following factors to perform optimally in a region: propagation models whose prediction



(a) Geo-location database implementation

(b) RF Explorer

Figure 1: A simple geo-location database implementation and a Spectrum analyzer

(a) Overview of geo-location database approach

(b) Overview of spectrum sensing

Figure 2: Overview of the TVWS identification approaches

in the area of interest is close to the measured data done in the same area; an Internet backbone infrastructure to facilitate efficient and frequent communication between a master WSD and a slave WSD and a master WSD and a geo-location database; existence of detailed centralised TV database information in the area of interest. For the spectrum sensing approach to perform optimally, the following factors must exist: detection threshold value that is optimal such that there is no harmful interference to the primary users or any missed opportunities by secondary users; minimal to no mult-path fading and shadowing to avoid hidden user problem; large blocks of TVWSs in the area of interest.

4. WHAT IS THE BEST APPROACH?

Since optimal performance of each approach is dependent on some factors that may not be present or exist in some regions, none can produce superior performance in all possible regions. For example, in regions or countries where propagation models have been tried and tested extensively such that their prediction results are close to ground truth data, there is reliable Internet backbone infrastructure and a centralized detailed TV database information is also available, the geo-location database approach is expected to perform better than spectrum sensing. That could be the case with developed regions such as the US and Europe where all the three factors are present. It is therefore not surprising that the geo-location approach is being given preference as the main technique of finding TV white spaces [8] in those regions. However, conditions in developing regions are quite different from developed regions. Internet backbone infrastructure is poor and unreliable, more especially in the rural areas; propagation models have rarely been tested here such that their behaviour is unclear at the moment; spectrum usage information is scattered and stored in many formats, electronic and paper, and the regulators have not collected it into a useful centralized database that is publicly available. Therefore, at these prevailing conditions, the use of a geo-location spectrum database approach may not produce optimal results.

Fading, shadowing and the *hidden user problem*, relevant to spectrum sensing, can be severe in the developed world because of many tall buildings, which are also close together. Consequently, primary TV signals are difficult to detect accurately by spectrum sensing alone and could result in harmful interference. Therefore, performance of spectrum sensing in urban areas may be low. On the contrary, rural areas, especially those of developing world countries, where white space could be used to provide broadband connectivity, are sparsely populated with small isolated traditional building structures, which are unlikely to cause considerable fading, shadowing or bring about the *hidden user problem*. In terms of the Fresnel zone concept, these small isolated building structures could block the Fresnel zone to a maximum obstruction allowable level of not more than 40%, which produces little to no interference.

Most rural areas of developing regions, for example sub-Saharan Africa, have vast tracts of unused spectrum in the ultra-high frequency (UHF) band [9], which do not require stringent restrictions like the developed regions. This makes the use of the spectrum sensing approach a suitable alternative because problems limiting its use in developed regions may be considered to be much more forgiving than the circumstances associated with dense urban areas. Furthermore, cooperative spectrum sensing reduces errors in spectrum sensing caused by multi-path fading and is a possible solution to the *hidden user problem*. Therefore, even these problems may exist in the rural regions, the use of cooperative spectrum sensing can improve its performance.

There are some additional factors that can come into play when deciding which technique to use, apart from the above-mentioned performance factors. One such metric is the cost to implement, maintain, and administer the approaches. Geo-location approach requires a complex centralized structure and even more complex logistics, and as such, its implementation cost, maintenance cost and administrative cost is higher than spectrum sensing approach [8]. This too is likely to make spectrum sensing a more favourable approach to the developing world as most countries in this region may not have the necessary infrastructures present and must be built especially for the DSA, which is expensive.

5. GROUND TRUTH EVALUATION

As argued in the paper that there are no propagation models clearly known to perform better in regions of the developing world that can be used in geo-location databases, the first step that countries in these regions should take is to perform extensive spectral measurements and compare values of the path losses obtained from the measurements against those estimated by propagation models suggested for use in geo-location databases. Therefore, the authors did a limited physical evaluation of the approaches by conducting spectral measurements and comparing values of the path losses obtained from the measurements against those estimated by some common propagation models. The experiments were conducted in the city of Cape Town in South Africa. This section gives a detailed discussion of how the whole process was carried out.

5.1. Propagation models examined

Five propagation models were examined and compared with values from measurement data; Longley-Rice (Irregular Terrain Model), Hata for urban areas, Egli, Ericsson 9999 and Free Space Path Loss (FSPL).

5.2. TV transmitter used

An analog terrestrial television (ATT) transmitter of one of the public TV broadcasters in South Africa called South Africa Broadcasting Corporation 2 (SABC2), located on latitude 33°52′31″S and longitude 18°35′44″E, was used as a base station (BS) transmitter. Its transmission parameters obtained from the Terrestrial Broadcasting Frequency Plan 2013 document by the Independent Communications of South Africa (ICASA) [10] are: UHF channel = 22, frequency = 479.25 MHz, Effective Radiated Power (ERP) = 2 KW = 63.01 dBm, antenna polarisation = vertical.

5.3. Measurements points

Twelve locations located at different distances from the BS transmitter site were identified where measurements were done. Tables 1 shows the geographical positioning system (GPS) coordinates of the sites and their distances away from the BS transmitter. The table also shows values of the height above avarage terrain (HAAT) of the sites calculated using GLOBE 1 km Base Elevation database [11] with the number of evenly spaced radials equal to 360° in each case. Figure 3 shows the measurement points relative to the BS transmitter generated using google maps.

5.4. Spectrum measurements setup

Outdoor spectrum measurements in the UHF ATT frequency band were done at the locations using a hand-held RF Explorer model WSUB1G, which has a measurement frequency range of 240 MHz to 960 MHz. The model was fitted with a Nagoya NA-773 wide band telescopic antenna with vertical polarization, which has wide band measurement capability. The RF Explorer was connected to an Android phone installed with an android code that starts to measure spectrum immediately after the RF Explorer is connected using an On-The-Go (OTG) cable. At each site, spectrum monitoring was done for 8 hours from 08:00 in the morning to 16:00 in the afternoon.

Figure 3: Measurement sites relative to BS transmitter

5.5. Results

The 3-hour measurements at each measurement location in channel 22 from which the BS transmitter was broadcasting were averaged and the mean value was taken as the received signal power R_x. As a starting point of our analysis, we decided to confirm the square law dependence of power loss with distance first. We used the average value of the measured power at the closest point to the BS transmitter (SITE 1) and estimate from there the power received at longer distances along the same approximate radial by using the Friis transmission equation [12] (equation 1).

$$Pr(d) = Pr(d_o) + 20 * log(d_o/d) \quad (1)$$

where $Pr(d)$ is the received power at distance d in the same radial where d_o is calculated, $Pr(d_o)$ is the received power at a close-in-reference-distance d_o.

In that way, we were able to compare the measurements with those values obtained using equation 1. As Table 2 shows, there is reasonable agreement between the measured and calculated values. This confirm that the square law dependence of power loss with distance is adequate.

The accuracy of an Effective Isotropic Radiated Power (EIRP) of a transmitter, the gain and return loss of a receiving antenna determines the real path loss. In this experiment,

the BS transmitter antenna parameters such as pointing direction, pattern and gain were unknown, which brings in a degree of uncertainty about the real EIRP dissipated by the transmitter in the direction where the measurements were made. Therefore, we had to make the following assumptions in order to be able to analyse the results further:

1. The published ERP of the TV transmitter (63.01 dBm) minus 2.15 dB is the EIRP dissipated in the direction where the measurements were taken and attribute the difference between the received power at the antenna input of the spectrum analyzer and the actual power measured by the spectrum analyser at Site 1 as a resultant effect of the return loss and the antenna gain of the Nagoya NA-773 wide band telescopic antenna at the broadcasting frequency of 479.25 MHz of the TV transmitter.
2. The path loss at 0.54 Km distance from the TV transmitter is equal to the free-space path loss.

Assumption 1 is based on the fact that the square law dependence of power loss with distance is adequate and also that the measurements at similar distances from the TV transmitter are similar as shown in Table 2. The 2.15 dB subtracted from the ERP is the gain of the half-wavelength dipole antenna, assumed to be the antenna used by the TV transmitter. A dipole antenna, which is electrically one half wavelength long, in free space, exhibits a gain in its direction of maximum radiation of 2.15 dB over a theoretical isotropic radiator because it concentrates the energy in a certain direction so that the radiation in that direction is greater than the radiation from an isotropic source with the same input power.

The assumed resultant effect of the return loss and the antenna gain of the Nagoya NA-773 wide band telescopic antenna was regarded as a correction factor (CF) to every measurement. Using the FSPL equation 2, the FSPL at 0.54 Km distance and at the broadcasting frequency of 479.15 MHz was calculated as 80.71 dB, and the CF as 60.54 dB (Assumed EIRP - measurement at Site 1 + assumed FSPL at Site 1).

$$P_L = 32.45 + 20 * log(f) + 20 * log(d) \qquad (2)$$

where P_L is the free space path loss in dB, f is the frequency in MHz and d is the distance in kilometers.

The calculated CF was added to every measurement power at each location to get the corresponding antenna input power. Table 3 shows the measured power by the spectrum analyser at each measurement location and the corresponding spectrum analyser's input power.

Table 3: Measured power and antenna input power

No.	Name	d_{Km}	Measured (dBm)	Antenna input (dBm)
1	Tygerberg Natural Reserve	0.54	-80.39	-19.85
2	Harl Bremmer Hospital	2.37	-97.30	-36.76
3	Bellville Business Park	2.98	-86.45	-25.91
4	Parow Centre	3.21	-93.34	-32.80
5	Tyger Valley Shopping Centre	3.51	-94.52	-33.98
6	Bellville Market	3.89	-92.58	-32.04
7	Tygerberg Hospital	4.18	-96.43	-35.89
8	Bellefleur Flats	5.71	-97.75	-37.21
9	Parow Industrial Area	6.04	-91.77	-31.23
10	UWC	7.26	-97.33	-36.79
11	Unibell	7.55	-94.25	-33.71
12	HPR	8.04	-94.23	-33.69

Table 1: GPS coordinates of measurement sites and their distances away from BS transmitter

Site name	**Latitude Longitude**	**HAAT (m)**	d_{km}	**Site name**	**Latitude Longitude**	**HAAT (m)**	d_{km}
Tygerberg Natural Reserve (**SITE 1**)	-33°52′41″ 18°36′1″	227	0.54	Tygerberg Hospital (**SITE 7**)	-33°54′32″ 18°36′56″	-22	4.18
Harl Bremer Hospital (**SITE 2**)	-33°53′36″ 18°36′34″	27	2.37	Bellefleur Flats, Bellville (**SITE 8**)	-33°54′17″ 18°38′47″	-14	5.71
Bellville Business Park (**SITE 3**)	-33°53′59″ 18°36′30″	-4	2.98	Parow Industrial Area (**SITE 9**)	-33°55′36″ 18°37′1″	-19	6.04
Parow Centre (**SITE 4**)	-33°54′15″ 18°35′52″	-26	3.21	University of the Western Cape (**SITE 10**)	-33°56′2″ 18°37′49″	-9	7.26
Tyger Valley Shopping Centre (**SITE 5**)	-33°52′30″ 18°38′1″	30	3.51	Unibell Train Station (**SITE 11**)	-33°56′15″ 18°37′42″	-7	7.55
Bellville Market (**SITE 6**)	-33°54′12″ 18°37′14″	-11	3.89	Henry Peterson Residence (**SITE 12**)	-33°56′28″ 18°37′54″	-9	8.04

Table 2: Comparison of calculated vs measured power

No.	Name	d_{Km}	Measured (dBm)	Calculate ($Pr(d)$) (dBm)	Measured - Calculated (dBm)
1	Tygerberg Natural Reserve	d_0= 0.54	$Pr(d_o)$= -80.39	Ref. power	-
2	Harl Bremmer Hospital	2.37	-97.30	-93.24	-4.06
3	Bellville Business Park	2.98	-86.45	-95.23	8.78
4	Parow Centre	3.21	-93.34	-95.87	2.53
5	Tyger Valley Shopping Centre	3.51	-94.52	-96.65	2.13
6	Bellville Market	3.89	-92.58	-97.54	4.96
7	Tygerberg Hospital	4.18	-96.43	-98.17	1.74
8	Bellefleur Flats	5.71	-97.75	-100.01	2.26
9	Parow Industrial Area	6.04	-91.77	-101.36	9.59
10	UWC	7.26	-97.33	-102,96	5.63
11	Unibell	7.55	-94.25	-103,30	9.05
12	HPR	8.04	-94.23	-103,85	9.62

The path loss is calculated by subtracting the received signal at the antenna input at each measurement location from the EIRP (60.86 dBm). For each propagation model, the path loss from the BS transmitter is estimated for distances corresponding to those at which measurements were taken using their formulas. The path losses from the measurements and those estimated by the propagation models are shown in Table 4. To get a clearer picture of the pattern of the losses, they were plotted in graphs as shown in Figure 4. Path loss errors (average error, average absolute error, standard deviation) between measurements and the propagation models are shown in Table 5. The RMSE between the measured received signal power and that estimated using each of the propagation models is shown in Table 6.

Table 4: Path losses

No.	Name	d_{km}	Path Losses (dB)					
			Measured	FSPL	L-R	Ericsson 9999	Egli	Hata
1	Tygerberg Natural Reserve	0.54	80.71	80.71	99.70	90.68	86.32	108.88
2	Harl Bremmer Hospital	2.37	97.62	93.56	97.10	110.18	112.01	131.17
3	Bellville Business Park	2.98	86.77	95.55	99.50	110.18	115.99	134.62
4	Parow Centre	3.21	93.66	96.19	95.40	114.18	117.28	135.75
5	Tyger Valley Shopping Centre	3.51	94.84	96.97	99.70	115.36	118.84	137.09
6	Bellville Market	3.89	92.90	97.86	98.80	116.72	120.62	138.64
7	Tygerberg Hospital	4.18	96.75	98.48	99.60	117.66	121.87	139.73
8	Bellefleur Flats	5.71	98.07	101.19	109.70	121.78	127.29	144.43
9	Parow Industrial Area	6.04	92.09	101.68	102.30	122.52	128.27	145.27
10	UWC	7.26	97.65	103.28	105.50	124.94	131.46	148.05
11	Unibell	7.55	94.57	103.62	104.70	125.46	132.14	148.64
12	HPR	8.04	94.55	104.17	105.40	126.29	133.23	149.58

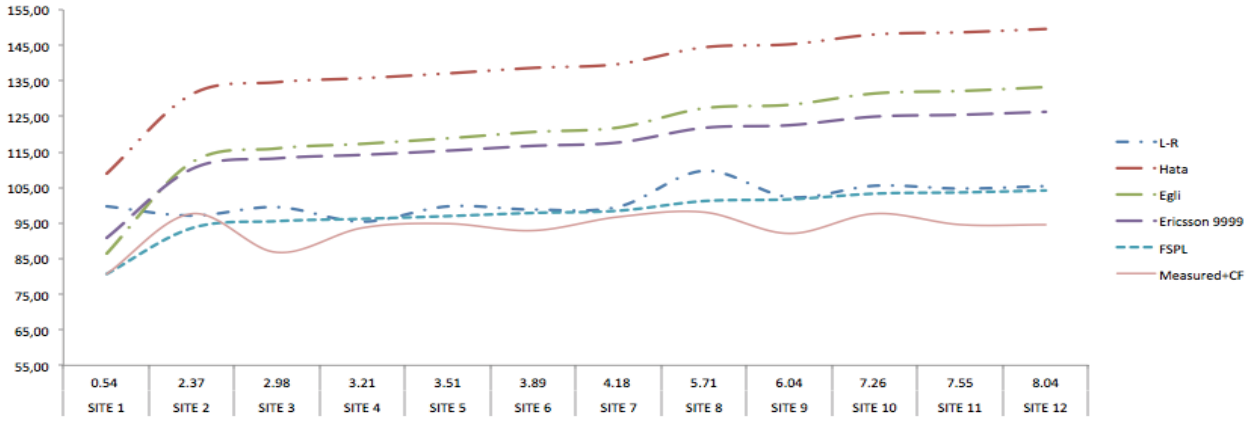


Figure 4: Plots of path losses

Table 5: Mean error, mean absolute error and standard deviation

No.	Name	Path loss Errors (dB)				
		Measured & FSPL	Measured & L-R	Measured & Ericsson 9999	Measured & Egli	Measured & Hata
1	Tygerberg Natural Reserve	0.00	18.99	9.98	5.61	28.17
2	Harl Bremmer Hospital	-4.06	-0.52	12.56	14.40	33.55
3	Bellville Business Park	8.78	12.73	26.43	29.22	47.86
4	Parow Centre	2.53	1.74	20.52	23.63	42.09
5	Tyger Valley Shopping Centre	2.13	4.86	20.52	24.00	42.25
6	Bellville Market	4.96	5.90	23.82	27.72	45.74
7	Tygerberg Hospital	1.74	2.85	20.91	25.12	42.98
8	Bellefleur Flats	3.12	11.63	23.71	29.22	46.36
9	Parow Industrial Area	9.59	10.21	30.43	36.18	53.18
10	UWC	5.63	7.85	27.29	33.81	50.40
11	Unibell	9.05	10.13	30.89	37.57	54.07
12	HPR	9.62	10.85	31.74	38.69	55.04
	Mean	4.42	8.10	23.23	27.10	45.14
	Mean absolute	5.10	8.19	23.23	27.10	45.14
	Standard deviation	4.32	5.45	6.87	9.67	8.11

Table 6: RMSE between measured and models' estimation power

Propagation Model	**RMSE** (dB)
FSPL	6.06
L-R	9.64
Ericsson 9999	24.15
Egli	28.64
Hata	45.80

5.6. Discussion of the results

From the plots of the path losses in Figure 4 and the path loss errors in Table 5, the FSPL is the closest model to the measurements. The accurancy of the FSPL can be attributed the possiblity of clear line of site between the TV transmitter and the measurement locations since the measurement locations were just few kilometers away from the TV transmitter site. The L-R (ITM), using Radio Mobile [13], was also accurate as it uses terrain elevation data of an area to compute the path loss. The Ericsson 9999, Egli and Hata underestimated the received signal power such that their estimated path losses are greater than the derived path losses from the measurements at each location.

Although the FSPL model is closest to the measurement data as its RMSE (6.06 dB) from Table 6 falls within the acceptable range of 6-7 dB for urban areas [14], we cannot conclusively say that is the best-fit model for the area within the distances where the measurements were taken. Extensive long-hours spectral measurements are needed and also more measurement sites need to be included to confirm the validity of a model for the area, which may be costly and time consuming.

6. CONCLUSION AND FUTURE WORK

Discussions in the literature have placed much emphasis on the limitations of the spectrum sensing approach and mainly based on the developed world environment ignoring the limitations of the geo-location database approach and its impact on the developing world. This paper considered a broader analysis of the approaches by looking at factors that can affect the performance of each approach and how the presence or absence of these factors in a developed region and/or in a developing region can affect their performance. In so doing, the paper has highlighted the need to conduct more research on the performance of spectrum sensing in regions where plenty of white spaces are available, e.g., rural areas of developing world countries.

Our analysis shows that information that is needed by the geo-location database approach to perform optimally may not exist in most developing world countries especially in the rural areas where telecommunication infrastructure is lacking and white spaces are abundant. Based on the absence of factors that are needed by the geo-location approach to perform optimally in developing world regions, presence of abundant white spaces in rural areas of developing world regions and also considering the implementation, maintenance and administrative cost of geo-location approach, we have concluded that spectrum sensing is more favourable to use in rural areas of developing world countries than geo-location database approach.

Since cooperative spectrum sensing eliminates the limitation of local spectrum sensing, the authors would like to investigate a white space network detection framework based on cooperative spectrum sensing using hand-held spectrum analyzers as part of their future work, and also extend the analysis to find the available white spaces at UWC measurement site, which could be used for campus networking.

REFERENCES

[1] I.F. Akyildiz, W. Lee, M.C. Vuran, and S. Mohanty, "Next generation/dynamic spectrum access/cognitive radio wireless networks: a survey," *Computer networks*, vol. 50, no. 13, pp. 2127–2159, 2006.

[2] R. Kaniezhil, C. Chandrasekar, and S. NithyaRekha, "Performance evaluation of qos parameters in dynamic spectrum sharing for heterogeneous wireless communication networks," 2012.

[3] Y. Yuan, P. Bahl, R. Chandra, T. Moscibroda, and Y. Wu, "Allocating dynamic time-spectrum blocks in cognitive radio networks," in *Proceedings of the 8th ACM international symposium on Mobile ad hoc networking and computing*. ACM, 2007, pp. 130–139.

[4] J. Carlson, N. Ntlatlapa, J. King, F. Mgwili-Sibanda, H. Hart, C. Geerdts, and S. Song, "Studies on the use of television white spaces in south africa: Recommendations and learnings from the cape town television white space trial," electronic, https://www.tenet.ac.za/tvws/recommendation-and-learnings-from-the-cape- town-tv-white-spaces-trial, 2013.

[5] T. Yücek and H. Arslan, "A survey of spectrum sensing algorithms for cognitive radio applications," *Communications Surveys & Tutorials, IEEE*, vol. 11, no. 1, pp. 116–130, 2009.

[6] T. Brown, E. Pietrosemoli, M. Zennaro, A. Bagula, H. Mauwa, and S. Nleya, "A survey of TV white space measurements," in *e-Infrastructure and e-Services for Developing Countries*. 2014, pp. 164–172, Springer International Publishing.

[7] L. Doyle, *Essentials of cognitive radio*, Cambridge University Press, 2009.

[8] V. Gonçalves and S. Pollin, "The value of sensing for TV white spaces," in *New Frontiers in Dynamic Spectrum Access Networks (DySPAN), 2011 IEEE Symposium on*. IEEE, 2011, pp. 231–241.

[9] E. Pietrosemoli and M. Zennaro, *TV White Spaces. A pragmatic approach*, vol. 1, chapter 4, pp. 35–40, ISTB, December 2013.

[10] ICASA, "Terestrial broadcasting frequency plan 2013," Tech. Rep., Independent Communications Authority of South Africa, 2013.

[11] Federal Communications Commission et al., "Antenna height above average terrain (haat) calculator," *Online, Retrieved*, vol. 4, no. 7, pp. 16.

[12] S. Rao, "Estimating the zigbee transmission-range ism band-designers of short-range wireless devices in the 900-mhz and 2.4-ghz band need to understand what and how parameters affect the transmission range," *EDN*, vol. 52, no. 11, pp. 67–74, 2007.

[13] R. Coudé, *Radio Mobile - RF propagation simulation software*, http://radiomobile.pe1mew.nl/, 1998.

[14] N. Blaunstein, D. Censor, D. Katz, A. Freedman, and I Matityahu, "Radio propagation in rural residential areas with vegetation," *Progress In Electromagnetics Research*, vol. 40, pp. 131–153, 2003.

TASK-BASED PROCESS MODELING FOR POLICY MAKING IN SMART CITIES

Leonidas Anthopoulos

Associate Professor, University of Applied Science (TEI) of Thessaly, Greece, lanthopo@teilar.gr

George Giannakidis

Center for Renewable Energy Sources and Saving, Greece, ggian@cres.gr

ABSTRACT

Several competitive standards have been introduced for smart city quite recently, which define the architecture and its components or city key performance indicators. However, these standards do not discuss smart service formulation, nor policy making process modeling. Standardization assists in achieving process automation by introducing "best practices" as standard process models. Policy making mainly follow non-standardized procedures, even if it is supported by various tools (i.e., Multi-Criteria Decision Methods (MCDM)). Inspired by the Task-Based Modeling method (TBM) this paper focuses on policy making process standardization for smart cities. It utilizes the case-study of the InSmart (Integrative Smart City Planning) coordination action in the smart city of Trikala, Greece, in an attempt to define and introduce a model for such a process.

Keywords— standardization, smart city, policy making, decision making, case study, TBM.

1. INTRODUCTION

Smart city has evolved since its initial appearance in 1996 to an emerging market and to a multidisciplinary scientific domain. Quite recently, most standardization bodies around the globe have introduced several competitive standards in their attempt to clarify the smart city context and corresponding industrial products.

However, standardization has left out so far smart service modeling as well as corresponding policy making. For instance, smart city owners claim that smart urban solutions address significant challenges (i.e., climate change and local growth) and in this regard, they plan corresponding policies that deal with city's energy efficiency and sustainability. Such policy making is mainly supported by decision making tools (i.e., Multi-Criteria Decision Methods (MCDM)) and has not been modeled yet. This lack in standardization mainly occurs due to the broad context of city policy making and due to the extensive scope of smart services. This paper addresses the above problem and aims to discuss the policy making process that deals with one of the most important smart city challenges: city energy efficiency. In this regard, this paper aims to provide with answers the following research questions:

RQ1: do existing smart city standards provide guidelines for city's energy efficiency policy making?

RQ2: how can smart city energy efficiency policy making be modeled and standardized?

Both the above questions are very important to be answered, since smart cities and corresponding standards are being evolved, while several energy efficiency solutions appear to deal with city's sustainability. In this respect, it is important for standards to cover important smart city issues that have not been addressed yet. Moreover, recent studies [1] show that the urban system is so complex that existing eco-friendly solutions can succeed only temporarily in their targets and in this regard a long-term and standardized policy making has to be performed.

In order to provide with answers the above research questions, this paper focuses on standardization of the policy making process for city energy efficiency. It utilizes the Task-Based Modeling (TBM) method [2] as a means to map the corresponding policy making process. Moreover, it uses findings from the project InSmart (Integrative Smart City Planning), which is a European Coordination Action and it is being implemented with the collaboration of 10 partners (3 of which are cities and 2 universities) coming from 4 European countries. More specifically, it uses the outcomes from experiments that took place in the city of Trikala for this purposes of this project.

The remainder of this paper is structured as follows: section 2 concerns a background section, where both the smart city standardization and the process standardization topics are investigated. Next, section 3 contains this paper's research methodology, which concludes to the task-based process model. Finally, section 4 contains some conclusions and future thoughts.

2. BACKGROUND

Standardization refers to the consistent use of methodologies, procedures, tools, and techniques specified above the level of individual projects [3;4]. Standards contain specification documents, rules and guidelines for product or process development [3], while they establish technological convergence in different industrial sectors [5].

Smart city on the other hand, has been quite recently defined as *innovation –not necessarily but mainly based on information technologies-, which enhances urban living in terms of people, governance, economy, mobility, environment and living* [6]. Moreover, standardization bodies have also given corresponding definitions: the International Telecommunications Union (ITU) [7] emphasizes on ICT and considers *a smart sustainable city as an innovative city that uses information and communication technologies (ICTs) and other means to improve quality of life, efficiency of urban operation and services, and competitiveness, while ensuring that it meets the needs of present and future generations with respect to economic, social and environmental aspects*. Similarly, the International Standards Organization (ISO) [8] recognizes smart city as *a new concept and a new model, which applies the new generation of information technologies, such as the internet of things, cloud computing, big data and space/geographical information integration, to facilitate the planning, construction, management and smart services of cities*. Literature regarding "smart city" and "standardization" is poor (Table 1): a crawl in ScienceDirect for "smart city" AND "standard" returned only 16 journal articles on June 2016, all of which are irrelevant to developing standards for smart cities. Only some (i.e., [2]) discuss the determinants for smart service use, which have some relevance with the context of this paper. Scopus returned the triple size (43 articles), some of which investigated particular standards in the smart city nexus (i.e., Machine-To-Machine (M2M) Communications and Internet-of-Things (IoT) [9; 10], socio-economic issues [11;12], ubiquitous networks [13], wireless sensor networks [14] etc.).

Table 1: Findings from "smart city" and "standard"

Source	Results	Articles after screening
Scopus	43	[9; 10; 11; 12; 13; 14]
ScienceDirect	16	[2]

Nevertheless, several competitive smart city standards have been introduced recently, some of which are presented on (Table 2): ISO defined the 37120:2014 standard, which defines and establishes methodologies for a set of indicators to steer and measure the performance of city services and quality of life. Moreover, it provides the technical specifications document 37151:2015, which gives principles and specifies requirements for community infrastructure performance metrics, while it gives recommendations for analysis of community infrastructures. Furthermore, ISO provides several specification documents and reports for individual smart city components (i.e., ISO/TR 28682:2008 for Intelligent Transport Systems). ITU has delivered several technical reports and specifications for smart sustainable cities. The National Institute of Standards and Technology (NIST) in the U.S.A. has triggered 2 initiatives (Global City Teams Challenge and the working group for Internet-of-Things (IoT) Enabled Smart City Framework) in its attempt to define a smart city standard. On the other hand, the European standardization organizations CEN/CELENEC/ETSI grounded their Smart and Sustainable Cities and Communities' Coordination Group (SSCC-CG) to define a smart city standard. In this regard, it also funded a coordination act in 2015 and invited standard institutes from member states to contribute on a common smart city standard, which considers smart city as a system. The British Standards Institute (BSI) has organized its Smart City Advisory Group and drafted a set of Publicly Available Specifications (PAS) 180, 181 and 182. From the remainder, European states only the Spanish Standards (AENOR) have initiated corresponding standardization activities and resulted to the UNE 178301 and UNE 178303 standards, while it adopted the ISO 37120 document with urban sustainability indicators.

Table 2: Smart city standards

Standardization Body	Standard
ISO [8]	ISO/TR 37120:2014, Sustainable development of communities - Indicators for city services and quality of life ISO/TS 37151:2015, Smart community infrastructures — Principles and requirements for performance metrics
ITU [7]	Smart Sustainable Cities
NIST [15; 16]	IoT-Enabled Smart City Framework Global City Teams Challenge[1]
CEN/CELENEC/ ETSI[2][3] [17]	Development of system standards for smart cities and communities' solutions
BSI[4] [18]	PAS 180 Smart city terminology PAS 181 Smart City Framework PAS 182 Smart city data concept
Spanish Standards (AENOR)[5]	UNE 178301 on Open Data UNE 178303 requirements for municipal assets' management. UNE-ISO 37120 adopts ISO urban sustainability indicators

On the other hand, process standardization has attracted scientific attention since the late 1970s (Table 3). ScienceDirect returned 246 journal articles on June 2016, with the keywords "process" AND "standardization", with no time limits but with a focus on the topics combining "standard, model, process, system, project, technology and

[1] http://www.nist.gov/public_affairs/releases/nist-global-city-teams-challenge-aims-to-create-smart-cities.cfm

[2] http://www.cencenelec.eu/standards/Sectors/SmartLiving/smartcities/Pages/default.aspx

[3] https://ec.europa.eu/research/participants/portal/desktop/en/opportunities/h2020/topics/353-scc-03-2015.html

[4] http://www.bsigroup.com/en-GB/smart-cities/Smart-Cities-Standards-and-Publication/

[5] https://eu-smartcities.eu/content/new-set-smart-cities-standards-spain

company". This focus attempted to leave out articles that discuss biology and health, as well as food and agriculture. A screening process followed and left out articles discussing standardization or process alone, as well as articles irrelevant to the context of this paper such as the ones that discuss the standardization process. Furthermore, Scopus returned 4,332 journal articles on June 2016. After the application of filters to demonstrate works regarding Engineering, Computer Science, Environmental Science and Mathematics, this number decreased to 447 journal articles, on which a similar screening process was followed. Screening left out irrelevant articles, like the ones discussing processing in general (i.e., textile processing).

Table 3: Findings from "process" and "standardization"

Source	Results	Citations after screening
Scopus	447	[20; 21; 22; 23; 24; 25; 26; 27; 28]
ScienceDirect	246	[3; 19]

Process standardization is defined in alternative ways [20], all of which agree with regard to the identification and unification of variants and the establishment of information interchange between different systems or components. Moreover, process standardization describes the extent to which the organization follows recurrent processes and adheres to established standards [5]. Process standardization enables performance measurement and sets the basis for continuous improvement. Different process modeling approaches are located in literature, which have been applied on different sectors (i.e., construction [20;21]; car industry [22], business and management [23; 24], Information Technology (IT) [25; 26; 27] and Health [28] etc.). In fact, software process standardization appears to have positive impact on software flexibility and project performance [5]. Some important models that were discovered in this literature review concern the TBM [2] and the Plan-do-check-act cycle [29; 30].

The above literature analysis returned useful findings: the competitive standards presented on (Table 2) indicate the existence of an "open race" regarding smart city clarification and the standardization of corresponding solutions. However, this evidence shows that smart service and policy making standardization have been left out so far. ITU's [7] is the only standard that defines a set of primary smart services, but there's still much work to be done in this regard [31]. On the other hand, policy making standardization has not been modeled according to literature findings and to the existing smart city standards. These findings provide with answer the research question RQ1.

3. RESEARCH METHODOLOGY: THE CASE OF THE INSMART PROJECT

Defining a policy making process is not a simple procedure, since it varies according to the context of the drafted policy. This paper focuses on policy making regarding transforming a city to a more energy efficient one. Energy efficiency is one of the primary smart city challenges and many solutions have been suggested for cities [1] like smart buildings, renewable energy unit and smart grid installation etc., which seem to have only temporary effect on their mission achievement [1]. This argument obliged the authors to look for alternatives in an attempt to define a proper policy making process.

In this respect, this paper uses the case study of the project InSmart (Integrative Smart City Planning) (http://www.insmartenergy.com/), which is a coordination action that is funded by the 7th European Framework Program (FP7). This project started in the early 2014 and will last until the early 2017, and it is being developed with the contribution of 10 partners from 4 countries (UK, Italy, Portugal and Greece). Four (4) representative and different European cities participate in the consortium (Nottingham (UK), Cesena (Italy), Evora (Portugal) and Trikala (Greece)). Each of the cities has special needs, while it has developed different types of smart technologies (Nottingham has emphasized on renewable energy; Cesena on smart public lighting; Evora on smart grids; and Trikala on smart transportation). The aim of this project is multi-dimensional and its tasks concern the following: (a) it investigated the potentially different sources of energy supply and demand within the involved cities; (b) it defined a reference framework (baseline) for energy demand calculation in the project's start, with the use of data coming from 2012; (c) it collected scenarios from all the involved cities regarding policy making for energy efficiency; (d) it developed a model for energy demand prediction by 2030, which can test the contributed scenarios; (e) it involved city stakeholders in all the partner cases in order to execute a MCDM for scenarios' prioritization; (f) it calculated scenarios' effect on policy targets. Today, this project finalizes the outcomes from the final calculation in all the involved cities (task f). In the city of Trikala in Greece, the above tasks resulted to the following outcomes:

(a) **Reference framework:** the baseline accounted that Trikala is organized in 20 zones, inhabited by a smoothly increasing population, while the majority of buildings are mainly used for housing purposes.

(b) **Energy demand sources:** buildings (organized in 4 typologies); water and sewage process; waste chain; and transportation.

(c) **Energy supply sources:** heating oil; transportation diesel and gasoline, natural gas; solar panels; and biomass.

(d) **Scenarios definition:** 15 alternative scenarios were tested by the model and the calculated outcomes were compared with the reference framework (baseline). These 15 scenarios concerned alternative energy savings policies that are being considered or being developed by the Municipality of Trikala, in order to comply with the Covenant of Mayors for Climate & Energy objectives[6], which had signed. This mix of scenarios (Table 4) concerns various activities that address all the 5 energy demand sources and result to energy savings.

[6] http://www.covenantofmayors.eu/index_en.html

(e) **Scenarios execution:** calculations regarding the estimated cost and the energy efficiency's outcome of each scenario were performed with the use of the project's models (Fig. 1), accompanied by maintenance costs and potential income. The analysis was performed using a City Energy System Model based on the TIMES model generator [32]. TIMES (an acronym for The Integrated MARKAL-EFOM System) is a least cost optimization model generator developed by the ETSAP IEA Technology Cooperation Programme (http://iea-etsap.org/). The TIMES energy economy is made up of producers and consumers of commodities such as energy carriers, energy services, and emissions. The model's objective is to minimize the total cost of the energy system, satisfying a number of constraints related to the availability of energy resources, availability of funds to invest in energy efficiency measures and limits of green-house-gas (GHG) emissions etc. The model outputs include the flows of energy forms and their prices, the required level of investments in energy technologies and the associated cost, the annual operation and maintenance costs of each component of the energy system and GHG emissions. The model was adopted in order to represent the energy system of Trikala and then to run the alternative scenarios presented in Table 4. The model results were used to calculate the quantitative criteria (first five criteria in Table 5) of the MCD analysis (Fig. 1). These results show that each policy result to different outcomes, which have to be evaluated and selected properly. The calculation functions' presentation is beyond the purposes of this paper.

(f) **Scenarios' criteria selection definition:** a MCDM process was followed with the contribution of the city stakeholders (Municipality, Commercial Chamber, Technical Chamber and Commercial Association), which play significant economic roles within the city. The purpose of the MCDM was to define and prioritize the criteria for scenarios selection and then to demonstrate how this prioritization affects the selection. A set of 9 criteria were defined and evaluated by the stakeholders (Table 5) with the use of the PROMETHEE Decision Making methodology. The first five criteria in the table are quantitative criteria and are defined as the difference of the value of the parameter in the baseline scenario compared to its value in the scenario under consideration. For instance energy savings are defined as total energy consumption in the baseline scenario minus total energy consumption in the policy scenario, while implementation cost is defined as the difference between investment costs in the policy scenario minus investment cost in the baseline scenario.

Table 4: Scenarios definition

Group	Scenario
Buildings	1. Municipal building renovation (20% improved efficiency) 2. 80% of city buildings connected with the natural gas network 3. Renovation of all city buildings grounded before 1950 4. Energy efficient upgrade of all city buildings
Public lighting	5. Lights' replacement with LED
Renewable Energy	6. Renewable energy production by 10% of total demand
Green Spaces	7. Green Open Space creation (5% cooling demand reduction)
Transportation	8. Mobility Ring-Road (8C) and Cycle Lane Network with 5-10 Km (8R) 9. Replacement of 10 municipal vehicles with electrical ones 10. Encouraging hybrid and electrical vehicle use (i.e., with tolls in the city entrance)
Water and sewage	11. Biomass landfill (950KWh production capacity) 12. Sewage treatment with bacteria (25% decrease of energy demand) 13. Dam construction (200KWh)
Systemic	14. Solar panels on all terraces. 15. 20% of CO2 production decrease

The implementation cost efficiency is the ratio of these two quantities expressing the Euros spend per kWh saved in the implementation of each scenario. Generated income refers to the income that is produced for the citizens and the municipality from the sales of electricity produced by renewable energy sources (mainly solar PVs). In the MCDM process stakeholders' inputs were given to the corresponding software application and all the scenarios were prioritized accordingly (Fig. 3). The PROMETHEE MCDM is suitable to problems with limited options (alternatives) and results to prioritized decisions, according which the decider makes a selection. It uses both quantitative and qualitative criteria for options' prioritization, while variables (weights) are calculated and assigned to these criteria. Moreover, this method compares the alternatives in pairs, according to each criterion and to decider's preferences. In order for the criteria weights' to be calculated according to decider's preferences, a 2-dimensional matrix is used, where each criterion is compared with all the remainders.

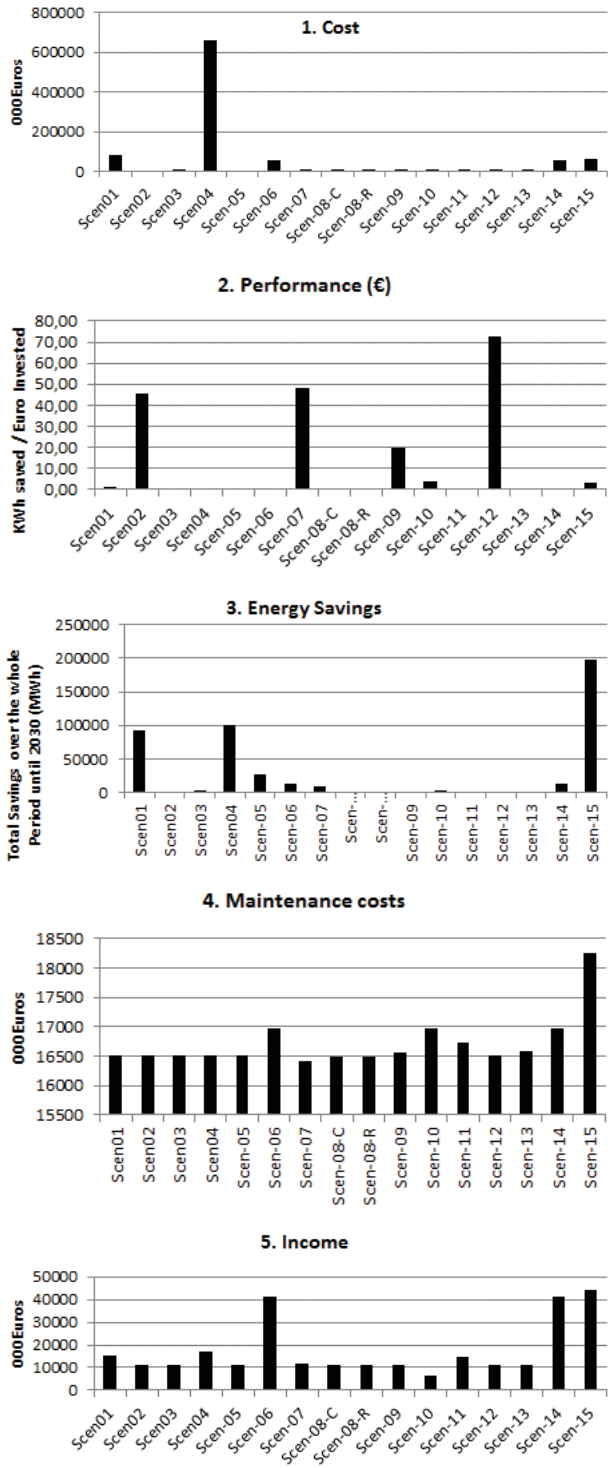


Figure 1. Scenarios simulation

Similar outcomes to Trikala's have been generated in all the involved cities, but their presentation is beyond the purposes of this paper. The followed policy making method has significant advantages: it calculates policy in terms of its expected efficiency. In the examined case, policy had to do with city's energy efficiency. Since there are several alternatives with different impacts, scenarios have to be calculated with regard to their estimated impact, in terms of money and potential. Moreover, policy success has to do with outcomes' adoption. In this regard, an important contribution of the above methodology is that social engagement occurred in practice with a bottom-up method, since the most important city stakeholders contributed with their perspectives and prioritized policy options. To this end, they affected policy making and the selected policies are more likely to succeed in terms of social acceptance.

Table 5: MCDM criteria

1. Implementation Cost
2. Implementation Cost Efficiency
3. Energy savings
4. Operation and Maintenance Cost
5. Revenue Production
6. Ease of Implementation
7. City's Quality of Life Improvement
8. City's Economic Development Improvement
9. Social Acceptance

4. TBM AND THE POLICY MAKING PROCESS

According to the Task-Based Modeling method (TBM) [2] a process is analyzed in key management tasks and each task is accompanied by an execution method and by detailed attributes. Instead of trying to standardize a business process, TBM intends to focus on standardizing lower level management tasks. TBM enables the clear definition and the outcome's measurement of a process. A management task represents a management action or a series of actions to be taken in order to achieve specific object or function. To this end, each management task has the following features, regardless its complexity [2]:

- Action: it takes an action or a series of related actions to execute a task. An action is characterized by a "verb";
- Method: it may be conducted with certain methods;
- Object: the action may work on an object that affects the means of executing a task.

Management tasks can be *fully* or *semi-automated* or *fully manual* [2]. Fully automated tasks can be conducted by computers via the appropriate software (i.e., the action send_email[what, whatElse] sends an e-mail message with content "what" to recipients "whatElse"). Semi-automated tasks on the other hand, require the interaction between a decider and the computer during a task (i.e., retrieving historical spending data with adjustments made by a human). Finally, manual tasks concern complex decision making tasks, which can be only performed by qualified humans.

A management process is driven by a request -which is a question, a process or a need- and it is executed with the following steps [2]: (a) raise a request; (b) create an instance of a process model; and (c) execute the model -task by task-, and trace the dynamic status of the model.

All the above information can be adjusted with the process that was presented in the previous section 3, where policy making was analyzed in 6 specific tasks (a-f). The extracted TBM consists of a total of 17 tasks, it is depicted on (Fig. 2) and may vary between cities. Although this model has been defined for energy efficiency purposes in smart cities, it is quite generic and it may be applicable in other

corresponding policy making problems too, since it incorporates MCDM and defines decision options and corresponding criteria. In (Table 6) an analysis of all the steps is demonstrated. In this process model, there are 10 different tasks as follows:

- DefineBaselineStructure: a manual task. It requires a manager to define the structure of the baseline.
- SearchDataBase: an automated task. It searches records from a given database;
- UpdateDatabase: an automated task. It updates a database with new records.
- Search_Manual: a manual task. It involves a manual process to search for a list of qualified items;
- DefineScenarios: a manual task. It requires a manager to define the alternative scenarios.
- ExecuteScenarios: an automated task. It uses the model and the energy data to calculate policies' efficiency.
- DefineScenarioSelectionCriteria: a manual task. It requires a manager to define the selection criteria.
- RankCriteriaWeights: a manual task. It requires stakeholders to rank the importance of the selection criteria.
- ScenariosPrioritization: an automated task. It uses MCDM methodology to evaluate the scenarios according to the criteria and to their ranking.
- SelectScenario: a manual task. It requires a manager to select the optimal scenario.

The identified TBM provides with answer RQ2 and can standardize a typical decision making process.

5. CONCLUSIONS – FUTURE THOUGHTS

This paper focused on the problem of policy making in smart cities and more specifically, it focused on policies for city energy efficiency. It defined RQ1 and RQ2 research questions and followed different methodologies to provide them with answers. More specifically, it used literature evidence to answer RQ1, which show that existing smart city standards do not provide with rules or guidelines policy making, but they contain specifications for parts of energy efficiency components (i.e., buildings, smart grids etc.). In order to answer RQ2, the TBM was utilized for modeling the policy making process. Using the task components as the basic modeling elements and the experimentation that took place in the city of Trikala in Greece for the purposes of the project InSmart, a process model was created by connecting the needed tasks. The identified model has been tested in Trikala and remains to be tested in the other participating cities too until the end of 2016. Such a standardized process can become a guide for similar tasks that deal with smart city energy efficiency policy making. Some future thoughts -beyond this further testing- concern the identification of similar process models for other decision making procedures in smart cities.

In order to answer RQ2, the TBM was utilized for modeling the policy making process. Using the task components as the basic modeling elements and the experimentation that took place in the city of Trikala in Greece, a process model was created by connecting the needed tasks. The identified model remains to be tested in the other participating cities too until the end of 2016. Such a standardized process can become a guide for similar tasks that deal with city energy efficiency policy making. Some future thoughts concern the identification of similar process models for other decision making procedures in smart cities.

6. ACKNOWLEDGEMENTS

This paper was supported by the project InSmart (Integrative Smart City Planning) (http://www.insmartenergy.com/), which is a coordination action that is being funded by FP7 European sources.

REFERENCES

[1] N. Tsolakis and L. Anthopoulos, "). Ecocities: An Integrated System Dynamics Framework and a Concise Research Taxonomy," *Sustainable Cities and Society*, vol. 17, pp. 1-14, 2015.

[2] J.J. Shi, D-E Lee, and E. Kuruku, "Task-based modeling method for construction business process modeling and automation," *Automation in Construction*, vol. 17, pp. 633-640, 2008.

[3] S. Gauch and K. Blind, "Technological convergence and the absorptive capacity of standardisation," *Technological Forecasting & Social Change*, vol. 91 pp. 236–249, 2015.

[4] J. Wettig, "New developments in standardisation in the past 15 years - product versus process related standards," *Safety Science*, vol. 40, pp. 51–56, 2002.

[5] G.E. Moore, "Standardization," *Microelectronics Journal*, vol. 10, No. 2, pp. 30-34, 1979.

[6] L. Anthopoulos and Ch. Reddick, Understanding electronic government research and smart city. *Information Polity*, vol. 21, no. 1, pp. 99-117, 2015.

[7] International Telecommunications Union (ITU), "Smart sustainable cities: An analysis of definitions," 2014.

[8] International Standards Organization (ISO), "Smart cities Preliminary Report 2014," 2014, retrieved, May 2016 from http://www.iso.org/iso/smart_cities_report-jtc1.pdf

[9] D. Belanche, L.V. Casaló, and C. Orús, "City attachment and use of urban services: Benefits for smart cities," *Cities*, vol. 50, pp. 75-81, 2016.

[10] P.K. Verma, R.Verma, A. Prakash, A. Agrawal, K. Naik, R. Tripathi, M. Alsabaan, T. Khalifa, T. Abdelkader and A. Abogharaf, "Machine-to-Machine (M2M) communications: A survey," *Journal of Network and Computer Applications*, vol. 66, pp. 83-105, 2016.

[11] R. Neisse, G. Steri, I.N. Fovino and G. Baldini, "SecKit: A Model-based Security Toolkit for the Internet of Things," *Computers & Security*, vol. 54 pp. 60-76, 2015.

[12] M-L Marsal-Llacuna, "Building Universal Socio-cultural Indicators for Standardizing the Safeguarding of Citizens' Rights in Smart Cities," *Social Indicators Research*, In Press, pp. 1-117, 2015.

[13] M-L Marsal-Llacuna, "City Indicators on Social Sustainability as Standardization Technologies for Smarter (Citizen-Centered) Governance of Cities," *Social Indicators Research*, In Press, 2015, DOI 10.1007/s11205-015-1075-6.

[14] C.S. Lee, G.M. Lee and W.S. Rhee, "Standardization and challenges of smart ubiquitous networks in ITU-T," *IEEE Communications Magazine*, vol. 51 no. 10, pp. 102-110, 2013.
[15] NIST, "IoT – Enabled Smart City Framework," 2016, retrieved July 2016 from https://s3.amazonaws.com/nist-sgcps/smartcityframework/files/IoT-EnabledSmartCityFrameworkWP.pdf
[16] NIST, "SMART CITIES/CPS AT NIST," 2014, retrieved July 2016 from http://www.nist.gov/public_affairs/releases/upload/smartcities_cps_budgetsheet.pdf
[17] CEN-CENELEC-ETSI, "Smart and sustainable cities and communities: A role for European standardization," 2016.
[18] BSI, "PAS 181 Smart City Framework Standard," 2014.
[19] T. Watteyne and K.S.J. Pister, "Smarter cities through standards-based wireless sensor networks," *IBM Journal of Research and Development*, vol. 55 no. 1-2, pp. 1-10, 2011.
[20] J. Villalba-Diez and J. Ordieres-Mere, "Improving Manufacturing Performance by Standardization of Interprocess Communication", *IEEE Transactions On Engineering Management*, vol. 62, no. 3, pp. 351-360, 2015.
[21] E.W.T. Ngai, D.C.K. Chau, J.K.L. Poon, and C.K.M. To "Energy and utility management maturity model for sustainable manufacturing process," International Journal of Production Economics, vol. 146, no. 2, pp. 453-464, 2013.
[22] S. Herrmann, H. Rogers, M Gebhard and E. Hartmann, "Co-creating value in the automotive supply chain: an RFID application for processing finished vehicles," *Production Planning & Control*, vol. 26, no. 12, pp. 981-993, 2015.
[23] T-H Chou, C-C Lee, and C-W Lin, "Using ontological e-services framework to orchestrate e-business process services," *International Journal of Enterprise Information Systems*, vol. 11, no. 2, pp. 65-89, 2015.
[24] R. Hahn, "ISO 26000 and the Standardization of Strategic Management Processes for Sustainability and Corporate Social Responsibility," Business Strategy and the Environment, vol. 22, pp. 442-455, 2013.
[25] T. Egyedi and S., Muto, "Standards for ICT: A green strategy in a grey sector," Source of the DocumentInternational Journal of IT Standards and Standardization Research, vol. 10, no. 1, pp. 34-47, 2012.
[26] P. Teachout, "Process standardization," *Flexo*, vol. 34, no.3, pp. 28-32, 2009.
[27] Reyneri, C. "Operational building blocks for business process modelling," *Computers in Industry*, vol. 40, pp. 115-123, 1999.
[28] J. Young, C. Rinner and D. Patychuk, "The Effect of Standardization in Multicriteria Decision Analysis on Health Policy Outcomes," *Advances in Intelligent Decision Technologies*, vol. 4 of the series Smart Innovation, Systems and Technologies, pp 299-307, 2010.
[29] J.P. Russel, "How to implement and audit the process approach," *Quality Progress*, vol. 35, no. 12, pp. 71-73, 2003.
[30] J. Yu-Chih Liu, V. J. Chen, C-L Chan and T. Lie "The impact of software process standardization on software flexibility and project management performance: Control theory perspective," *Information and Software Technology*, vol. 50, pp. 889–896, 2008.
[31] L. Anthopoulos, M. Janssen and V. Weerakkody, "Smart Service Portfolios: do the Cities Follow Standards?," in the Proceedings of the WWW2016 ACM 25th World Wide Web International Conference, Montreal, Canada 2016, DOI: 10.1145/2872518.2888618
[32] "Documentation for the TIMES model, Part I", April 2005, available at: http://www.iea-etsap.org/web/Docs/TIMESDoc-Intro.pdf
[33] J.P. Brans and P.H. Vincke, "A preference ranking organization method: the PROMETHEE method for multiple-criteria decision-making," *Management Science*, vol. 31, no. 6, pp. 647–656, 1985.

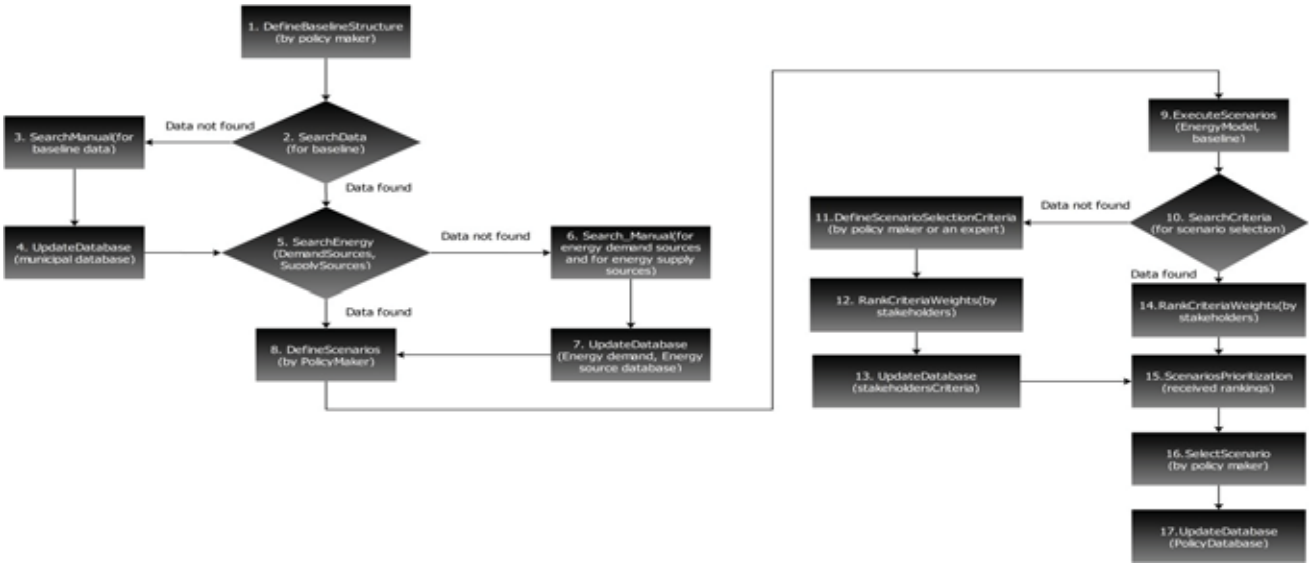


Figure 3. The policy making process model

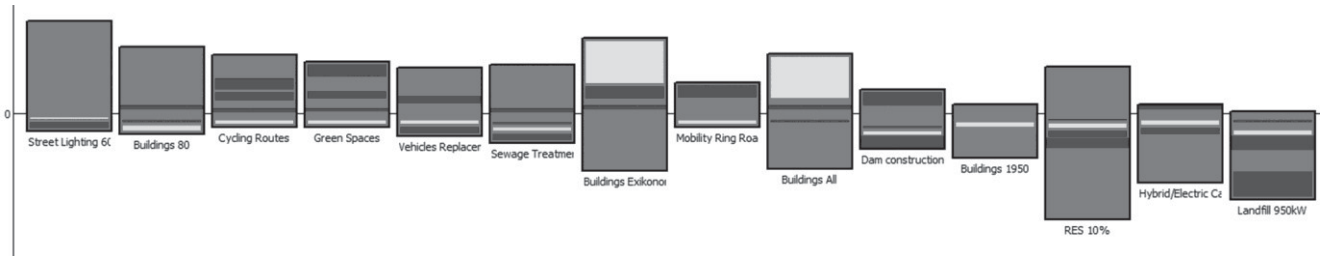


Figure. 2: Scenarios prioritization according to the PROMETHEE MCDM in Trikala city

Table 6: The analysis of the proposed TBM process

Task Code	Action	What	WhatElse	Description of action
1.	DefineBaselineStructure	DefinedBaselineStructure	Policy maker or an expert	Policy maker or an expert defines the baseline structure
2.	SearchDatabase	6. BaselineDataExist = yes 4. BaselineDataExist = no	BaselineDatabase – a database name	Select a route based on whether or not baseline information is found
3.	Search_Manual	City, Population, Energy Suppliers, Energy Consumers, Traffic Data, Emission Production, Water and Sewage, Waste Management	Policy maker or an expert	Policy maker investigates the city for the required information for the baseline
4.	UpdateDatabase	ID#|City |BaselineParameter |BaselineValue	BaselineDatabase – a database name	Automated task: Update the database with the Baseline data
5.	SearchDatabase (5.1, 5.2)	9: SuppliersExist = yes AND ConsumersExist = yes 7: SuppliersExist = no OR ConsumersExist = no	EnergySource – a database name EnergyDemand – a database name	Select a route based on whether or not energy suppliers and consumers are found
6.	Search_Manual	Energy Demand, Energy Production Source	Nil	Policy maker or an assigned expert identifies a set of energy suppliers (sources) and energy consumers (demands)
7.	UpdateDatabase (7.1, 7.2)	ID#|EnergySource ID#|EnergyDemand	EnergySource – a database name EnergyDemand – a database name	Automated tasks: Update the first database with the identified energy supplier sources and the second database with the identified energy demands
8.	DefineScenarios	DefinedScenarios	Policy maker or an expert	Policy maker or an assigned expert defines a set of scenarios (policy options)
9.	ExecuteScenarios	EnergyModel, baseline	Nil	Automated task: the identified scenarios are tested regarding their efficiency with the identified model
10.	SearchCriteria	15: CriteriaExist = yes 12: CriteriaExist = no	Nil	Select a route based on whether or not selection criteria are found
11.	DefineScenarioSelection Criteria	DefinedCriteria	Policy maker or an expert	Policy maker defines a set of criteria, for scenario selection
12.	RankCriteriaWeights	RankedCriteria	Stakeholders	The stakeholders rank the MCDM criteria
13.	UpdateDatabase	ID#|Criterion| CriterionWeight	StakeholdersCriteria – a database name	Automated task: Update the selected criteria database
14.	RankCriteriaWeights	RankedCriteria	Stakeholders	The stakeholders rank the MCDM criteria
15.	ScenariosPrioritization	RankedCriteria	Nil	Automated task: scenarios are prioritized according to criteria weights
16.	SelectScenario	Top ranked Scenario	Policy maker	The policy maker makes decision based on the optimal scenario according to MCDM results
17.	UpdateDatabase	ID#|Scenario|Rank	PolicyDatabase – a database name	Automated task: Update the selected scenario database

CLEANWIFI: THE WIRELESS NETWORK FOR AIR QUALITY MONITORING, COMMUNITY INTERNET ACCESS AND ENVIRONMENTAL EDUCATION IN SMART CITIES

Gómez Ruíz Carlos Andrés

Universitaria Agustiniana. carlos.gomezr@uniagustiniana.edu.co

ABSTRACT

This work presents a new development model for wireless community networks, framed within the sustainable development goals, SDG 11, to achieve the transformation of cities to environmentally sustainable areas and to take measures to fight climate change; this model is based on the use of renewable energies, mesh routing protocols, the monitoring of air quality and environmental variables, IoT, and the application of educational methodologies in order to reward less polluted areas. There are proposed some ideas about standardization of building networks for measuring air quality in smart cities, which provide a great source of information for location-based services (LBS) that promote environmental awareness services.

The CleanWiFi network constantly monitors the air for pollutant gases, uses that information to feed a Big Sensor Data system, and uses the same data for the automatic configuration of the public WiFi service, displaying information about the quality of the air to the user, and rewarding less oolluted areas with a better service. That way it raises public awareness about state of air pollution and how important it is to reduce it; promotes the use of renewable energies and brings WiFi connectivity to the people.

Keywords— WiFi node, air quality, sensor network, public LED lighting, big data, IoT.

1. INTRODUCTION

The development of WiFi networks has been, in the last few years, one of the most impactful aspects within the urban, industrial and academic environments; the great number of existing wireless networks have become a problem for the environment, because there are more electronic devices consuming more electricity. However, the increasing need for connectivity makes it impossible to bring down the existing wireless networks. Not many people have developed an interest in bringing together the need for wireless connectivity with measures to ensuring environmental protection.

This paper presents a viable option to combine the development of wireless networks, environmental monitoring, use of renewable energy and to promote the culture of environmental protection.

As of the writing of this paper, system modules have been developed and implemented, and site surveys with the system completely integrated will follow shortly after.

2. STATE OF THE ART

The sensor networks may include seismic, thermal, visual, infrared and acoustic sensors with the capability to monitoring the environmental conditions. The sensor nodes can be used to perform continuous measuring, event detection, localization, and local actuators control. Furthermore, with the idea of microsensing and wireless networking, there are more possibilities for implementations. [1]

For a sensor network, these sensor nodes are the building blocks that help to assemble it. The IEEE created a standard for the development of wireless networks aimed to the exclusive use of sensor networks, the standard IEEE 802.15.4, that provides a new protocol for sensor communication, designed for low powered devices, for personal area networks PAN [2], a technology known as Zigbee. However, the implementation of the Zigbee standard, results in the installation of a dedicated and exclusive hardware. Given the great popularity of WiFi networks, some works showing the implementation of these sensor networks are beginning to appear.

In the year 2014, Nyugen [3] proposed the creation of small wireless sensor nodes to be used for Wireless Embedded Internet applications. Nyugen proposed the use of 802.11 WiFi networks on low power mode, instead of the traditional 802.15.4, to build measurement micronodes.

The combination of sensors, embedded computing systems and the access to data networks is what makes possible the technology behind de Internet of Things IoT.

2.1. Air quality measurement.

The measurement of personal exposure to air pollution has traditionally been directly and indirectly, where the direct method involves personal sampling of air, and the indirect method involves measurements of the environment, mathematical models and surveys. [4]

There are two approaches on how to make real-time measurements on air pollution: individual based and place based. The individual measurement focuses on the study exposing a person to pollution, identifying the concentration of pollution and the time of exposure over a period of time, yielding valuable data to produce conclusions on the possible connection of pollution air diseases. This approach can rely on location technologies such as GPS, to create information of air pollution through people's paths.

Place-based real-time assessment methods investigate pollution by subdividing the area into simple objects or homogeneous as neighborhoods, schools, classrooms, etc., allowing easier handling of information. It has the disadvantage that it is assumed that air pollution is homogeneous in space.[4]

2.2. Public hotspots

One the most implemented techniques for public WiFi areas is the public hotspot, which presents a web interface with terms and conditions to access the WiFi, which the user has to agree with to be able to navigate through the internet, and also to restrict access for those users that have the possibility to buy a ticket or have a contract with a company. One of the first captive portal solutions was the NoCatAuth, a software that could run in small computers because of its capability to store usernames and passwords in a local database, this software is no longer in use. WiFiDog is another open-source software that can run in small computers and embedded systems [5].

Given the proliferation of open source firmware based routers, captive portal solutions like NoDogSplash have become famous for their capability to offer basic information within their web interface, without having to access external resources.[6]

Chillispot is an open source software solution that allows the use of captive portals located on cloud web servers compatible with the RADIUS protocol, which are used for centralized solutions like dynamic captive portals, one of which is being developed by the Colombian company DataWiFi, with the capability to get demographic information, like people's needs and interests, through the continuous use of WiFi hotspots. [7]

2.3. Big sensor data

Nowadays, sensor networks have become the biggest data generators in the world, 1,250 billion gigabytes worth of data was generated in 2010. Furthermore, this data is taken in real time, and can contain a big list of variables [8]. Wu [8] developed a cloud-based platform with the capability to collect data generated by sensors from anywhere, enabling quick data analysis and storage. These kind of centralized data analysis solutions offer valuable tools to those smarter cities that, for instance, understand the real impact inflicted by vehicular pollution.

2.4. Renewable energies for public use.

A few companies have started to implement the use of renewable energies, like solar energy, smartphone charging areas and resting places, as it is the case for the company Sologic - Renewable Energy Systems and its product eTree[9]. Though solutions like these are innovative and promote the use of renewable energies, they are glamorous and expensive, as shown in Figure 1.

Figure 1. An eTree installation. [9]

3. DESIGN

CleanWiFi is a multi-purpose community wireless network. CleanWiFi has a small and economic design that is easy to replicate. It is composed by wireless routers interconnected to form a self-configuring Mesh network where the information travels through the nodes and to the users of CleanWiFi.

Each node has a data acquisition system based on a low powered and small embedded system with sensors including the capability to determine the air quality by measuring the concentration of gases like methane, isobutene, ethanol, hydrogen, nitrous dioxide (NO2) and carbon monoxide (CO).

Each node has a solar panel, a light sensor, a rechargeable battery and a LED lamp that turns on when the natural lighting is low. Also, each node has a public WiFi area to provide free access to the internet.

CleanWiFi combines its sensor network, its mesh interconnection and its public WiFi network to offer an environmental information and education system. The WiFi network has a captive portal with a splash page showing information about the air quality in real time and a Round Robin statistical graphic representing the air quality on that node during the last day, the last week and the last month. It can also provide information about the air quality in the area or in the city thanks to its centralized platform for data capture and analysis.

The splash page within the captive portal is smart, and is able to show alerts for high levels of air pollution to raise awareness in order for the people to take steps to solve the problem.

Lastly, the captive portal configures itself to reward those areas with lower levels of air pollution. Therefore, if an area has low levels of polluting gases, the users will enjoy a higher bandwidth and more time to navigate the internet. If the environmental pollution increases, the captive portal will reduce the bandwidth and the time available for the users to access the internet. If the pollution levels exceeds historic highs, the captive portal will show alerts to the users and restrict their free access to a very low bandwidth until pollution levels decrease. Figure 2 shows a CleanWiFi node block diagram.

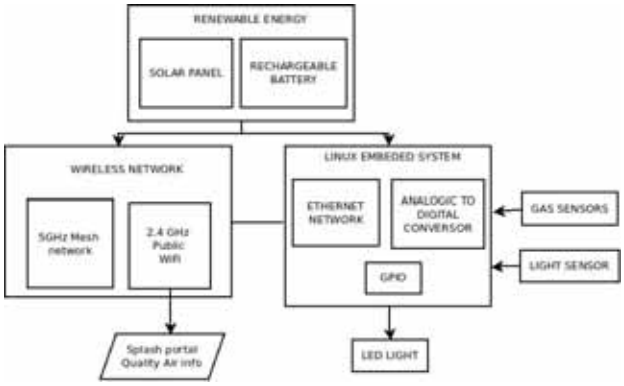


Figure 2. Design of the CleanWiFi network

The supply system provides electrical energy to all the modules from its solar cells and its rechargeable battery. It contains two logic blocks, one is the network unit responsible for establishing the mesh interconnections, and the other block deals with the measuring of air quality variables and LED light control.

3.1. Air quality measurement.

This approach is based on place-based real-time assessment methods in order to characterize air pollution at fixed sites and make continuous measurements to create a reliable source of information for future surveys.

The main component on the node is its air quality measurement system, which provides the user with information, and feeds the Big Sensor Data server and the hotspot system in order to establish internet access rules in concordance with the air quality on that node.

The measurement system contains the following components:

- Air pollution sensor TGS 2600. It is capable of measuring the concentration of gases like methane, carbon monoxide, ethane, isobutene and hydrogen, but it just provides a general measurement, with no distinction between gases.

- MiCS-2714 sensor. It measures the concentration of nitrous dioxide (NO2), one of the main polluters within the nitrous oxides.

- MiCS-5134 sensor. It measures the concentration of carbon monoxide (CO) a dangerous, odorless and tasteless gas.

It is possible, with these sensors, to know how is the air quality and the concentration level of polluting gases.

The Big Sensor Data server is an open source GNU/Linux operating system with a database engine, a web server, and software tools to generate graphical statistics. It allows to store a large amount of data, and to manage, not only individual nodes, but also a group of nearby nodes to generate air quality reports from a range of areas.

3.2. Wireless community network.

Figure 3 shows the operating diagram of the CleanWiFi wireless network, representing three interconnected nodes and the data flow between the sensors and the user navigation.

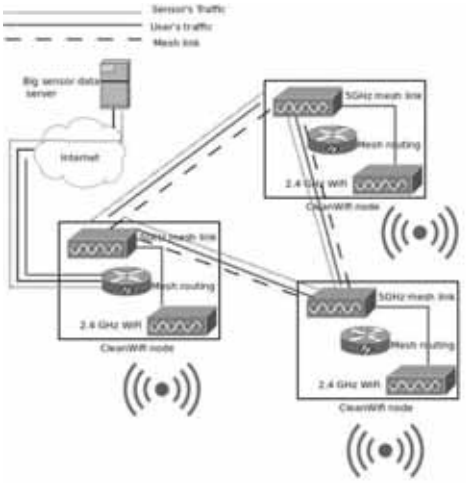


Figure 3. Network diagram of three CleanWiFi nodes

Each node has a wireless router including the GNU/Linux based operating system OpenWRT. The router has two network interfaces, one of them works with the 5 GHz band and it's in charge of the interconnection with nearby nodes.

To operate the mesh network, the routing protocol B.A.T.M.A.N-ADV (better approach to mobile ad-hoc networking -Advanced) has been implemented. This dynamic routing protocol works within the layer 2 of the OSI model by emulating a virtual switch on each participating node, and encapsulating all the traffic inside Ethernet frames [10]. Thanks to this implementation, all the nodes have their own IPv6 addresses which can be quickly and dynamically routed, achieving the continuous operation of the network.

User traffic is encapsulated and sent to any node that has an internet connection, and sensor traffic is encapsulated and sent to the Big Sensor Data server, which can be installed locally into the wireless network or into a centralized server within the internet cloud.

Each router has the software NoDogSplash embedded into their systems, which is used to show a custom captive portal as an authentication method to access the internet via WiFi network.

The splash page works dynamically as it has an algorithm that reads recent air quality measurements and shows the related information to the user. NoDogSplash also allows to set some QoS (Quality of Service) variables in order to control de bandwidth available to the user. The algorithm establishes bandwidth rates that can be used in concordance to the air pollution levels, rewarding those areas that manage to reduce pollution and penalize those with high pollution levels. Figure 4 shows the algorithm logic.

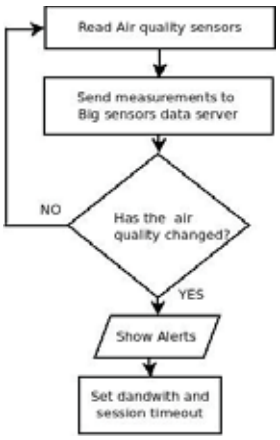


Figure 4. Flowchart of the captive portal control algorithm

Figure 5. A splash page displaying a Nitrous Dioxide contamination alert.

Figure 5 shows a regular splash page as seen from a smartphone.

3.3. Lighting and supply systems.

The electrical supply system on each node is equipped with a solar panel and a rechargeable battery.

The solar panel has the following specifications:

Maximum power: 60W

Open-circuit voltage: 22v

Material: Monocrystalline silicon

The battery has the following specifications:

Type: Li-NMC

Capacity: 168 Wh

Current: 15.6 Ah

Each node also includes a 12 W LED lamp in order to reduce the electrical consumption devoted to lighting the areas with CleanWiFi nodes.

Table 1 shows the electrical consumption from each of the nodes components.

Table 1. Electrical consumption.

Component	Consumption
Router	5.5 W max / 460 mA
Air measurement	1.2 W max / 240 mA
LED lamp	12 W max / 1A

The capacity of the electrical supply system ensures that each node will operate at a full capacity with solar energy and also during night-time with an autonomy of:

Up to 16 hours with the air quality measurement system and wireless communications active, and the LED lighting system inactive.

Up to 6.5 hours with all systems active, including the LED lighting system.

Considering the flexibility of this system, it is possible to add movement sensors to reduce the electrical consumption by deactivating the LED lighting system on the node when there is no people around.

4. STANDARDIZATION OF MEASURING NETWORKS AIR QUALITY FOR PEDESTRIANS

The results of this study will allow contribute to recommendations for building networks to monitor air pollution in order to have a model that can be fed from many points on the planet and can contribute to an information model that collect and analyze large amounts of information in order to create a knowledge base to develop policies and practices according to the objectives of sustainable development.

ITU-T Study Group 20 is currently studying the "Requirements of the network for the Internet of Things". This paper is consistent with the definitions and structures raised to ensure that there is harmonious with future recommendations.

4.1. API - Application Programming Interface

Study Group 20 defines the local network IoT as a network of devices for the IoT and gateways interconnected through local connections. [11]; In draft recommendation it's proposed a first API focused on the physical layer transport

information, however this paper aims to make a first draft standard in structuring data sent by the measurement nodes, at level information analysis.

The data is sent to a Big Sensor Data server through a JSON composite object like this one:

```
{
 "node": {
   “provider”: “0002”,
   "ID": "001122",
  "cont_gas": “50”,
  "no2": "34",
  "co":"0",
  "date": “2016-07-01",
  "hour": "15:23",
 }
}
```

Where "cont_gas", "no2" and "co" represent the reading of the sensors which are expressed in ppm (parts per million).

This simple structure allows collecting information from multiple measurement nodes and its organization using models as big data and business intelligence.

4.2. KPI: Key Performance Indicators

ITU-T Study Group 5 set up the ITU-T Focus Group on Smart Sustainable Cities has defined the KPIs in smart sustainable cities. KPI D2.1 has been defined to measure air quality. [12]

The results of this study, after collecting a lot of information, will create the knowledge base to define the technical specifications of KPI D2.1 for characterizing air quality measurements, which enable smart cities determine their behavior and create actions for improvement.

4.3. Standardizing facilities

As an important part of defining the technical specifications of D2.1 KPI, the results of future work of the network CleanWiFi will pose a model of technical implementation for measuring points of air quality and its interconnection with WiFi access networks. Among other things it aims to define:

- Correct height for the installation of sensors to measure pollution with direct impact on pedestrians.
- Characteristics of the spaces to be monitored.
- Type and minimum required quality of digital connection.
- Physical Installation recommendations.

5. RESULTS

Even though, as of the writing of this paper, complete integration tests are yet to be carried out, there are some individual results:

Air pollution measurements have been acquired, showing the presence of gases like NO2, and storing the data on the Big Sensor Data server to further develop studies and establish hypothesis.

Figure 6 shows the 10-day evolution of the NO2 concentration on one node.

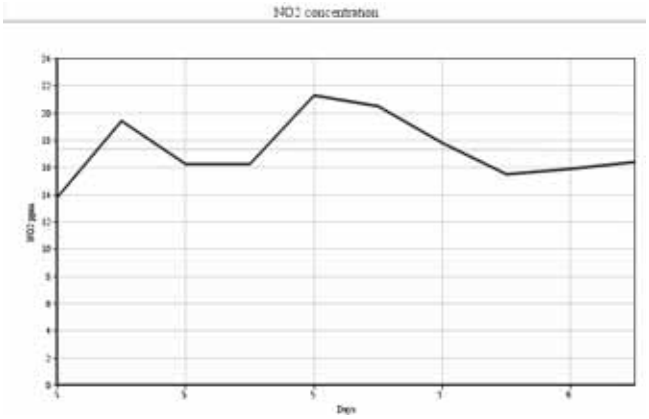


Figure 6. NO2 measurement of a CleanWiFi node. The graphic was elaborated with data stored inside the Big Sensor Data server.

The linking between the measurements and the splash page control software has been achieved to control the quality of service within the WiFi network in concordance with the air quality.

An eye-catching web interface has been designed to send dynamic messages to the user via the splash page, in order to inform them about the air quality and to display alerts with any increase in the pollution levels.

It has been possible to control the variables of upstream and downstream’s bandwidth, as well as the session time of connections to the WiFi network, based on measurements of air quality.

There has been confirmation about the effectiveness of the communication through the mesh network using the B.A.T.M.A.N-ADV protocol between 5 nodes. Achieving an approximate throughput of 30 Mbps between nodes.

The CleanWiFi network has been implemented over a extended area of over 5,000 square meters, where it has been possible to obtain a complete outlook of the air quality.

The implementation of a big sensor data server has been achieved, based on a GNU/Linux operating system and open-source tools, to organize, store and compare measurements in order to analyze the evolution of the air quality in large areas.

6. CONCLUSIONS

The CleanWiFi proposal shows that it is possible to link the evolution of physical variables, specifically the air quality, with technical communication systems, in order to connect all people using public WiFi networks to the monitoring of

the environment and to promote actions to avoid its deterioration.

It is possible to install a model that is easy to replicate, low cost and environmentally-friendly, to create a vast network of measurement instruments to test the air quality and to provide internet access.

The data and sensor network that has been designed works perfectly with renewable energies. That way, the CleanWiFi network contributes to the development of smarter and sustainable cities, involving their citizens in the environmental protection.

With the development of the idea of CleanWiFi, the necessary knowledge bases are created to make recommendations for massive monitoring of air quality, storage and processing of data, and technical specifications of the KPI, as part of the study groups of ITU-T in the areas of Environment and climate change and IoT and applications, smart cities.

REFERENCES

[1] I.F. Akyildiz, W. Su, Y. Sankarasubramaniam, E. Cayirci. 2002. Wireless sensor networks: a survey. Broadband and Wireless Networking Laboratory, School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA

[2] Deshpande P, Madankar M. 2015. Techniques Improving Throughput of Wireless Sensor Network: A Survey. 2015 International Conference on Circuit, Power and Computing Technologies [ICCPCT].

[3] Nguyen M, L Trung. 2014. Low-Power and Cost-Effective Wifi Sensor Motes for Wireless Embedded Internet Application. The 2014 International Conference on Advanced Technologies for Communications (ATC'14)

[4] Tianfang Bernie Fang & Yongmei Lu (2012) Personal real-time air pollution exposure assessment methods promoted by information technological advances, Annals of GIS, 18:4, 279-288, DOI: 10.1080/19475683.2012.727866

[5] Hortelano J. Et al. 2007. A Wireless Mesh Network-based System for Hotspots Deployment and Management. Third International Conference on Networking and Services(ICNS'07). IEEE

[6] NoDogSplash on OpenWRT. https://wiki.openwrt.org/doc/howto/wireless.hotspot.nodogsplash

[7] DataWiFi. Internet con propósito. http://www.datawifi.co

[8] Wu C, et al. Concinnity: A generic Platform for Big Sensor Data Applications. 2014. IEEE Cloud Computing (Volume:1 , Issue: 2)

[9] eTree by Sologic - Renewable Energy Systems. 2016. http://solargiving.com/the-etree/

[10] B.A.T.M.A.N. Advanced https://www.open-mesh.org/projects/batman-adv/wiki/Wiki

[11] ITU. 2016. Draft Recommendation ”Requirements of the network for the Internet of Things” (Y.IoT-network-reqts) - Draft output of Q2/20 meeting, Geneva, 25 July-5 August 2016.

[12] ITU. 2014. Overview of key performance indicators in smart sustainable cities.

CLOUD BASED PATIENT PRIORITIZATION AS SERVICE IN PUBLIC HEALTH CARE

A. Bagula[1], C. Lubamba[1], M. Mandava[1], H. Bagula[2], M. Zennaro[3], E. Pietrosemoli[3]

[1]ISAT Laboratory, Computer Science Department, University of The Western Cape (UWC),
Private Bag X17, Bellville, 7535, South Africa.
Email: abagula@uwc.ac.za
[2]Department of Public Health, School of Medecine, University of Cape Town,
Cape Town, South Africa.
Email: BGLHER002@myuct.ac.za
[3]T/ICT4D Laboratory, The Abdus Salam International Centre for Theoretical Physics,
Via Beirut 7, 34151 Trieste, Italy.
Email: {mzennaro, ermanno}@ictp.it

ABSTRACT

This paper proposes and evaluates the performance of a Cyber-healthcare system which is aimed at providing patient prioritization over the cloud as a public health service for the rural and urban communities of the developing world. The underlying cloud-based Internet-of-Things (IoT-Cloud) infrastructure is aimed to be implemented in the city of Lubumbashi in the republic Democratic Republic of the Congo (DRC) with the objective of setting up a community health care network around a mesh of health kiosks. We propose a deployment model for the proposed Cyber-healthcare system, and describe a patient prioritization process as part of its situation awareness component. The results obtained from an experimental prototype reveal the field readiness of the off-the-shelf bio-sensor technology used by the system, the performance achieved when using a solar powered subsystem, the relative communication capabilities provided by its protocols and the network engineering feasibility of the planned community health care network. The relative efficiency of using supervised machine learning compared to unsupervised machine learning when performing patient prioritization, is also revealed through two popular algorithms: Multivariate linear regression (MLR) and K-means clustering (KMC).

Keywords— E-health, Cyber-healthcare, Internet-of-Things, Patient prioritization, Situation recognition

1. INTRODUCTION

The recent advances in sensor, actuator, radio frequency identification (RFID), wireless/mobile and cloud computing technologies have changed the way information is produced and exchanged between communities. These advances combined with the use of artificial intelligence for mining medical records have revolutionized the way health-care is delivered in the developed world and spun out the Cyber-healthcare field in medicine with positive impacts on patient care. Cyber-healthcare aims at enabling Cyber physicians consultation and treatment of patients via the Internet as well as providing better healthcare management through digitization of all aspects of clinical work: technology, imaging, medications, surgery, rehabilitation, preventive measures, physical therapy, nursing homes, and medical supplies. It provides new opportunities for enhancing health care in the developing world thanks to low acquisition costs and flexible deployment, while improving accuracy by replacing manual operations with fully digitized processes.

The public health sector in both rural and urban settings of the developing world can leverage the Cyber-healthcare technology to improve health care management and service delivery. Leapfrogging from poorly prepared to adequately equipped communities, bio-medical researchers in the developing world can also take advantage of the tools provided by these technologies to conduct research and thus reduce the scientific divide in the medical field. Some of the issues associated with Cyber-healthcare systems deployments include:

Bio-sensor field readiness. While not aimed at replacing the medical practitioner, a Cyber-healthcare system is assumed to offer medical decision support by providing accurate and calibrated vital sign values and build upon these values to implement patient condition recognition and prioritization. The field readiness of the bio-sensor devices used by the system, especially when consisting of low-cost off-the-shelf equipment, is an important parameter upon which the accuracy of the system depends.

Sensor readings dissemination. In many deployment scenarios, the vital signs captured from patients are routed over a network to a processing place where patient's condition recognition is achieved. The efficiency of the bio-sensor readings dissemination is another important parameter upon which service delivery depends. While ZigBee and WiFi have been widely used as legacy protocols in the digital

Thanks to XYZ agency for funding.

healthcare field for the transport of bio-sensor data in the industrial scientific and medical (ISM) frequency bands, many other communication protocols are emerging in other frequency bands. Selecting the most appropriate among these protocols can lead to efficient Cyber-healthcare system designs.

Cyber-healthcare Power Supply. It is well recognized that power supply is one of the main barriers to the ICT's expansion in the developing world and one of the main contributors to the technological divide between developed and developing countries. Wind and solar energy are emerging as alternative solutions to the lack or issue of intermittent power supply in the developing countries. When used to supply power in a healthcare setting, a solar/wind subsystem can become a key component of a Cyber-healthcare system whose supply and demand need to be balanced adequately for efficient service delivery.

Patient condition recognition. Besides using field-ready and calibrated bio-sensor devices, the Cyber-healthcare system is assumed to provide patient condition recognition and medical decision support to both patients and medical practitioners. Both objectives can be reached only through the use of intelligent software systems usually driven by machine learning algorithms. The selection of the type of machine learning algorithms and their designs is an important issue upon which successful patient condition recognition and prioritization depends.

Many other important issues associated with digital health systems include security, privacy, and interoperability when deployed in a cloud-based infrastructure. These issues are beyond the scope of this paper.

1.1. Contribution and Outline

The main goal of this paper is to present and evaluate the performance of a Cyber-healthcare system that combines lightweight cloud computing and Internet-of-Things concepts to achieve patient prioritization also known in the medical field as the Triage system. The underlying cloud-based infrastructure is aimed to be implemented in the city of Lubambashi in the Democratic Republic of the Congo (DRC) with the objective of setting up a community health care network of health kiosks. This paper includes four contributions which are aimed to provide answers to the issues associated with Cyber-healthcare deployments. Firstly, we assess the field readiness of the sensor devices used by the proposed Cyber-healthcare system by benchmarking these sensors against the world health organization (WHO) patient scoring standard. Secondly, we evaluate the performance of the information dissemination model underlying the proposed system to select the best communication protocol for the Cyber-healthcare infrastructure and evaluate the network engineering feasibility of the mesh of health kiosks. Thirdly, we evaluate the energy yield produced by the healthcare infrastructure when powered by a solar energy to extend the lifetime of the healthcare sensor network. Lastly, machine learning techniques are compared to select the most suitable algorithm for the proposed Cyber-healthcare system. The proposed prioritization system expands the work done in [1] to consider a hybrid communication model where both IEEE 802.11 and IEEE 802.15.4 protocols operating in the ISM frequency band are used on different communication links of the Cyber-healthcare system. Furthermore, this current work assess the relevance of using an unsupervised learning algorithm as an alternative to the supervised learning models presented in [1] for patient prioritization process.

The remainder of this paper is organized as follows. Section 2 presents the Cyber-healthcare framework and reveals the main components of the Triage system. The algorithmic solutions to the prioritization problem are presented and discussed in Section 3 while section 4 presents the performance evaluation and section 5 our conclusions.

2. THE CYBER-HEALTHCARE SYSTEM

A digital healthcare system is a platform that should empower people with limited or no medical training to capture and store clinical data into a digitized form, process, analyze this data and share it over the cloud as a service to the public health sector. It should also enable the capture of data in different other forms including crowd sensed data on mobile phones and on-body bio-sensed data. The cloud infrastructure will be equipped with intelligent data analysis algorithms capable of performing situation recognition in terms of patient and process prioritization and decision support using an expert system engine to help the concerned health professionals in the decision making. The medical health workers should periodically take the relevant readings of all the patients that have not been attended to by the doctor. The system should also allow doctors to periodically monitor and access the patient's data remotely from their smart devices; tablets and smart phones with no time delay constraint. The information collected by the system should also be shared by health care planners for evaluation and planning and biomedical researchers to achieve predictive patient analytics, while abiding to security and personal data privacy protection schemes. When deployed as an interconnected sets of medical databases, Cyber-healthcare systems provide an unprecedented opportunity to advance the discovery and treatment of new diseases and a better understanding of how the human body works[2]. Such advances are boosted by the use of cloud computing technologies [3, 4, 5], [6] to provide three service models to patients and the medical communities. These include (a) Software as a service (SaaS) by providing software applications which are hosted by the cloud and made available to users over a network (b) Platform as a service (PaaS) where development tools such as operating systems hosted in the cloud are accessed through a browser and (c) Infrastructure as a service (IaaS) where the resources of the cloud including storage, hardware, servers and network components are outsourced to users. While the three models may be equally critical in the developing world settings, this paper's focus lies more on the SaaS model to enable sharing a patient prioritization application by a number

of users through a mesh network of health kiosks.

2.1. A Mesh of Health Kiosks

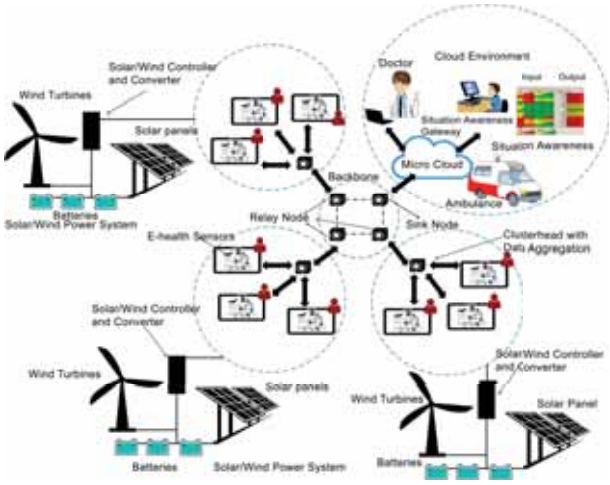


Figure 1: The Cyber Healthcare Model

A Cyber-healthcare deployment model is depicted by Figure 1. It is built around the idea of a mesh of health kiosks infrastructure where 1) the bio-sensors of an e-Health kit are used in a health kisok to capture patient vital signs 2) they are stored in the nodes of the sensor network and aggregated at cluster heads and 3) they are relayed to a sink node connected to a micro cloud where data analytics are performed to achieve situation awareness (condition recognition and patient prioritization) and where the information is shared among a number of organizations or entities. The objective is to provide doctors, institutions, emergency workers, public health planners and bio-medical researchers access to an integrated health care system for planning, utilization and advanced research, while preserving the privacy of individual's data. Some of the advantages of such a deployment when applied to the developing world context include:

- System accuracy. A medical bio-sensors e-Health kit provides a way of replacing the error-prone manual patient vital signs capture process by more accurate automated procedures. Furthermore, the use of intelligent software for situation recognition has the advantage of providing accurate approaches used as support to the healthcare providers in order to reduce diagnostic errors.
- Cost saving. When deployed in a hospital or health care centre, a simple e-health kit could be used by many patients or shared by a community to reduce cost. Furthermore, using the cloud can result in a collaborative economic environment where the overhead costs are shared among the participants and the costs for exchanging and sharing patient's data are tremendously reduced.
- Data availability. Using a cloud-based IoT infrastructure allows easy storing and remote access to medical data irrespective of the patient and clinician locations and time-of-the-day to provide to the healthcare practitioners a comprehensive view of the patient's history for better patient care.
- Data analytics. As illustrated by Figure 1, the doctor or ambulance may access the situation awareness gateway via the Internet from anywhere and anytime. In the rural settings with shortage of healthcare professionals and facilities, the data analytics provided over-the-cloud by a situation awareness gateway can be a great service to the public health sector as it can speed up medical diagnosis, treatment and reduce the costs of laboratory tests which usually constitute a large portion of medical expenditures.
- Data sharing. The cloud-based IoT infrastructure is an infrastructure spanning various settings and geographies, shared between authorized participants with the expectation of improving healthcare service delivery and improve operational efficiency. Such an infrastructure can enable participatory consultation, facilitating the interaction of different medical specialists, when required, to improve the medical diagnosis, health care support and many other services that could not be availed to citizens without its presence.
- Real-time updates. Using a cloud-based IoT infrastructure also enables real-time updates of patients' medical history (consultations, prescriptions, hospitalization) which are useful for future treatment validation.

As presented in figure 1, the Cyber-healthcare relies on a networked digital health infrastructure where a) the bio-sensor devices are equipped with different sensors aimed at capturing different body vital signs b) communication between nodes of the network is achieved indoor or outdoor depending on the localization of the vital signs capturing modules c) the routing of the sensor readings over these links is achieved by different protocols including WiFi and the 802.15.4 protocols d) the micro-cloud infrastructure is equipped with a patient prioritization server and can be a component of a federated cloud infrastructure shared by several hospitals in rural settings of the developing world and e) owing to the lack or intermittent energy supply in the developing world, the system proposed in this paper is endowed with a wind/solar power source.

2.2. The Patient Prioritization System

As illustrated by Figure 2, the patient prioritization system proposed in this paper has four main components: a) a database storage used to store the medical record history of the patients, time stamped patient physiological parameter readings from the bio-medical sensors and also the Triage results (or scores) for every patient record b) a scoring system which builds from the WHO standardized table of vital parameter risk zones [1] to acquire domain knowledge in

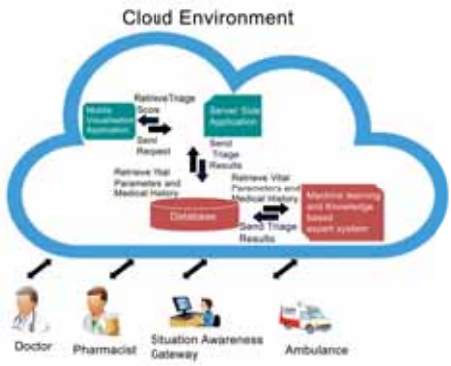


Figure 2: The Patient Prioritization System.

order to analyze the patient data stored in the cloud database storage c) a mobile visualization application applying rule based data analysis in order to perform situation awareness and d) a server application interfacing with the cloud storage and the Mobile Visualization application. The triage scoring system adopted by our system was described in [1]. It is based on a model that assigns scores to vital signs used as Triage parameters in order to quantify their severity level. As depicted by figure 2, upon determination of the quantitative measure of a patient medical condition, the feedback from the patient prioritization software can be provided to the users (patients and/or doctors) through the mobile visualization application revealing the condition of the patient in green, red or orange statuses and different actions to be taken: do nothing if green status, plan a visit to the doctor if orange status and see the doctor immediately if red status.

2.3. Information Dissemination and Power Supply

Different frequency bands have been recently recommended by the 802.15.6 standard to mitigate the interference in the ISM band resulting from co-location of bio-sensors with other devices. While interference is very significant in the urban settings of the developing world, the less crowded ISM band of the rural settings of the developing world can tolerate less complex mitigation schemes used to overcome wireless interference. The focus of our work will therefore lie on the IEEE802.11 and IEEE802.15.4 communication standards when operating in the ISM band, as they are commonly available and have a built in interference mitigation mechanism. The power supply model adopted by a Cyber-healthcare system is as important as its communication model since both have an impact on its efficiency and the lifetime of the underlying communication network infrastructure. The Cyber-healthcare power supply system proposed in this paper is based on photovoltaic energy. It consists of a solar powered sensor subsystem depicted in Figure 3 where a photovoltaic panel converts light into variable DC power, a charge controller is used to charge the battery and to supply the proper voltage to the load. The same scheme can be applied to provide large amounts of electricity, but in the case of a sensor board, the charge controller need not to be a separated device, as in the system depicted which consists of a 3W solar panel that provides 5V to the Arduino board and to the 2300 mAh Li-Ion battery.

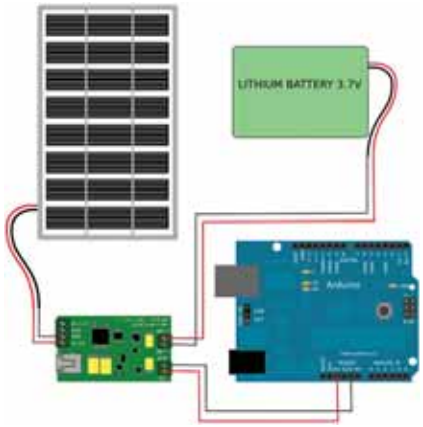


Figure 3: Solar Powered Sensor Subsystem [8].

3. MACHINE LEARNING ALGORITHMS

The aim of a patient prioritization system also called Triage system is to build upon the WHO scoring system and machine learning techniques borrowed from the field of artificial intelligence to determine a quantitative measure of patient's medical conditions (condition recognition) and then give priorities to the most urgent cases. The Triage algorithm should be easily interpretable, fast to compute, simple, accurate and scalable in order to be portable to small devices, e.g. smart phones, tablets, iPad, and/or work with biomedical sensors and smart watches without any problems. Two machine learning algorithms were selected to solve the patient prioritization problem. Their basic characteristics are described below and their performance compared in section 4.

3.1. Multivariate Linear Regression

Figure 4: Multivariate Linear Regression.

The Matrix Algebra method (MAM) has been often used as an alternative to the the multivariate linear regression (MLR) algorithm. However, the MLR by gradient descent was used in our work because MAM tends to run slower with the increase of the amount of data.

Algorithm description. As illustrated by Figure 4, this algorithm uses the knowledge based system to score the training data before training. In other words, it is an improved

expert system knowledge based algorithm which learns from the data, calculates the weights for each variable or generates a linear hypothesis which it uses to score the vital parameters.

3.2. K-means Clustering Algorithm

The K-means clustering (KMC) algorithm considered in this paper partitions n observations into k sets ($k < n$) $S = \{S_1, S_2, \ldots, S_k\}$ so as to minimize the within-cluster sum of errors squares (WCSS), which is expressed by:

$$arg_s \min \sum_{i=1}^{k} \sum_{x_j \in S_i} || \, x_j - \mu_i \, ||^2$$

where μ_i is the mean of points in S_i.

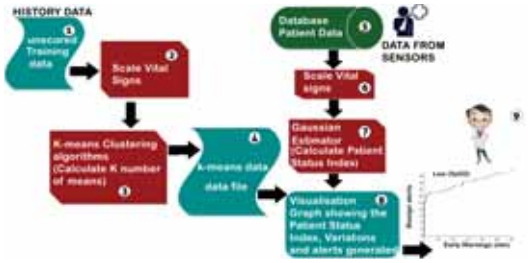


Figure 5: K-means clustering

Algorithm description. As described by Figure 5, the KMC algorithm considered in this paper partition the data into clusters and uses the Gaussian estimator (Parzen window estimator) algorithm to estimate a probability density function p(x) correspondent to a particular status which is then used to calculate a patient status index (PSI) by means of the equation $PSI = log_e[1/p(x)]$.

4. PERFORMANCE EVALUATION

The Cyber-healthcare system used in our work leverages the off-the-shelf e-Health kit from Libelium as a low cost device that can be easily and quickly deployed in a rural environment. The development environment used in our experiments included: Ubuntu 13.10, Android SDK, Apache2 server (on Ubuntu), MySQL server, MySQL -Java-Bridge (MySQL-J-Connector). Different languages were used during our development, including Java, PHP, SQL, JSON, and XML.

4.1. Sensor Field Readiness

We conducted a first set of experiments to evaluate the field readiness of the off-the-shelf e-health sensor technology with the objective of making sure that the sensor readings fall in acceptable ranges. To overcome the lengthy ethical clearance procedures aimed at protecting patient privacy through confidentiality, we used in this experiment two healthy users whose vital signs were monitored for four days. The results presented in tables 1 confirmed a normal healthy state for both individuals with 1) normal vital signs indication according to the WHO norms during the four days and 2) very little daily variations since the users did not fall sick during the experimentation. These values were calibrated against those obtained from medical equipment used by nurses in hospitals and bench-marked against the WHO values[1]. They revealed similar values and performance patterns for non-sick individuals.

4.2. Information Dissemination

We conducted a set of experiments using a worst case deployment scenario for a rural hospital represented by an overcrowed building complex in Cape Town with many tenants using WiFi devices (laptops, tablets and phones) to access the Internet and communicate through social media. The experimental results are depicted by Figures 6 and 7 for the RSSI and throughput respectively. The signal strength at the receiver's side for the IEEE 802.11 communication is constant over different distances while the signal strength in the 802.15.4 links decreases with distance. This is in line with the fact that the 802.11 protocol has been designed for longer communication ranges than the 802.15.4 protocol. Furthermore, the IEEE 802.11, basis of WiFi, reveals similar results for both indoor and outdoor communication. In contrast, the 802.15.4 shows a difference between indoor and outdoor scenarios where outdoor links reach longer ranges than indoor links and the indoor RSSI strength reduces with the number of walls. To measure the throughput achieved over indoor and outdoor communication links, we transmitted a number of packets configured to contain the bio-sensor readings as payload and measured the ratio of the number of packet successfully received and acknowledged to the number of packets sent. Figure 7 reveals a performance pattern similar to the received signal strength indicator (RSSI) where the throughput achieved indoor and outdoor are the same for the 802.11 protocol. The 802.15.4 reveals different performance patterns between indoor and outdoor communication and for the indoor communication with different number of walls. Thanks to the high RSSI, a constant percentage of packets was received for the 802.11 protocol while the indoor 802.15.4 links achieved a higher throughput for one wall compared to two walls. The 802.11 protocol achieved higher throughput compared to the 802.15.4 in both indoor and outdoor communications. Note that although the 802.11 protocol outperformed the the 802.15.4 on both performance parameters, the 802.15.4 deployment is a cheaper option and more frugal option in terms of energy consumption even when using the lightweight version of the IEEE802.11 protocol often referred to as WiFi-lite.

4.3. Health Kiosks Mesh Network Engineering

The information dissemination experiments presented above were complemented by a network engineering study of the feasibility of a 45 nodes community health care network in Lubumbashi as described earlier. The results produced using

Table 1: Sensor Field Readiness: Subject One

Subject one	Day 1	Day 2	Day 3	Day 4
Systolic blood pressure	Max: 120mmHg	Max: 137mmHg	Max: 127mmHg	Max: 131mmHg
	Min: 111 mmHg	Min: 117 mmHg	Min: 121 mmHg	Min: 127 mmHg
Diastolic blood pressure	Max: 93mmHg	Max: 86mmHg	Max: 91mmHg	Max: 78mmHg
	Min: 77mmHg	Min: 76mmHg	Min: 83mmHg	Min: 69mmHg
Pulse	Max: 66 bpm	Max: 86 bpm	Max: 71 bpm	Max: 70 bpm
	Min: 61 bpm	Min: 61 bpm	Min: 65 bpm	Min: 66 bpm
SPO2	Max: 99%	Max: 90%	Max: 95%	Max: 99%
	Min: 98 %	Min: 88 %	Min: 89 %	Min: 93 %
Temp (^oC)	Max:37.00	Max:36.94	Max:36.98	Max:36.99
	Min:36.98	Min:36.64	Min:36.90	Min:36.81
Subject two	Day 1	Day 2	Day 3	Day 4
Systolic blood pressure	Max: 134mmHg	Max: 131mmHg	Max: 134mmHg	Max: 130mmHg
	Min: 132 mmHg	Min: 129 mmHg	Min: 131 mmHg	Min: 126 mmHg
Diastolic blood pressure	Max: 79mmHg	Max: 82mmHg	Max: 81mmHg	Max: 83mmHg
	Min: 78mmHg	Min: 76mmHg	Min: 75mmHg	Min: 63mmHg
Pulse	Max: 88 bpm	Max: 79 bpm	Max: 73 bpm	Max: 74 bpm
	Min: 82 bpm	Min: 67 bpm	Min: 67 bpm	Min: 66 bpm
SPO2	Max: 99%	Max: 93%	Max: 97%	Max: 95%
	Min: 95 %	Min: 89 %	Min: 95 %	Min: 93 %
Temp (^oC)	Max:36.87	Max:37.06	Max:36.98	Max:36.99
	Min:36.53	Min:36.64	Min:36.92	Min:36.91

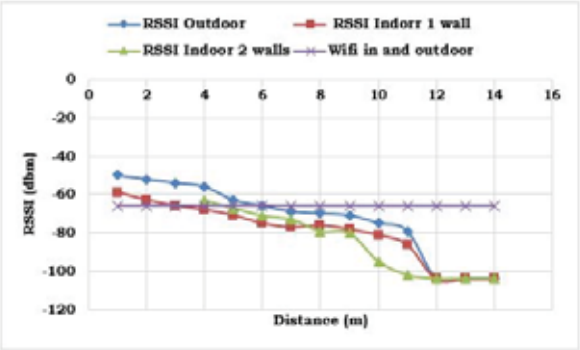


Figure 6: RSSI of WiFi and 802.15.4 Vs distance.

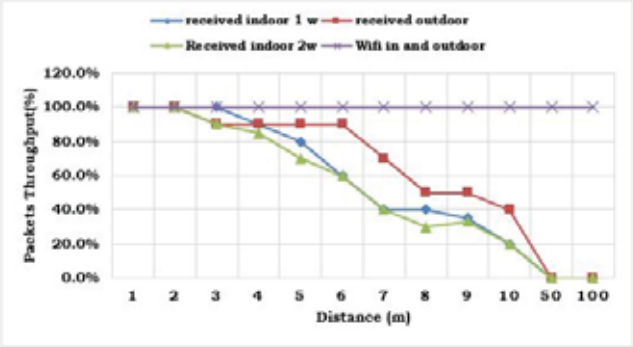


Figure 7: Packets throughput of WiFi and 802.15.4 Vs distance.

Radio mobile revealed a fully mesh network exempted of any obstructions shown by Figure 8, with the best link reflecting a link margin of 49 dB as depicted by Figure 9. When compared to a similar 45 nodes community health care network for the city of Cape Town, it was found that a mesh network was not feasible when using the 2.4Ghz band and the best link margin for such network was of only 15.4 dB, a third of the performance achieved by the best link in the Lubumbashi network. This is in agreement with our initial assumption of using WiFI in the proposed community healthcare infrastructure for settings where the ISM band is lightly crowded. Note that while both are cities of the developing world, Cape Town is a city with a better ICT infrastructure and a wireless/mobile penetration much higher than Lubumbashi.

4.4. Solar Power System

The Cyber-healthcare system presented in this paper was designed to be deployed in regions of the developing countries with no or intermittent power supply. The solar power was therefore selected as an alternative source of energy for powering both the sensor nodes used for vital signs collection and the micro-cloud server used to host the situation recognition gateway/server. Starting with a fully charged battery attached to an arduino mote, we conducted a set of four experiments to evaluate the performance of the solar powered sensor subsystem. Two types of experiments were conducted, one with the sensor mote running only on its lithium rechargeable battery, and the other with the sensor mote running on battery but with solar panel connected. For each experiment, two scenarios were considered, one with the 802.15.4/ZigBee protocol and another one with the 802.11/WiFi protocol. The experiments consisted of sensor devices communicating with the gateway to exchange packets with the following information about the sender: MAC address, On body temperature, Sender remaining battery power (%), Sender node Name, Sender node ID and a “hello” string. In both cases, the sensor mote was left running, broadcasting and receiving broadcast messages while writing the information received on a file. The results shown in figure 11 revealed that during a day, using solar panels could result in huge energy harvesting. The resulting energy savings could be considerably increased through implementation of Sleeping mode (turning off the device when not active) to further increase the lifetime of the cloud-based Cyber-healthcare system. This has been reserved for future work.

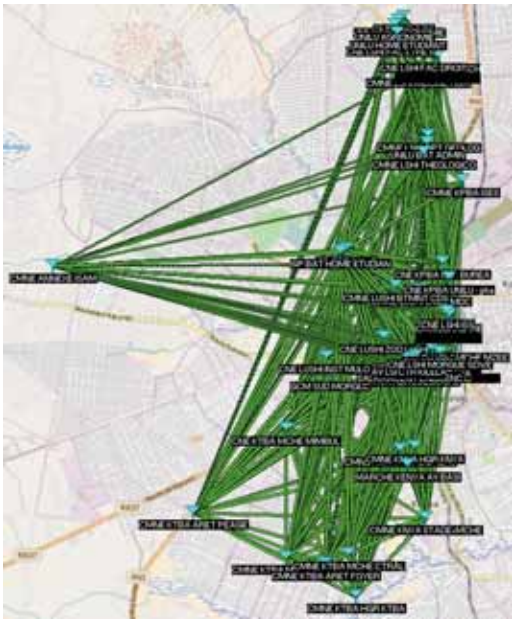


Figure 8: Cyber-Healthcare Mesh Network.

Figure 9: Best Fresnel Zone in Lubumbashi.

Figure 10: Best Fresnel Zone in Cape Town.

4.5. Patient Condition Recognition

We conducted another set of experiments to compare the two machine learning algorithms in order to select one that will be deployed as algorithm of choice for our Cyber-healthcare system. Four different performance parameters were used to compare the algorithms: Coefficient of determination, Accuracy, Runtime and the Time Complexity. The Analysis of Variance (ANOVA) method was used to evaluate the models in this paper. The most important parameter in this method is the Coefficient of determination, denoted R^2 or r^2. It indicates how well data fit a statistical model. This value ranges from 0 to 1; the value one means the data fits the model perfectly. A value less than 0.5 indicates that the data do not fit the model.

To avoid using healthy users as in our previous experiment, we selected for this experimentation a real patients' dataset found from an MIT website (http://www.physio.net). This dataset was used and adapted to train and compare the two different machine learning algorithms used in this paper: Multivariate linear Regression and K-means Clustering. The experimental results presented in table 2 reveal that the Multivariate Linear Regression (MLR) algorithm takes approximately 5 seconds to compute the Triage priority score and has a very high accuracy of approximately 90%. The K-means clustering is an unsupervised learning algorithm which is not associated with an accuracy value but has a run time of 14.22 seconds which almost the triple of the MLR algorithm. These results complement those provided in [1] where its was found that the MLR algorithm outperformed the support vector machine (SVM). They also reveal its relative efficiency as best algorithm to be used for patient prioritization in the Cyber-healthcare infrastructure.

5. CONCLUSION AND FUTURE WORK

A Cyber-healthcare system using off-the-shelf equipment for patient prioritization was presented in this paper as a first step towards the implementation of low cost healthcare systems for the developing countries. The off-the-shelf e-Health kit

Table 2: Condition Recognition Results.

Parameters	Multivariate Linear Regression	K-means clustering
Coefficient of determination	0.903	n.a for unsupervised learning
Accuracy (%)	90.30	n.a for unsupervised learning
Runtime (seconds)	5.01	14.22 (for only 10 clusters exponentially grows as the number of clusters increases)
Time Complexity	O(pn+kn) where p is the dimension of each observation (input), k is the number of tasks (dimension of outputs) and n is the number of observations	Big(O) for Kmeans + Big(O) for Parzen Window $O(knT) + O(n^2)$, where k is the number of clusters, ,n is the number of points and T is the number of iterations.
Recal / Detection	0.769231	n.a for unsupervised learning
Precision	0.833333	n.a for unsupervised learning
False Rate	0.6	n.a for unsupervised learning

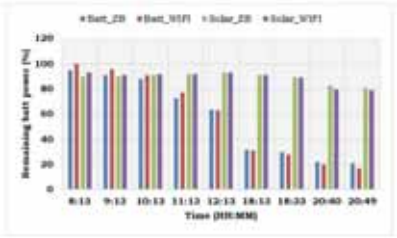


Figure 11: Solar Powered Node Performance.

used in our experimentation was tested and found ready for field deployment. Two machine learning algorithms to solve the patient prioritization problem were described and compared. They revealed the relative efficiency of the supervised Multivariate Linear regression learning algorithm compared to the unsupervised K-Means algorithm.

The situation recognition system presented in this paper has been built on top of a communication platform that considers single hop routing to disseminate the healthcare information from their points of collection to the micro-cloud server that handles the Triage system. When considering a larger network configuration with multi-hop routing paths, multi-path routing techniques such as presented in [9, 10] can be redesigned to support QoS by having different forms of healthcare data propagated over different paths from a source to a destination. The cost-based traffic engineering techniques proposed in [11, 12, 13] will also be redesigned to balance traffic over the Cyber-healthcare communication platform to increase throughput and reduce communication delays. Deploying a long distance sensor network [14, 15] to support Cyber-healthcare network deployment in the rural settings of the developing world is another key issue that needs to be addressed as future research work.

REFERENCES

[1] M. Mandava, C. Lubamba, A. Ismail, H. Bagula. and A. Bagula. Cyberhealthcare for public healthcare in the developing world. in the Proceedings of the ISCC 2016 Conference, pages 14-19, June 2016. Processing: Opportunities and Challenges, in IEEE International Conference on Services Computing (SCC), Jun 2015, pp. 285-292.

[2] E. Eskin. Discovering Genes Involved in Disease and the Mystery of Missing Heritability, Communications of the ACM, vol. 58, no. 10, pp. 80-87, 2015.

[3] M. Hassanalieragh, A. Page, T. Soyata, G. Sharma, M. K. Aktas, G. Mateos, B. Kantarci, and S. Andreescu, Health Monitoring and Management Using Internet-of-Things (IoT) Sensing with Cloud-based

[4] S. Earley. The Promise of Healthcare Analytics, IEEE Computing Edge, pp. 27-29, June 2015.

[5] L. Wang and R. Ranjan. Processing distributed internet of things data in clouds,IEEE Computing Edge, pp. 12-16, Jun 2015.

[6] S. Babu, M. Chandini, P. Lavanya, K. Ganapathy, and V. Vaidehi, Cloud-enabled remote health monitoring system, in Int. Conf. on Recent Trends in Inform. Tech. (ICRTIT), July 2013, pp. 702-707.

[7] J. Augustyn. The south african triage scale: a tool for emergency nurses, viewed 8 january 2016. http://www.pntonline.co.za/index.php/PNT/article/download/579/849 June 2016.

[8] Cooking Hacks, https://www.cooking-hacks.com/flexible-solar-panel-7-2v-connector, October 2016.

[9] A. B. Bagula. Modelling and implementation of qos in wireless sensor networks: a multiconstrained traffic engineering model. EURASIP Journal on Wireless Communications and Networking, (1), 2010.

[10] A. B. Bagula and A. Krzesinski. Traffic engineering label switched paths in IP networks using a pre-planned flow optimization model. In Proceedings of the IEEE 2001 MASCOTS Symposium, pages 70-77, 2001.

[11] A. B. Bagula. Hybrid traffic engineering: the least path interference algorithm. In Proceedings of the 2004 SAICSIT conference, pages 89-96, 2004.

[12] A. B. Bagula. Hybrid routing in next generation Ip networks. In Computer Communications, 29(7):879892, 2006.

[13] A. B. Bagula. On achieveing bandwidth-aware lsp/lambdasp multiplexing/separation in multi-layer networks. IEEE Journal on Selected Areas in Communications, 25(5):987-1000, 2007.

[14] M. Zennaro, A. Bagula, D. Gascon, and A. Noveleta. Long distance wireless sensor networks: simulation vs reality. In Proceedings of the 4th ACM Workshop on Networked Systems for Developing Regions, (12), 2012.

[15] A. Bagula, M. Zennaro, G. Inggs, S. Scott, and D. Gascon. Ubiquitous sensor networking for development (usn4d): An application to pollution monitoring. In Sensors, 12(1):391-414, 2012.

ABSTRACTS

(Papers marked with an "*" were nominated for the three best paper awards.)

Session 1: Role of ICT in environmental sustainability

S1.1 How organisations can assess and improve their green ICT activities in a standard and efficient way.

Albert Hankel, Patricia Lago (VU University Amsterdam, The Netherlands)

This study demonstrates how a maturity model on Green ICT can help organisations improve themselves and become more environmental sustainable in a standard and efficient manner. For this we have used the SURF Green ICT Maturity Model and facilitated the use of this model in four organisations. These organisations participated in a maturity scan, evaluation session to discuss the results of the scan and a questionnaire on the use of the model. This field study showed that individual participants were very positive about the use of such a model and that it provided inspiration for improvement, both to reduce the environmental impact of ICT as well as to use ICT as an green solution for other business processes.

S1.2 Mobile signal extension in deep sea - towards a safe and sustainable fisheries.

Dineshkumar Singh; Sanjay Kimbahune (Tata Consultancy Services Ltd., India); Veerendra Veer Singh (Mumbai Research Center of CMFRI, India)

Despite having one million active fishermen, India lacks a scalable boat-to-shore communication framework. Small boat operators, which form 80% of entire fishing community, are more vulnerable to the wilder nature of the ocean, due to dynamic sea current, wind speed, direction and wave heights, etc. This is also one of reason for fishermen entering accidently in other country's territory, getting caught and sometimes costing their life. This poses a serious question to fishermen and their family, whether they would return the next time they venture out in to the sea. Deep sea mobile connectivity and Information, given in advance about the sea conditions such as wind speed and wave height can help fishermen to plan their trip effectively and avoid the danger zone. This infrastructure could provide path of building "knowledge society" to fishermen deprived of real time and usable information. For realizing this, integration of the existing communication technologies like VHF transceivers, satellite communication and extension of mobile communication network is needed. A consortium of partners like Central Marine Fisheries Research Institute, Tata Consultancy Service and Tata Tele Services piloted a project for mobile signal extension into the sea upto 30kms. This enabled an extended network accessibility, which led to create an Early Warning System over a digital highway of 30 by 120 KM, along the coast. This paper captures the experiences and recommendations derived from this pilot.

S1.3 Human safety considerations in the emerging ICT environment.

Shailendra K. Hajela (ITU-APT Foundation of India, India)

This paper deals with the challenges posed to human safety in the emerging IoT and ubiquitous network scenario by super intelligent machines/ cognitive robotics that would coexist in the upcoming heterogeneous world comprising humans and super intelligent machines. Questions of trust in the super intelligent machines have been raised by scientists, sociologists, industry stalwarts, etc. After deliberating on the issues, aligning AI to human interests has been suggested as a plausible option. Agenda for further action proposed is that the ITU may constitute a multidisciplinary umbrella group, collaborating with existing technical and policy groups working on AI standards at technology and policy levels, for in-depth study of all facets, bridging the technology innovations and the policy considerations, and bring out Guidelines for AI aligning with human interests adhering to ethical code.

Session 2 - Service and quality standards

S2.1 Invited paper: Ageing, well-being and technology: from quality of life improvement to digital rights management. A French perspective.

Nathalie Devillier (Grenoble Ecole de Management, France)

ICTs are a critical enabler for a sustainable development that helps with unlocking human capabilities. Technology related to ageing belongs to several sustainable development targets and action lines: social inclusion, access to scientific knowledge, capacity building, cyber-security and ehealth. While the objective is the improvement of quality of life for citizens, a special monitoring is needed so as to regulate ICT use in compliance with the rights and freedoms of data subjects. How can we ensure that these tools are implemented in a way that is relevant and appropriate? This article aims at finding the right balance between multiple stakeholders' interests: in particular the relationship between risk management and respect for private life and lighting providers about integrating data protection into the design of their products so as to improve the quality of life for citizens and preserve their privacy.

S2.2 Universal Service, quality caps and net neutrality.*****

Emilio Carrera Félix (Université Paris II - Panthéon Assas, France)

In the actual network configuration, content providers are not involved in the universal service programs with a corresponding participation being taxes, funds or compensations. We propose a regulated Internet contract with a free or very small economic participation to access into a limited version of the service with content providers financing participation to broadcast their high bandwidth content. We study a minimum quality of a service as a strategy of public policy over a broadband telecommunication services to create better absorption of technological benefits as a welfare measure for the users. As results we show a positive effect to propose universal service by quality as complement of universal service obligation, the conditions to determine the prices have to be ruled ex-ante by authorities and finally the best market scenario for welfare superior is determined by competition.

S2.3 Quality and standardization in technology-enhanced learning.

Irina Tal; Gabriel-Miro Muntean (Dublin City University, Ireland); Eva Ibarrola (University of the Basque Country-UPV/EHU, Spain)

Education and technological advances have enabled digital learning technologies to become a key pillar holding up one of the main United Nations Sustainable Development Goals (SDGs): Quality Education. Technology-enhanced learning strategies have led to innovative ICT applications for the development of new improved learning and teaching practices aiding to guarantee inclusive and equitable quality education and promote opportunities for all. Nonetheless, there is still a big challenge to get hold of all the capabilities of these technology-enhanced learning strategies: improving the learner quality of experience (QoE). This paper presents an innovative technology enhanced learning initiative that aims to attract students to STEM education and improve access for students with disabilities through the analysis of the learner QoE. Moreover, in this paper some of the required standards and specifications to be used for developing this initiative are identified. Another key contribution of the paper is that it explores the necessity of new related standards and introduces some novel proposals for standardization in this area.

Session 3 - Spectral efficiency in wireless networks

S3.1 Space division multiplexing technology: next generation optical communication strategy.*

Kazuhide Nakajima; Takashi Matsui; Kotaro Saito; Taiji Sakamoto; Noriyuki Araki (NTT Corporation, Japan)

Space division multiplexing (SDM) is expected to be a key technology both for dealing with the future capacity crunch facing traditional single-mode fibre (SMF) and for realizing a sustainable optical network that can accommodate the various data streams originating from, for example, future 5G communication, the Internet of things (IoT), and machine to machine (M2M) networks. This paper describes the potential of SDM as regards optical fibre and cable technology. We focus on the potential of multi-core fibre (MCF), and investigate the reality of MCF based SDM optical wiring as the first example of an SDM application taking the latest research and development into consideration. Finally, we show that MCF based SDM optical fibre cable will be a promising technology for next generation optical networks, and the key technology behind MCF based SDM optical wiring is ready for discussion as the near future standard.

S3.2 Resource allocation for device-to-device communications in multi-cell LTE-advanced wireless networks with C-RAN architecture.*

Sajjad Mehri Alamouti; Ahmad R. Sharafat (Tarbiat Modares University, Iran)

Device-to-device (D2D) communications underlaying LTE-Advanced wireless networks reuse cellular frequency spectrum to establish direct links between users without traversing base stations or the cellular network. In this paradigm, there is a need to optimally allocate resources with a view to maximizing the utility, e.g., the total throughput, and mitigating the interference caused by sharing the same spectrum between cellular users (CUs) and D2D pairs. This paper proposes a scheme for optimally allocating transmit power levels and channels to maximize the total number of active D2D pairs and reused channels while minimizing the aggregate transmit power pertaining to CUs and D2D pairs. We consider a multi-cell scenario in which the transmitter and the receiver of each D2D pair can be in the same cell or in two different cells, and each user can simultaneously transmit over multiple reused channels. The optimization is done via a centralized baseband processing in the cloud radio access network (C-RAN) architecture. Simulations show that via our proposed scheme, more users (both cellular users and D2D pairs) can simultaneously communicate and the total system throughput is also significantly increased.

S3.3 PAPR reduction in SC-FDMA via a novel combined pulse-shaping scheme.

Naser Ahmadi Moghaddam; Ahmad R. Sharafat (Tarbiat Modares University, Iran)

Peak-to-average-power-ratio (PAPR) is an important parameter that affects the cost of end-user devices in next generation wireless networks. When PAPR is high, the end-user power amplifier's dynamic range should also be high, resulting in costly power amplifiers. Single-carrier frequency-division-multiple-access (SC-FDMA) is used as the air-interface in LTE-Advanced, and this paper proposes a novel and efficient technique for PAPR reduction via pulse shaping for interleaved-FDMA (IFDMA) subcarrier mapping in SC-FDMA. By way of simulations, we show that PAPR can be reduced by 2.11 dB for our novel pulse shaping compared to raised cosine (RC) pulse shaping with QPSK modulation.

S3.4 Accelerating the introduction of spectrum sharing using market-based mechanisms.

Fernando Beltrán (University of Auckland, New Zealand)

Spectrum management needs to be effective, in that spectrum must be allocated to the right uses, and efficient, in that spectrum must be assigned to those that value it the most. Technological advances and demands for further spectrum availability from mobile broadband operators (among others) require spectrum management to timely and firmly incorporate schemes to increase the technical efficiency of spectrum utilisation. One such scheme is spectrum sharing which has the potential to result in higher spectrum utilisation and greater spectrum value. In such context allocation and assignment, two critical functions to manage the spectrum, are also discussed. It is argued that in the course of deciding about allocation and assignment of spectrum, a spectrum authority can and should include market-based mechanisms that incentivise incumbents to share spectrum needed by entrants.

Session 4: Network evolution

S4.1 Invited paper: 5G in rural and low-income areas: are we ready?

Luca Chiaraviglio; Nicola Blefari-Melazzi (CNIT / University of Rome Tor Vergata, Italy); William Liu; Jairo A. Gutierrez (AUT, New Zealand); Jaap Van De Beek (Lulea University of Technology, Sweden); Robert Birke; Lydia Chen (IBM Research, Switzerland); Filip Idzikowski (Faculty of Electronics and Telecommunications, Poznan University of Technology, Poland); Daniel Kilper (The University of Arizona, USA); Paolo Monti (KTH Royal Institute of Technology, Sweden); Jinsong Wu (University of Chile, Chile)

Current trends in telecommunication networks foresee the adoption of the fifth generation (5G) of wireless networks in the near future. However, a large number of people are living without coverage and connectivity. To face this issue, we consider the possibility of deploying 5G networks in rural and low-income zones. After detailing the current state-of-the-art, we consider the main challenges that need to befaced. Moreover, we define the main pillars to follow in order to deploy 5G networks in such zones, as well as a proposal of a future network architecture.

S4.2 Design of scalable directory service for future IoT applications.

Ved P. Kafle; Yusuke Fukushima; Pedro Martinez-Julia; Hiroaki Harai (National Institute of Information and Communications Technology, Japan)

Unprecedentedly a massive number of devices are getting connected in the coming era of the Internet of Things (IoT). For discovery, remote access and management of these IoT devices, an IoT directory service is needed to store and provide their various attributes such as location, generated data types, owner's name, and security keys. In this paper, we present the architectural design of the IoT directory service that is capable to store a huge number of heterogeneous records and provide fast lookup (latency of few milliseconds) and dynamic update (latency of few seconds), while fully complying with owner- or user-centric security and privacy policy. To meet the performance requirements despite fluctuations in the workload and networking environment, we leverage tools of network function and resource virtualization to dynamically allocate and adjust the computational and network resources assigned to the directory service.

S4.3 A stack4things-based platform for mobile CrowdSensing services.

Salvatore Distefano (Kazan Federal University, Russia); Antonio Puliafito; Giovanni Merlino; Francesco Longo; Dario Bruneo (Università di Messina, Italy)

As mobiles grow pervasive in people's lives and expand their reach, Mobile CrowdSensing (MCS) and similar paradigms are going to play an ever more prominent role. There is a pressing need then to ease developers and service providers in embracing the opportunity, and that means offering a platform for such efforts. This in turn means providing a solid foundational architecture with abstractions and sound layering for MCS application designs to be mapped over it. This should base on a flexible infrastructure able to provide resources to MCS applications according to their requirements, hopefully on-demand. A service-oriented/Cloud model can perfectly fill this gap. This paper is a first step in this direction, proposing to adopt Stack4Things (S4T), an OpenStack-based platform for managing sensing and IoT nodes, for runtime customization of resources and their functions to support MCS services and applications. This implies developing and extending the S4T platform further to the specific requirements coming from off-the-shelf, e.g., Android-based, mobiles, as well as describing an example S4T-powered MCS application, Pothole Detection Mapping, to highlight the role of the platform.

S4.4 A popularity-based caching strategy for the future Internet.

Suhaidi Hassan; Ikram Ud Din; Adib Habbal; Nur Haryani Zakaria (Universiti Utara Malaysia, Malaysia)

Information-Centric Networking (ICN) is an attractive network model receiving increasing consideration by the research community because of its inspiring features. To better manage the Internet usage move from host-centric communication to receiver-driven content retrieval, revolutionary ICN architectures have been proposed. A distinguished characteristic of these innovative architectures is to provide ubiquitous and transparent in-network caching to enhance network resource utilization and accelerate content dissemination. With the exponential increase of Internet traffic, the issue of content storage is a growing concern in ICN. In this paper, we present a caching strategy that considerably increases cache hit rate and reduces stretch ratio, which are the most important metrics in the evaluation of ICN caching. Through extensive simulations, it is shown that our proposed work is a favorable and realistic contribution for the standardization exercise of data caching for achieving accurate and valid network performance in the future Internet.

S4.5 Multi-path chunked video exchanges over OF@TEIN SDN cloud playground.

Phyo May Thet; Chaodit Aswakul (Chulalongkorn University, Thailand); JongWon Kim (Gwangju Institute of Science & Technology, Korea)

Recent explosion of SDN (software-defined networking) paradigm is steadily facilitating the open evolution of legacy protocol-biased networking. In this paper, we design and develop middle-box splitting functionalities for chunked video exchanges (i.e., file transfer and streaming) over multiple concurrent paths over South Korea, Malaysia, and Thailand parts of OF@TEIN SDN-Cloud playground. To reduce middle-box processing delay and to enable multi-path capacity leverage, we propose to combine the multi-path file-transfer function and Tsunami protocol. This combination allows the multi-path-based pre-fetch transfer of 4K-video chunks from Tsunami server in South Korea to Thailand's SmartX Box. The experiment results show that the proposed scheme can reduce 23~72 seconds in transferring 10-minute 4K-video file by effectively utilizing the available international multi-path capacity.

Session 5: Services and implementation-related issues

S5.1 Implementation of tele-rehabilitation system combined with video call center.*

Kotaro Suzuki; Yoshitoshi Murata (Iwate Prefectural University, Japan)

Japan's low birthrate and rapidly aging population are causing medical expenses to take up ever more of the national budget and leading to a shortage of young medical professionals. As a result, rehabilitation therapy is being shifted from hospital-care to home-care. Several other countries will also face the same situation in the near future. Thus, we propose a tele-rehabilitation system combined with a video call center to make up for the shortage of rehabilitation therapy done by visiting physiotherapists. A video call center operator coaches a patient instead of a physiotherapist, and a physiotherapist supervises multiple operators. The system focuses on cerebrovascular patients who have a home-visit rehabilitation or an outpatient one and uses Microsoft KINECT to measure strain of the upper body. In this paper, implementation of this system is mainly described.

S5.2 Intricacies of implementing an ITU-T X.1303 cross-agency situational-awareness platform in Maldives, Myanmar, and the Philippines.*

Biplov Bhandari; Angga Bayu Marthafifsa; Manzul Kumar Hazarika (Asian Institute of Technology, Thailand); Francis Boon; Nuwan Waidyanatha; Lutz Frommberger (Sahana Software Foundation, USA)

Maldives, Myanmar, and the Philippines are vulnerable to natural disasters. Sendai Framework of Action calls for risk reduction by implementing early warning systems. A prevailing challenge is for authorities to coordinate warnings across disparate communication systems and autonomous organizations. Cross-Agency Situational-Awareness platforms and the ITU-T X.1303 Common Alerting Protocol (CAP) interoperable data standards presents themselves as solution for diluting the inter-agency rivalries and interconnection disparities. The Sahana Alerting and Messaging Broker (SAMBRO) was designed to overcome these issues by providing a Common Operating Picture and a platform for all Stakeholders to share early warnings. To that end, the CAP-on-a-MAP project is implementing SAMBRO and the CAP standard along with the policies and procedures in the Maldives, Myanmar and Philippines. The project is applying an agile development methodology with a design, build, test, and redesign strategy for implementing the cross-agency situational-awareness and warning system in the respective countries. This paper discusses the country context implementation challenges and discusses strategies fostered through the introduction of the CAP content standard for warning system designers to consider for overcoming similar challenges.

S5.3 A community-driven information system to develop next generation collaborative and responsive rural community (NCoRe).

Jayanta Basak; Rishikesan Parthiban; Somprakash Bandyopadhyay (Indian Institute of Management Calcutta, India); Siuli Roy (Heritage Institute of Technology, India)

Much of the ICT interventions for rural transformation are exogenous in nature (development from outside), in the sense that they use a "push" approach towards development, without considering the nature and problems of an individual member of the rural community. Information and knowledge transactions, especially with dis-empowered people and groups, are a complex process and ICT needs to be appropriated and used in a way that helps resolve daily concerns. With this perspective in mind, this paper proposes NCoRe, an interactive community-driven information system platform to harness the potential of community participation in governance. In a digitally-connected global society, each individual in a community of people is not only a consumer of information but also a producer of information: a potential contributor in many ways to build a better community. NCoRe exploits the potential of community knowledge, making them available to the community and empowering the communities to interact, collaborate and participate in the development of society and transforming the way they live, learn and work. NCoRe is an ongoing initiative to build next-generation collaborative and responsive community by empowering the rural community of India with an ICT-enabled "capability framework" involving the self-help groups (SHG): the micro-communities within a village community.

S5.4 Toward authenticated caller ID transmission: the need for a standardized authentication scheme in Q.731.3 calling line identification presentation.*

Huahong Tu; Adam Doupé; Ziming Zhao; Gail-Joon Ahn (Arizona State University, USA)

The rising prevalence of phone fraud is hurting consumers and businesses. With about a half million reports each year in the United States, phone fraud complaints have more than doubled since 2013. In the current calling line identification presentation scheme, the caller ID is trivially spoofed. Scammers are using spoofed caller IDs to trick their victims into answering unwanted calls and further a variety of scams. To provide a solution to this problem, this paper proposes an authentication scheme that provides the possibility of a security indicator for the current Q.731.3 calling line identification presentation supplementary service. The goal of this proposal is to help prevent users from falling victim to phone impersonation scams, as well as provide a foundation for future defenses to stop unwanted calls based on the caller ID information. This work will help to guide the future development of a standardized scheme in authenticating SS7 identities.

Session 6: Sustainability and smartness

S6.1 Certified security systems for sustainable cities of the 21st century.

Simone Wurster (Berlin University of Technology, Germany); Irene Kamara (Vrije Universiteit Brussel, Belgium); Thordis Sveinsdottir (Trilateral Research & Consulting, United Kingdom)

The United Nations formulated 17 sustainable development goals to "transform our world". Goal 11 aims to "make cities and human settlements inclusive, safe, resilient and sustainable". As such, security systems have become increasingly relevant, particularly in the past several years as significant and dangerous threats have emerged throughout the world. In addition to said risks, a number of security solutions, for example in the field of CCTV, are linked with significant privacy risks. Therefore, an appropriate certification scheme for security systems that not only considers security aspects but also additional issues, e.g. data protection and privacy, is needed in Europe. The EU Project CRISP (Evaluation and Certification Schemes for Security Products) aims to facilitate this process via the development of pan-European certification. This paper shows CRISP's solutions based on the current outcomes of the project and its specific contribution to research and practice.

S6.2 WiFi networks on drones.

Antonio Guillen-Perez, Maria-Dolores Cano; Juan Carlos Sanchez-Aarnoutse; Joan Garcia-Haro (Universidad Politécnica de Cartagena, Spain); Ramon Sanchez-Iborra (Universidad Politécnica de Cartagena, Spain / Universidad Técnica Federico Santa María, Chile)

The huge growth in the number of connected wireless devices leads to an increasing demand for network connectivity. In this context, aerial networks may play an important role by widening the concept of access networks. This paper describes and analyzes one of the most promising applications of Unmanned Aerial Vehicles, commonly known as drones, in the field of communications: Extending the capacity or coverage of wireless systems through the deployment of aerial communication networks. We present a comprehensive characterization study of an experimental system to deploy an aerial WiFi network. To do so, an Intel Galileo development board is appropriately configured and equipped as a WiFi node playing either the role of an access point in the infrastructure mode or of an intermediate hop in the ad-hoc operational mode. This device is then integrated onboard a drone. We compare both WiFi modes in terms of coverage area, throughput, and energy efficiency. Preliminary results reveal that there is a trade-off between coverage and data rates, for which the infrastructure mode performs better, and energy efficiency, where the ad-hoc mode is more responsive.

Poster Session

P.1 Adaptive video streaming over HTTP using stochastic bitrate prediction in 4G wireless networks.

Dhananjay Kumar; S. Aishwarya; A. Srinivasan; L. Arun Raj (Anna University, India)

Video streaming over Hypertext Transfer Protocol (HTTP) used in multifarious applications creating a multimedia environment faces a challenge in 4G wireless network due to the fluctuating nature of internet traffic and variable capacity of wireless channel over time. The existing Dynamic Adaptive Streaming over HTTP (DASH), though works well for stored video up to some extent, poses a complication in live transmission thereby depreciating the streaming quality due to high link bit rate fluctuation. In this paper, we have proposed an efficient ARIMA Based Bit Rate Adaptation (ABBA) model in the receiver/client side that estimate the link traffic based on the incoming packet bit rate to predict the subsequent future link capacity in order to notify the sender/server. Based on the response from the receiver the server adapt its outgoing stream as per forecasted link data rate, and hence eliminate the degradation of video due to channel throughput variations. The proposed ABBA algorithm was implemented on IP over 4G wireless network and the streaming quality was evaluated on several full reference metrics of video quality. The test result outperformed an existing buffer based approach and also a fuzzy based adaptation algorithm. For example, the ABBA algorithm exhibited an average increase of 22 % in PSNR and 9% in SSIM than a buffer based method.

P.2 Assessing Internet performance over mobile networks: from theory to practice.

Eneko Atxutegi; Jose Oscar Fajardo; Eva Ibarrola; Fidel Liberal (University of the Basque Country-UPV/EHU, Spain)

The proper execution of performance tests is of utmost importance for the analysis and evaluation of Internetrelated technologies, protocols and deployment strategies. There are plenty of tools available for experimenters ranging from simulation tools, emulation equipment, to small and large scale experimentation testbeds. Each of these performance evaluation frameworks introduces a series of capabilities and drawbacks. Additionally, there is a need for using a common methodology to perform different performance tests and to create comparable outcomes. Seeking for the standardization of reliable and comparable Internet speed measurements and trying to provide end-users with trustworthy measurement tools, different SDOs have standardized different measurement methodologies. This paper focuses on the applicability of the distinct measurement frameworks and methodologies in the field of mobile Internet, with special focus on the impact of using TCP over mobile broadband connections. The paper describes a series of experiments over different frameworks and, based on the obtained results, identifies the implications of the different types of performance tests into the evaluation outcomes and states the need for large scale measurements.

P.3 Systematic analysis of geo-location and spectrum sensing as access methods to TV white space.

Hope Mauwa; Antoine Bagula (University of The Western Cape, South Africa); Marco Zennaro; Ermanno Pietrosemoli (The Abdus Salam International Centre for Theoretical Physics, Italy); Albert A. Lysko (Council for Scientific and Industrial Research, South Africa); Timothy X Brown (Carnegie Mellon University, Rwanda)

Access to the television white space by white space devices comes with a major technical challenge: white space devices can potentially interfere with existing television signals. Two methods have been suggested in the literature to help white space devices identify unused channels in the TV frequency band so that they can avoid causing harmful interference to primary services legally protected to run on the band. These methods are geo-location spectrum database and spectrum sensing. Discussions in the literature have placed much emphasis on the limitations of the spectrum sensing approach and mainly based on the developed world environment ignoring the performance requirements of the geo-location database approach and how the absence of these requirements in a developing region could affect its performance. This paper considers a broader analysis of the approaches by looking at factors that can affect the performance of each approach and how the presence or absence of these factors in a developed region or developing region can affect their performance. In so doing, the paper highlights the need to conduct more research on the performance of spectrum sensing in developing regions where there are plenty of white spaces to ascertain its use in these regions.

P.4 Task-based process modeling for policy making in smart cities.

Leonidas Anthopoulos (University of Applied Science (TEI) of Thessaly, Greece); George Giannakidis (Center for Renewable Energy Sources and Saving, Greece)

Several competitive standards have been introduced for smart city quite recently, which define the architecture and its components or city key performance indicators. However, these standards do not discuss smart service formulation, nor policy making process modeling. Standardization assists in achieving process automation by introducing “best practices” as standard process models. Policy making mainly follow non-standardized procedures, even if it is supported by various tools (i.e., Multi-Criteria Decision Methods (MCDM)). Inspired by the Task-Based Modeling method (TBM) this paper focuses on policy making process standardization for smart cities. It utilizes the case-study of the InSmart (Integrative Smart City Planning) coordination action in the smart city of Trikala, Greece, in an attempt to define and introduce a model for such a process.

P.5 CleanWiFi: the wireless network for air quality monitoring, community Internet access and environmental education in smart cities.

Carlos Andrés Gómez Ruíz (Universitaria Agustiniana, Colombia)

This work presents a new development model for wireless community networks, framed within the sustainable development goals, SDG 11, to achieve the transformation of cities to environmentally sustainable areas and to take measures to fight climate change; this model is based on the use of renewable energies, mesh routing protocols, the monitoring of air quality and environmental variables, IoT, and the application of educational methodologies in order to reward less polluted areas. There are proposed some ideas about standardization of building networks for measuring air quality in smart cities, which provide a great source of information for location-based services (LBS) that promote environmental awareness services. The CleanWiFi network constantly monitors the air for pollutant gases, uses that information to feed a Big Sensor Data system, and uses the same data for the automatic configuration of the public WiFi service, displaying information about the quality of the air to the user, and rewarding less oolluted areas with a better service. That way it raises public awareness about state of air pollution and how important it is to reduce it; promotes the use of renewable energies and brings WiFi connectivity to the people.

P.6 Cloud based patient prioritization as service in public health care.

Antoine Bagula; Claude Lubamba; Munyaradzi Mandava (University of The Western Cape, South Africa); Marco Zennaro; Ermanno Pietrosemoli (The Abdus Salam International Centre for Theoretical Physics, Italy); Herman Bagula (University of Cape Town, South Africa)

This paper proposes and evaluates the performance of a Cyber-healthcare system which is aimed at providing patient prioritization over the cloud as a public health service for the rural and urban communities of the developing world. The underlying cloud-based Internet-of-Things (IoT-Cloud) infrastructure is aimed to be implemented in the city of Lubumbashi in the republic Democratic Republic of the Congo (DRC) with the objective of setting up a community health care network around a mesh of health kiosks. We propose a deployment model for the proposed Cyber-healthcare system, and describe a patient prioritization process as part of its situation awareness component. The results obtained from an experimental prototype reveal the field readiness of the off-the-shelf bio-sensor technology used by the system, the performance achieved when using a solar powered subsystem, the relative communication capabilities provided by its protocols and the network engineering feasibility of the planned community health care network. The relative efficiency of using supervised machine learning compared to unsupervised machine learning when performing patient prioritization, is also revealed through two popular algorithms: Multivariate linear regression (MLR) and K-means clustering (KMC).

INDEX OF AUTHORS

Index of Authors

ISBN 978-92-61-20441-9

Printed in Switzerland
Geneva, 2016