

International Telecommunication Union

Proceedings of the 2011
ITU Kaleidoscope
Academic Conference

**The fully
networked
human?**

Innovations for future networks and services
Cape Town, South Africa, 12-14 December 2011

Supporters:

Nokia Siemens
Networks



BlackBerry

Host:



Organizer:



Technical co-sponsor:



Proceedings of the 2011
ITU Kaleidoscope
Academic Conference

**The fully
networked
human?**

**Innovations for future
networks and services**

Cape Town, South Africa, 12-14 December 2011

Supporters:

**Nokia Siemens
Networks**



BlackBerry.

Host:



Organizer:



Technical co-sponsor:



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.

© ITU 2011

All rights reserved. No part of this publication may be reproduced, by any means whatsoever, without the prior written permission of ITU.

Foreword

Malcolm Johnson
Director
ITU Telecommunication Standardization Sector



Standardisation is the channel through which most technological innovation must pass if it is to be applied on a global scale; and when we speak of ICT, it is always in a global context. This is precisely why ITU-T seeks to strengthen its relationship with academia, the source of so much crucial research and so many innovative ideas.

Launched in 2008, Kaleidoscope has always enjoyed enthusiastic participation by academia. This year was no different, with 84 papers submitted and 30 selected for presentation at the conference. Particularly encouraging was the proportion of these papers originating in the so-called developing world, a key indication of its desire and need for participation in ICT standardization discussions.

The fourth Kaleidoscope conference took place in Cape Town, South Africa, and very appropriately so. In ICT terms, Africa is the last great frontier. The digital age we inhabit presents Africa the opportunity to leapfrog many steps in technological, and thereby economic development, learning from and thus avoiding the mistakes made in the past by developed nations. Taking part in the standardization process is an excellent opportunity for developing countries to have their voices heard in the development of new technologies, ensuring they play a key role in the creation of our collective future.

This year's theme, *The fully networked human? – Innovations for future networks and services*, attracted many excellent submissions and proved fuel for an illuminating debate on the future of ICT.

Human-centric ICT, as it has come to be known, is technology intrinsically designed to place its user at the centre, with virtualised networks, other IT resources, services and applications automatically adapting to the specific circumstances of the user. Having ICT adapt and respond to our activities as we perform them, or to our preferences as we form them, will better synergise our physical and digital worlds; ensuring a simpler, more personalised engagement with the technology we use so frequently.

A profoundly interesting event, Kaleidoscope 2011 further solidified ITU's relationship with academia, and once again unearthed ideas and innovations expressly designed to better our world. ITU is immensely thankful for the valuable contribution academia continues to make to the ICT field, and we thank all Kaleidoscope's participants for ensuring the conference remains the exciting, informative event it has become.

On ITU's behalf, my sincerest thanks go to the Department of Communications of the Republic of South Africa for making this event possible; our gracious hosts, the University of Cape Town; our generous sponsors, Nokia Siemens Networks, Telkom SA, and BlackBerry; our hard working Organizing Committee and Programme Committee members; and of course our engaging chairman, Dr. Mostafa Hashem Sherif.

A handwritten signature in blue ink, which appears to read "Malcolm Johnson". The signature is fluid and cursive, written over a white background.

Malcolm Johnson
Director
ITU Telecommunication Standardization Sector



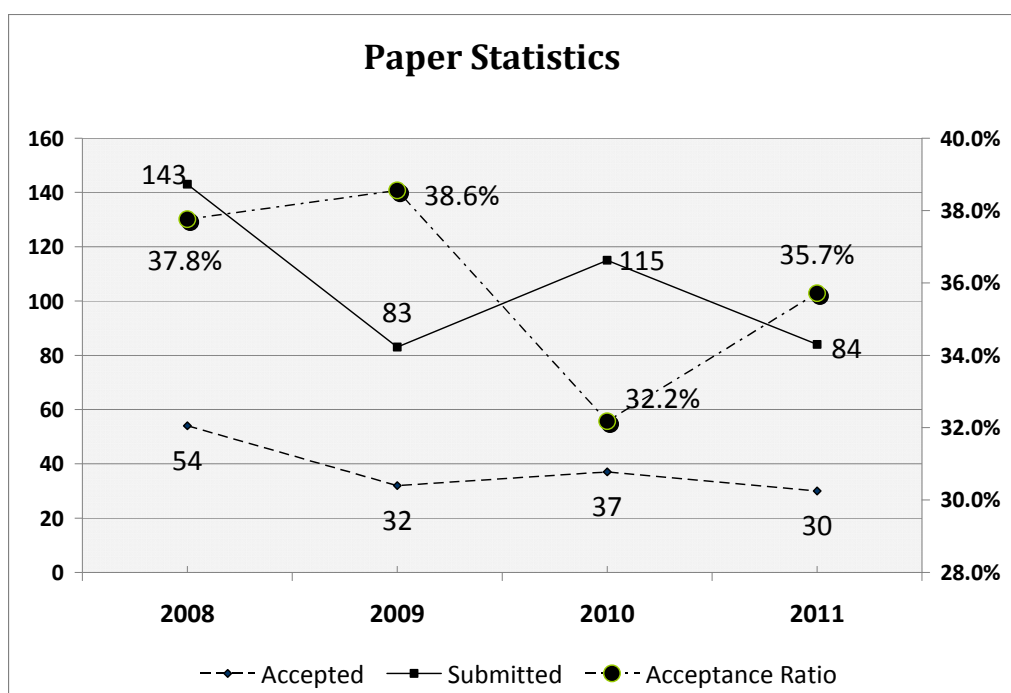
Chair's Message

Mostafa Hashem Sherif
General Chair

The ITU initiated its Kaleidoscope series of conferences in 2008 to provide a forum for practitioners and researchers on the standardisation of information and communication technologies (ICTs). Mr. Yoichi Maeda from NTT, Japan, chaired the first three conferences; in Geneva, Switzerland (2008), in Mar del Plata, Argentina (2009) and in Pune, India (2010). Under his stewardship, Kaleidoscope events have grown to become an important venue for Academia and for the ITU. This year, however, Mr. Maeda had to step aside because of his obligations as president of the Telecommunication Technology Committee (TCC) of Japan. It is a great honour for me to follow in his steps and chair the 2011 Kaleidoscope academic conference with the theme of "The fully networked human?—Innovations for future networks and services."

Kaleidoscope 2011 focuses on the multidisciplinary aspects of future human-centric systems. In this environment, virtualized networks, information technology resources, services and applications are adaptively and automatically configured to support users in their everyday activities. Whether technologies are competing or complementary becomes less significant; what counts is whether the service offer fulfils user needs and respects environmental and legal constraints.

The programme committee chaired by Dr. Kai Jakobs selected 30 papers out of 84 submissions from 29 countries on the basis of double-blind reviews with the help of 150 international experts. They also worked diligently to identify candidate papers for the awards. As seen in the Figure, the high quality of submissions made the selection challenging. Let me express here my deep gratitude to all reviewers and members of the programme committee for their generous contribution of time and effort to the success of this conference.



The geographic distribution of the accepted contributions using the country of the first author is shown in the following Table:

Country	Lecture	Poster
Canada	1	
Germany		1
Greece		1
India	1	
Italy	1	2
Japan	3	3
Malawi	1	
Malaysia		1
Senegal	1	
South Africa	7	1
Spain	1	
Switzerland	1	
United Kingdom	2	
United States	2	
Total	21	9

We are also delighted to have many distinguished speakers in the plenary sessions. Their talks cover academic, industrial and business aspects of a society with fully networked individuals. Moreover, Professor Thomas Magedanz from the Technische Universität Berlin, Germany, has generously volunteered to present a tutorial on next generation networks to complement the forward-looking aspects of the conference.

This year’s Jules Verne’s corner is on “chips in the brain” and how “fully networked humans” would behave 50 years from now.

By agreement with the IEEE Communications Society (ComSoc), selected papers from each year’s conference are typically considered for publication in a special feature section of the IEEE Communications Magazine. The tentative schedule for the 2011 vintage is the August 2012 issue of the Magazine. In addition, special issues of the International Journal of Technology Marketing (IJTMKT) and the International Journal of IT Standards and Standardization Research (IJITSR) are open to revised versions of papers from the conference.

As usual, all accepted papers are accessible through the IEEE Xplore digital library. The Proceedings from 2009 onwards can be downloaded at no charge from <http://itu-kaleidoscope.org>.

In closing, I would like to thank our hosts, the Department of Communications of the Republic of South Africa and the University of Cape Town. I would like also to express my gratitude to Nokia Siemens Networks and Telkom SA for their financial support; also to BlackBerry for donating two PlayBooks. Finally, the organization of such a conference would not have been possible without the extremely capable staff from the ITU-T Telecommunication Standardization Bureau (TSB), whose dedication and professionalism I wish to salute here.



Mostafa Hashem Sherif
General Chair

TABLE OF CONTENTS

	Page
Foreword.. .. .	i
Chair's message .. .	iii
Committees.....	ix
Keynote Summaries	
Rufus Andrew (Managing Director, Nokia Siemens Networks, South Africa)	3
Hirofumi Horikoshi (General Manager, Technology Planning Department, Nippon Telegraph and Telephone Corporation, Japan).....	4
Alfredo Terzoli (Professor, Rhodes University, South Africa).....	6
Session 1: ICTs helping Africa	
S1.1 The Role Of ICTs In Quantifying The Severity and Duration Of Climatic Variations - Kenya's Case..... <i>Muthoni Masinde; Antoine Bagula</i>	9
S1.2 ICT use in South African Microenterprises: An assessment of Livelihood outcomes. <i>Frank Makoza; Wallace Chigona</i>	17
S1.3 SM ² : Solar Monitoring System in Malawi..... <i>Mayamiko Nkoloma; Marco Zennaro; Antoine Bagula</i>	25
Session 2: Connecting rural regions	
S2.1 Proposal of a Wired Rural Area Network with Optical Submarine Cables..... <i>Yoshitoshi Murata; Hiroshi Mano; Hitoshi Morioka</i>	33
S2.2 Development of an ICT road map for eServices in rural areas..... <i>Robert Rangarirai Jere; Mamello Thinyane; Alfredo Terzoli</i>	41
S2.3 Investigating implementation of communication networks for advanced metering infrastructure in South Africa	49
Session 3: Reflections on a fully networked society	
S3.1 Invited paper: Cooperative Wi-Fi-Sharing: Encouraging Fair Play	59
<i>Hanno Wirtz; René Hummen; Nicolai Viol; Tobias Heer; Mónica Alejandra Lora Girón Klaus; Klaus Wehrle (RWTH Aachen University, Germany)</i>	
S3.2 Making things socialize in the Internet - Does it help our lives?..... <i>Luigi Atzori; Antonio Iera; Giacomo Morabito</i>	67
S3.3 Net-Centric World: Lifestyle of the 21st Century	75
<i>Daniel Kharitonov</i>	
S3.4 Reflexive Standardization of Network Technology	83
<i>Ian Graham</i>	

Session 4: Frequency and Spectrum Management

S4.1	Radio Resource Management in OFDMA-CRN Considering Primary User Activity and Detection Scenario.....	91
	<i>Dhananjay Kumar; Kanagaraj Nachimuthu Nallasamy</i>	
S4.2	Optimal Pilot Patterns Considering Optimal Power Loading for Cognitive Radios in the Two Dimensional Scenario.....	99
	<i>Boyan Soubachov; Neco Ventura</i>	
S4.3	Optimal Spectrum Hole Selection & Exploitation in Cognitive Radio Networks....	105
	<i>Mahdi Pirmoradian; Christos Politis</i>	

Session 5: Optimisation of Layers 1 – 3

S5.1	Transmission Analysis of Digital TV Signals over a Radio-on-FSO Channel.....	115
	<i>Chedlia Ben Naila; Kazuhiko Wakamori; Mitsuji Matsumoto; Katsutoshi Tsukamoto</i>	
S5.2	A Hybrid MAC with Intelligent Sleep Scheduling for Wireless Sensor Networks..	123
	<i>Mohammad Arifuzzaman; Mohammad Shah Alam; Mitsuji Matsumoto</i>	
S5.3	Route Optimization Based On The Detection of Triangle Inequality Violations.....	131
	<i>Papa Ousmane Sangharé; Bamba Gueye; Ibrahima Niang</i>	

Session 6: Architectures to support a fully networked society

S6.1	Invited Paper: Effective Collaborative Monitoring In Smart Cities: Converging Manet And Wsn For Fast Data Collection.....	141
	<i>Giuseppe Cardone; Paolo Bellavista; Antonio Corradi; Luca Foschini (University of Bologna, Italy)</i>	
S6.2	SOA Driven Architectures for Service Creation through Enablers in an IMS Testbed.....	149
	<i>Mosiuo Tsietsi; Alfredo Terzoli; George Wells</i>	
S6.3	A Virtualized Infrastructure for IVR Applications as Services	157
	<i>Fatma Belqasmi; Christian Azar; Mbarka Soualhia; Nadjia Kara; Roch Glitho</i>	
S6.4	Seamless Cloud Abstraction, Model and Interfaces	165
	<i>Masum Z. Hasan; Monique Morrow; Lew Tucker; Sree Lakshmi D. Gudreddi; Silvia Figueira</i>	

Session 7: Service Quality for a fully networked society

S7.1	Regulation of Bearer / Service Flow Selection between Network Domains for Voice over Packet Switched Wireless Networks.....	175
	<i>Nikesh Nageshar; Rex Van Olst</i>	
S7.2	Accessibility support for persons with disabilities by Total Conversation Service Mobility Management in Next Generation Networks	181
	<i>Leo Lehmann</i>	
S7.3	LabQoS: A platform for network test environments	189
	<i>Luis Zabala; Armando Ferro; Cristina Perfecto; Eva Ibarrola; Jose Luis Jodra</i>	

	Page
Poster Session: Showcasing innovations for future networks and services	
P.1 A Trust Computing Mechanism for Cloud Computing <i>Mohamed Firdhous; Osman Ghazali; Suhaidi Hassan</i>	199
P.2 The Energy Label A Need To Networks And Devices..... <i>Virgilio Puglia</i>	207
P.3 A distributed mobility management scheme for future networks..... <i>Ved P. Kafle; Yasunaga Kobari; Masugi Inoue</i>	215
P.4 Toward Global Cybersecurity Collaboration: Cybersecurity Operation Activity Model..... <i>Takeshi Takahashi; Youki Kadobayashi; Koji Nakao</i>	223
P.5 Context Representation Formalism and Its Integration into Context as a Service in Clouds <i>Boris Moltchanov</i>	231
P.6 Supporting technically the Continuity of Medical Care: Status report and perspectives..... <i>Vasileios B. Spyropoulos; Maria Botsivaly; Aris Tzavaras</i>	239
P.7 Coexistence of a TETRA System with a Terrestrial DTV System in White Spaces. <i>Heejoong Kim; Hideki Sunahara; Akira Kato</i>	247
P.8 Mobile cloud computing based on service oriented Architecture: embracing network as a service for 3rd party application service providers..... <i>Michael Andres Feliu Gutierrez; Neco Ventura</i>	253
P.9 RBAC for a configurable, heterogeneous Device Cloud for Web Applications. <i>Hannes Gorges; Robert Kleinfeld</i>	261
Abstracts	269
Index of authors.....	281

COMMITTEES

Organizing Committee

- General Chairman: Mostafa Hashem Sherif (AT&T, USA)
- Tohru Asami (University of Tokyo, Japan)
- Ashok Chandra (Ministry of Communications & IT, India)
- Christoph Dosch (IRT, Germany)
- Linda Garcia (Georgetown University, USA)
- Yoshikazu Ikeda (Otani University, Japan)
- Kai Jakobs (RWTH Aachen University, Germany)
- Mitsuji Matsumoto (Waseda University, Japan)
- Yushi Naito (Mitsubishi Electric, Japan)
- Ramjee Prasad (Aalborg University, Denmark)
- Felipe Rudge Barbosa (University of Campinas, Brazil)
- Helmut Schink (Nokia Siemens Networks, Germany)
- Alfredo Terzoli (Rhodes University, South Africa)
- Daniele Trincherò (Politecnico di Torino, Italy)

Secretariat

- Alessia Magliarditi, Project Head
- Martin Adolph, Project Technical Advisor
- Leslie Jones, Administrative support
- Pablo Palacios, Administrative support
- Toby Johnson, Promotion Advisor
- Simão Campos Neto, Project Advisor
- Stefano Polidori, Project Advisor

Programme Committee

- Chairman: Kai Jakobs (RWTH Aachen University, Germany)
- Vice Chairman, Track 1: Mitsuji Matsumoto (Waseda University, Japan)
- Vice Chairman, Track 2: Alfredo Terzoli (Rhodes University, South Africa)
- Vice Chairman, Track 3: Linda Garcia (Georgetown University, USA)

- Finn Aagesen (Norwegian University of Science and Technology, Norway)
- Marcelo F. Abbade (Pontifical Catholic University in Campinas, Brazil)
- Ahmad Zaki Abu Bakar (Universiti Teknologi Malaysia, Malaysia)
- Martin Adolph (Telecommunication Standardization Bureau, ITU)
- Artem Adzhemov (Moscow Technical University, Russia)
- Mohammad Alam (Waseda University, Japan)
- Bartosz Balis (AGH University of Science and Technology, Poland)
- Abdelmoula Bekkali (NTT Corporation, Japan)
- Paolo Bellavista (University of Bologna, Italy)
- Pierre-Jean Benghozi (Ecole Polytechnique - CNRS, France)
- Andreas Berger (Telecommunications Research Center Vienna, Austria)
- Vitor Bernardo (University of Coimbra, Portugal)
- Jose Everardo Bessa Maia (State University of Ceará, Brazil)
- Shiddhartha Bhandari (Institut Telecom SudParis, France)
- Mauro Biagi (University "La Sapienza" of Rome, Italy)
- Niklas Blum (Fraunhofer Institute FOKUS, Germany)
- Luiz Henrique Bonani (Federal University of ABC - UFABC, Brazil)
- Dario Bottazzi (Guglielmo Marconi Labs, Italy)
- Michael Bove (MIT, USA)
- Cagatay Buyukkoc (AT&T, USA)
- Marco Carugi (ZTE Corporation, China)
- Marcelo Carvalho (University of Brasilia, Brazil)
- Vicente Casares-Giner (Universitat Politècnica de Valencia, Spain)
- Piero Castoldi (Scuola Superiore Sant'Anna, Italy)
- Andre Cavalcante (Nokia Institute of Technology, Brazil)
- Isabella Cerutti (Scuola Superiore Sant'Anna, Italy)
- Jaeho Choi (Chonbuk National University, Korea)
- Antonio Corradi (University of Bologna, Italy)
- Noel Crespi (GET-INT Institut National des Télécommunications, France)
- Marilia Curado (University of Coimbra, Portugal)

- Salvatore D'Alessandro (University of Udine, Italy)
- Marc De Leenheer (Ghent University, Belgium)
- Alvaro Augusto de Medeiros (Federal University of Juiz de Fora, Brazil)
- Ugo Dias (University of Brasilia, Brazil)
- Luca Di Bert (University of Udine, Italy)
- Fadel Digham (National Telecom Regulatory Authority, Egypt)
- Christoph Dosch (IRT, Germany)
- Tineke Egyedi (Delft University of Technology, The Netherlands)
- Tamer ElBatt (Nile University, Egypt)
- Mahmoud T. El-Hadidi (Cairo University, Egypt)
- Dmitry Epstein (Cornell University, USA)
- Luis Carlos Erpen de Bona (Federal University of Paraná, Brazil)
- Mario Fanelli (University of Bologna, Italy)
- José Ewerton Farias (Federal University of Campina Grande, Brazil)
- Armando Ferro Vázquez (ETSI de Bilbao, Spain)
- Erwin J.A. Folmer (University of Twente, The Netherlands)
- Luca Foschini (University of Bologna, Italy)
- Miguel Franklin de Castro (Federal University of Ceará, Brazil)
- Ivan Gaboli (Italtel SpA, Italy)
- Alex Galis (University College London, United Kingdom)
- Ivan Ganchev (University of Limerick, Ireland)
- Molka Gharbaoui (Scuola Superiore Sant'Anna, Italy)
- Carlo Giannelli (University of Bologna, Italy)
- Katja Gilly (Miguel Hernandez University, Spain)
- Anahita Gouya (AFD Technologies, France)
- Victor Govindaswamy (Texas A&M University, USA)
- Ian Graham (University of Edinburgh, United Kingdom)
- Adam Grzech (Wroclaw University of Technology, Poland)
- Chris Guy (The University of Reading, United Kingdom)
- René Hummen (RWTH Aachen University, Germany)
- Eva Ibarrola (University of the Basque Country, Spain)
- Seong-Ho Jeong (Hankuk University of Foreign Studies, Korea)
- Nils Joachim (Bamberg University, Germany)
- Carlos Juiz (University of the Balearic Islands, Spain)
- Oliver Jung (Telecommunications Research Center Vienna, Austria)
- Ved Kafle (National Institute of Information and Communications Technology, Japan)

- Kamugisha Kazaura (Tanzania Telecommunications Company Limited, Tanzania)
- Tim Kelly (World Bank, USA)
- Hemanth Khambhammettu (Université du Quebec en Outaouais, Canada)
- Masafumi Koga (Oita University, Japan)
- Andrej Kos (University of Ljubljana, Slovenia)
- Katarzyna Kosek-Szott (AGH University of Science and Technology, Poland)
- Ken Krechmer (University of Colorado, USA)
- Claude Lamblin (France Telecom, France)
- Matti Latva-aho (University of Oulu, Finland)
- Gyu Myoung Lee (Institut Telecom SudParis, France)
- Heejin Lee (Yonsei University, Korea)
- Leo Lehmann (OFCOM, Switzerland)
- João Leite (University of Brasilia, Brazil)
- Fidel Liberal (ETSI de Bilbao, Spain)
- Luigi Logrippio (Université du Québec en Outaouais, Canada)
- Waslon Lopes (Federal University of Campina Grande, Brazil)
- Jose Giovanni López Perafán (University of Cauca, Colombia)
- Mónica Alejandra Lora (RWTH Aachen University, Germany)
- Giovanni Mancilla (Universidad Distrital, Colombia)
- Barbara Martini (CNIT, Italy)
- Peter Martini (University of Bonn, Germany)
- Lorne Mason (McGill University, Canada)
- Arturas Medeisis (Vilnius Gediminas Technical University, Lithuania)
- Florian Metzger (University of Vienna, Austria)
- Werner Mohr (Nokia Siemens Networks, Germany)
- Antonella Molinaro (University "Mediterranea" of Reggio Calabria, Italy)
- Pedro Henrique Juliano Nardelli (University of Oulu, Finland)
- Mohammed Nafie (Nile University, Egypt)
- Joan Olmos (Universitat Politecnica de Catalunya, Spain)
- Fumitaka Ono (Tokyo Polytechnic University, Japan)
- David Palma (University of Coimbra, Portugal)
- Riccardo Passerini (Telecommunication Development Bureau, ITU)
- Henrique Pequeno (Federal University of Ceará, Brazil)
- Francisco Portelinha (University of Campinas, Brazil)
- Louis Pouzin (Eurolinc, France)
- Albert Rafetseder (University of Vienna, Austria)

- Alessandro Raschellà (Universitat Politecnica de Catalunya, Spain)
- Peter Reichl (Telecommunications Research Center Vienna, Austria)
- Anna Riccioni (University of Bologna, Italy)
- Felipe Rudge Barbosa (University of Campinas, Brazil)
- Anthony Rutkowski (Georgia Institute of Technology, USA)
- Chiara Sammarco (University "Mediterranea" of Reggio Calabria, Italy)
- Alessandro Santiago dos Santos (Institute for Technological Research - IPT, Brazil)
- Diego Santos (São Paulo Federal Institute of Education, Science and Technology, Brazil)
- Ulrich Schoen (Nokia Siemens Networks, Germany)
- Florian Schreiner (Fraunhofer Institute FOKUS, Germany)
- Marko Schuba (RWTH Aachen University, Germany)
- Alexander Semenov (University of Jyväskylä, Finland)
- DongBack Seo (University of Groningen, The Netherlands)
- Jun-Bae Seo (University of British Columbia, Canada)
- Bartomeu Serra (University of the Balearic Islands, Spain)
- Riaz A. Shaikh (Université du Québec en Outaouais, Canada)
- Mostafa Hashem Sherif (AT&T, USA)
- Marek Sikora (AGH University of Science and Technology, Poland)
- Pierre Siohan (France Telecom, France)
- Lingyang Song (Peking University, China)
- David Stezenbach (University of Vienna, Austria)
- Ken-Ichi Suzuki (NTT Corporation, Japan)
- Szymon Szott (AGH University of Science and Technology, Poland)
- Kenzo Takahashi (University of Electro-Communications, Japan)
- Andrea Tonello (University of Udine, Italy)
- Ualsher Tukeyev (Al-Farabi Kazakh National University, Kazakhstan)
- Klaus Turowski (University of Augsburg, Germany)
- Kurt Tutschku (University of Vienna, Austria)
- Hiromi Ueda (Tokyo University of Technology, Japan)
- Hitoshi Uematsu (NTT Corporation, Japan)
- Mehmet Ulema (Manhattan College, USA)
- Manuel Uruña (Universidad Carlos III de Madrid, Spain)
- Geerten van de Kaa (University of Delft, The Netherlands)
- Jari Veijalainen (University of Jyväskylä, Finland)
- Fabio Violaro (University of Campinas, Brazil)
- John Visser (Canada)

- Hendrik vom Lehn (RWTH Aachen University, Germany)
- Michal Wagrowski (AGH University of Science and Technology, Poland)
- Marc Waldman (Manhattan College, USA)
- Nayer M. Wanas (Electronics Research Institute, Egypt)
- Klaus Wehrle (RWTH Aachen University, Germany)
- Hanno Wirtz (RWTH Aachen University, Germany)
- Norifumi Yamaguchi (Radiocommunication Bureau, ITU)
- Wilson Yamaguti (National Institute for Space Research, Brazil)
- Rachid Zagrouba (University of Manouba, Tunisia)
- Ahmed H. Zahran (Cairo University, Egypt)

KEYNOTE SUMMARIES

2020: THE UBIQUITOUS HETEROGENEOUS NETWORK - BEYOND 4G

Rufus Andrew

Managing Director, Nokia Siemens Networks, South Africa

The demand for mobile broadband communication continues to increase exponentially, fuelled largely by the proliferation of smarter and smarter devices. Smartphones, superphones, and tablets with powerful multimedia capabilities and applications are becoming increasingly popular and are creating new demands on mobile broadband.

This trend is expected to increase in momentum over the next decade and will be further fuelled by the arrival of billions of billions of machine devices and related machine-to-machine (M2M) applications and human to machine communications. As the interfaces and capabilities of mobile devices start accommodating broader radio-frequency (RF) capabilities and technologies such as RFID (radio-frequency identification), mobile networks will play an increasing role in the backhauling and fronthauling of signals that will be connected to the 'Internet of Things'

These factors are adding up to create an exponential increase in traffic volumes, number of transactions, and complexity of processing. Extrapolations of current growth trends predict that networks need to be prepared to support up to a thousand-fold increase in total mobile broadband traffic by 2020. This figure assumes a ten-fold increase in broadband mobile subscriptions and up to 100 times higher traffic per user

To meet such demand, the industry needs to look ahead of today's technology horizon to create 'Beyond 4G' Networks. In order to successfully achieve this, industry collaboration is needed to deliver on the following requirements.

1. Ten times more spectrum will need to become available for mobile broadband
2. Networks will need to use spectrum ten times more efficiently than existing mobile broadband technologies
3. Networks will require ten times more base stations

This presentation looks in detail at these requirements with possible solutions for the 'Beyond 4G' Network.

Finally we look at the impact of the 'Beyond 4G' Technologies bridging or further entrenching the Digital Divide.

RECOVERY FROM THE COMMUNICATION DISTURBANCE BY THE TOHOKU EARTHQUAKE AND ACTIONS TOWARD THE FUTURE

Hirofumi Horikoshi

General Manager, Technology Planning Department, Nippon Telegraph and Telephone Corporation, Japan

The Tohoku Earthquake and the following tsunami on 11 March 2011 disabled significant parts of NTT’s communication infrastructure in 385 central offices, 4,900 base stations of our cellular network and 150 million telephone circuits. This presentation describes: (1) NTT’s efforts for the recovery of its network, (2) how it kept communication services in the disaster regions, (3) how it provided livelihood support to the victims, and (4) possible future plans.

Despite many aftershocks, some of them over a magnitude of seven, as shown in Fig. 1, the communication system had recovered by the end of April, with the exception of 18 base stations. Even in the areas surrounding the nuclear power plants, people could communicate with each other. The major means for recovery were (i) new constructions or replacement of power supplies as well as communication facilities in central offices, (ii) reinstallations of transmission paths, (iii) repairs of entrance circuits for base stations, (iv) expansions of one base station’s coverage area, etc.

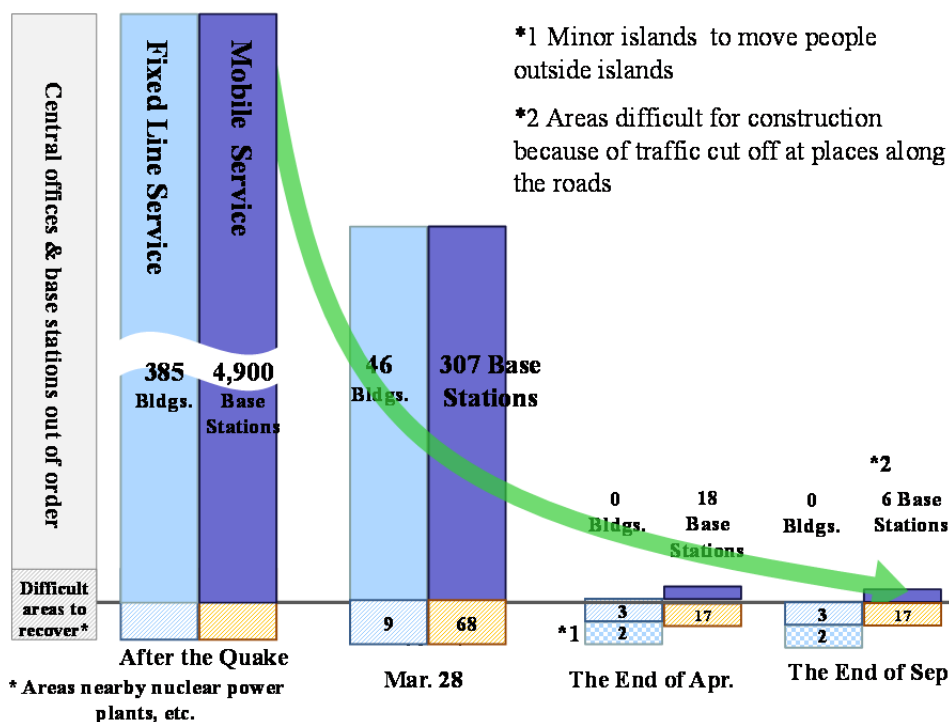


Fig. 1 Recovery of the Communication Systems from the Disaster

As part of the recovery effort, 3,900 public telephone booths with portable satellite links, 31 automobiles equipped with base station, 900 satellite cell-phones, 2,100 mobile phones and 410 battery charge booths were provided free of charge. In addition, 410 free internet corners, free WiFi services as well as 670 free tablet terminals were also provided for internet connectivity. A Web page was launched to show the map of recovered regions. Labor-intensive operations were also carried out. For the people in evacuation centers, we sent messengers to receive information about their well-being and to transfer these messages to their relatives, the mass media, and the Web page.

Several actions are planned toward a more fault-resilient network infrastructure. The first one is to make the power supply of base station uninterruptible. For this, 1,900 base stations will have long life batteries such as 24 hours, which covers 65% of the population. The second is to construct 100 large zone base stations in the densely-populated areas, which covers 35% of the population of Japan. The third is to increase the number of physical routes for each trunk circuit, to introduce the distribution of network functions, to improve waterproof central offices, etc.

From the lessons learned of this disaster, we will introduce a new voice message service as shown in Fig.2, which sends a message to the destination as a recorded file to avoid the congestion of telephone service. It is very useful to send a message as “I’m alive!” to our relatives after a disaster, and this is expected to be one of the most important life saving tools.

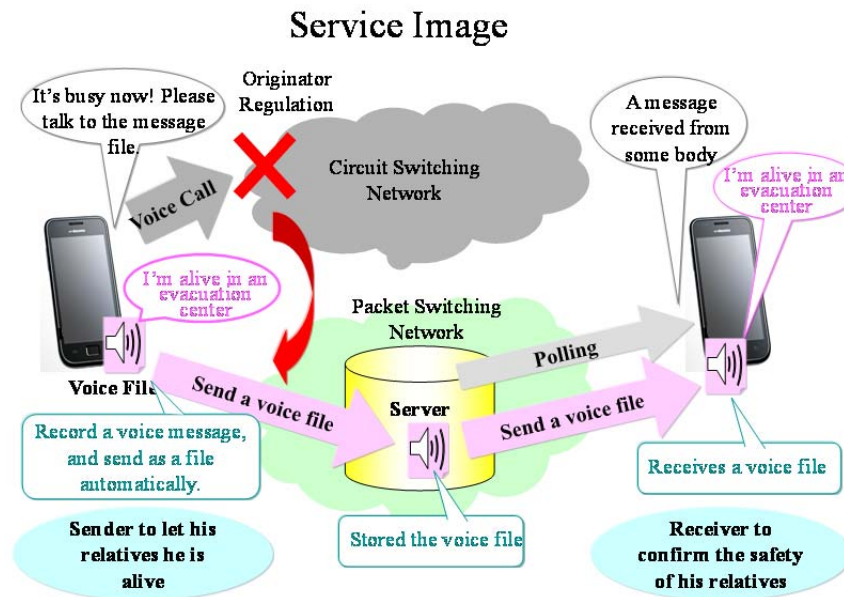


Fig.2 Voice Mail System suitable to communications at a disaster

**THE FULLY NETWORKED HUMAN-ITY? –
INNOVATIONS FOR THE UN-NETWORKED HUMAN**

Alfredo Terzoli

Professor, Rhodes University, South Africa

There is a growing perception that the many divides that fracture humankind are problematic. Divides do generate instability (to the point of revolutions or wars) and have obvious inefficiencies, deriving from the fact that, as humans our (collective or social) strength is in the group, not in the individual.

One such divide is between those that are connected to the global information cloud, the Internet or those that are not. On this divide, often referred to as 'digital divide', millions of words have been spent, but without any major concrete change: the problem is still with us, no matter how clichéd it sounds by now.

Interestingly enough, this divide runs along others, such as the one between the rich and the poor, and the educated and the uneducated. Maybe bridging one divide can help bridging the others? What type of innovations do we need to start bridging the digital divide for real - and so potentially other divides? What systems should be in place, what standards, what practices? This talk will try to answer these questions focusing on an ongoing experiment in Dwesa (Wild Coast, South Africa), the Siyakhula Living Lab, at the centre of a rich innovation ecosystem.

SESSION 1

ICTS HELPING AFRICA

- S1.1 The Role Of ICTs In Quantifying The Severity and Duration Of Climatic Variations - Kenya's Case
- S1.2 ICT use in South African Microenterprises: An assessment of Livelihood outcomes
- S1.3 SM²: Solar Monitoring System in Malawi

THE ROLE OF ICTS IN QUANTIFYING THE SEVERITY AND DURATION OF CLIMATIC VARIATIONS – KENYA’S CASE

Muthoni Masinde¹; Antoine Bagula²

¹ Hasso Plattner ICT4D Research School, University of Cape Town, mthonimasinde@yahoo.com

² ISAT Laboratory, Department of Computer Science, University of Cape Town, bagula@cs.uct.ac.za

ABSTRACT

For the last 2 decades, Kenya has consistently contributed the highest number of people affected by natural disasters in Africa. This is especially so for disasters triggered by climatic variations. The Kenya Meteorological Department has provided regular weather forecasts since the 60s. One of the shortcomings of this Department’s approach is the fact that their forecasts provide conceptual indications of droughts/floods without giving operational indicators. This makes it difficult for key stakeholders to develop solid strategic plans. Innovative use of ICTs can turn around this situation by realigning the forecasts to aid in answering questions such like, how long and how severe the predicted climatic variations will be. Use of cheaper wireless sensors can also help readdress the current poor coverage by weather stations. Based on analysis of 31 years of historical daily precipitation data from three weather stations, we prove that the Effective Drought Index can be used to quantify droughts/floods. We also present an effective web-based system that policy makers can use to monitor droughts/floods on daily basis. In the discussion, we explain how an on-going initiative aimed at integrating wireless sensor networks and mobile phones will further improve drought monitoring.

Keywords— Drought quantification, Effective Drought Index, Kenya’s weather forecast

1. INTRODUCTION

As Abdishakur Othowai put it [1], “Weather is upside down in Kenya”. He was referring to the noticeable climatic changes that have greatly undermined the traditional seasons that Kenyans have always relied on. Analysis of precipitation data for years 1979 to 2009 from three weather stations in Kenya attests to this fact. Rains no longer fall when they are expected; for instance, the March-April-May rains generally begin by mid-March but in the current year (2011), this did not happen until early May for most regions. The trend in Kenya is that when climate-related disasters strike, the Government leaves (like in many developing countries) relief operations in the hands of international agencies. This leads to the never-ending concern; when will Kenya build her own climatic-variations early warning systems to ensure effective and timely actions? The government has already mounted various initiatives towards addressing this, a good example being the creation of a Drought Management Authority (DMA) and a National Drought Contingency Fund (NDCF). Part of the answer can also be found in the adoption of innovative intervention approaches build around complimentary technologies such as mobile phone and wireless sensor networks. Intelligent drought prediction algorithms and perhaps invoking indigenous knowledge that Kenyan communities used in the yestercentury is

yet another way of achieving this. A related study (by the authors) is being carried out to find out if integrating wireless sensors/mobile phones with indigenous knowledge will improve predictions of droughts/floods.

Effective Drought Index (EDI) [2] is able to quantify droughts in absolute terms and also provide answers to: (1) the when, (2) the how long (onset to termination) as well as (3) the severity of droughts/floods. In this research, we demonstrate how this was possible, especially in quantifying the devastating drought that occurred 1983 to 1985 and the revenging floods of 1997-1998. EDI quantifies droughts in terms of droughts classes composed of positive and negative real values; for example, -2.50 indicates extreme drought, +3.28 indicates extreme floods and 0.98 indicates close to normal wetness. EDI further qualifies climatic/weather variations by providing Available Water Resource Index (AWRI) that can for example reliably inform a farmer of the amount of water in the soil at any given day.

In this paper, a user-friendly website was created to show how presenting daily precipitation values side by side with drought classes and AWRI values can aid in decision making process.

The rest of the paper is structured as follows: Section 2 gives the relevant definitions of drought and drought indices and describes EDI in details. Section 3 gives details of Kenya’s drought situation, weather forecasting and the existing gaps. The core contribution of this paper is described in sections 4, 5, 6 and 7 in terms of the methodology, results, the web-based system and the role of WSNs in bridging the gap resulting from sparse weather stations. Finally, the Conclusion and Further Work is presented in Section 8.

2. DROUGHTS AND DROUGHT INDICES

2.1. Definition and Types of Droughts

Drought is both a hazard and a disaster that can be classified together with earthquakes, epidemics, and floods. According to [3], drought qualifies as a hazard because it is a natural accident of unpredictable occurrence but of recognizable recurrence. As a disaster, drought corresponds to the failure of the precipitation regime, causing the disruption of the water supply to the natural and agricultural ecosystems. There is no one universally accepted definition of drought yet. Palmer[4] came to this conclusion as early as 1965 when he stated "drought means various things to various people depending on their specific interest". Since then, attempts have been made to define the term drought. The common element in all these definitions is “precipitation deficiency” which depending on how long and the intensity in turn affects soil moisture, streams, groundwater, ecosystems and human beings.

Authors have defined these effects as five ‘types’ of droughts ([5], ([6], [7] and [8]); namely *meteorological, hydrological, ground water, agricultural and socio-economic* droughts.

2.2. Drought Indices

There are several well-developed indices for quantifying effects of droughts in terms of parameters such as intensity, duration, severity and spatial extent. These indices further map the droughts to different time scales (daily, weekly, monthly, annually etc) and geographical regions to aid planning and decision-making process. One common feature among the indices is the consideration of precipitation as a contributing factor [8]. In [9], the following two broad categories of drought indices are presented:

- (a) Those that utilize an insufficient level of soil Moisture; e.g. Palmer Drought Severity Index (PDSI) [4]
- (b) Those using distribution of rainfall insufficiency; e.g. Standardised Precipitation Index (SPI)

2.3. Effective Drought Index

EDI is the drought Index that was applied in this research. It was developed by Byun and Wilhite [2] to address weaknesses they identified in the existing (at the time of their research) drought indices. Five of these weaknesses are:

- (a) They assess deficient of water for some predefined duration instead of consecutive occurrences. It is the consecutive occurrences that define drought severity;
- (b) Time unit of assessment is too long (a week, a month or a longer time period); none used daily unit yet a day’s rainfall can have great impact on drought;
- (c) They do not consider the two causes of drought; (1)shortage of soil moisture (recent deficiency of precipitation) and (2)shortage of water stored in the reservoirs (much longer deficiency of precipitation)
- (d) They made use of simple summation of precipitation in determining current shortage. This ignores the fact that the diminishing of water over time is a function of runoff and evapotranspiration.
- (e) Values of some parameters (such as soil moisture and evapotranspiration) used in the calculation of the indices are difficult to measure and often estimated. Furthermore, most of these are directly linked to precipitation and cannot be isolated from the latter. Precipitation alone is therefore adequate, easy to measure and their values are available for the longest period in any weather station in the world. Further, some parameters used may be caused/ accelerated/ affected by human activity making it even more difficult to approximate.

In EDI, daily precipitation height values and Effective Precipitation (EP) are used to compute deficiency or surplus of water resources for a particular date and place. EP is the summed value of daily precipitation with a time-dependant reduction function represented by the following equation:

$$EP_i = \sum_{n=1}^i \left[\frac{\left(\sum_{m=1}^n P_m \right)}{n} \right] \tag{1}$$

where P_m is the precipitation of m days before and the index i represents the duration of summation (DS) in days. Here $i=365$ is used, that is, summation for a year which is the most dominant precipitation cycle worldwide. The 365 can then be a representative value of the total water resources available or stored

for a long time. To be able to relate the EPs of a given weather station with climatological data, they are averaged along the day number (i.e., by calendar day) and used to compute the following:

Table 1. Characteristics of indices associated with EDI

Name	Calculation	Meaning
Mean of EP (MEP)	30-yr mean of EP for each calendar day	Climatological mean of water quantity
Deviation of EP (DEP)	DEP = EP - MEP	From climatological mean, the deficit of water quantity
Standardized value of DEP (SEP)	SEP = DEP/ST(EP)	From climatological mean, standardized deficit of water quantity
Consecutive days of negative SEP (CNS)	Consecutive days of negative SEP	Shows how long precipitation has been in deficit
Precipitation needed for a return to normal (PRN)	Calculated using Equation (2)	Precipitation needed for a return to normal conditions.
Effective Drought Index	Calculated using Equation (3)	The standardized deficit or surplus of stored water quantity.

$$PRN_j = \frac{DEP_j}{\sum_{N=1}^j \left(\frac{1}{N} \right)} \tag{2}$$

$$EDI_j = \frac{PRN_j}{ST(EDP_j)} \tag{3}$$

The EDI expresses the standardized deficit or surplus of stored water quantity. The EDI enables one location's drought severity to be compared to another location's, regardless of climatic differences. Depending on the value of EDI, the following drought classes are used:

- Extreme drought $EDI \leq -2.0$
- Severe drought $-2.0 \leq EDI \leq -1.5$
- Moderate drought $1.5 \leq EDI \leq -1.0$
- Near normal $1.0 \leq EDI \leq 1.0$

EDI has produced satisfactory results in measuring drought severity and was recently (2009) adopted (and adapted) to analyse 200-year drought climatology of Seoul, Korea[9]. Another example is the SPATial and Time Series Information Modeling (SPATSIM) which is a comprehensive software package developed at the Institute for Water Research (IWR), Rhodes University, South Africa. It calculates, displays, spatially plots, and exports/imports a variety of drought indices from rainfall time series data. EDI is one of the indices included but SPATSIM calculates monthly as opposed to the original daily effective precipitation (<http://www.ru.ac.za/institutes/iwr/>).

3. DROUGHTS IN KENYA

3.1. Overview of droughts

In Kenya, droughts are the most common disasters resulting from variations in weather/ climate, but there are occasional floods too. For instance, Kenya is among the countries at the Horn of Africa that are currently (August 2011) experiencing a devastating drought that has been described as “the worst drought in 60

years”(<http://www.bbc.co.uk/news/world-africa-13944550>). Such disasters cannot be avoided but can be managed through effective early warning systems. When they occur, droughts affect more than 25% of Kenya’s population not mentioning the ripple effects such as inadequate hydro-electric power supply, increased commodity prices and loss of jobs just to mention a few. The level of preparedness is determined by how well the disasters are defined and their characteristics quantified. In the period 1999 to 2008, Kenya contributed a whopping 32.85% of people affected by natural disasters in the Continent (Africa); Ethiopia was following by far at 15.17%[1]. Apart from the 1983-1985 drought and the 1997-1998 floods that are focus of this paper, more droughts occurred in some periods in 1980, 1982, 1999, 2000, 2001, 2003 and whole of 2008 to 2009. The 2008-2009 drought had in particular very catastrophic effects on the livestock-keeping communities.

3.2. Kenya’s Rainfall Seasons

Kenya has two main rainy seasons: (1)October-November-December (OND); and (2)March-April-May (MAM) - this is the main season. Sometimes the rainfall may occur in the period June-July-August (JJA). Among other factors (not part of this research paper), the amount of rain in each of the above season is used to classify Kenya into 12 Climatic Zones (<http://www.meteo.go.ke>). Using data completeness as the main criteria, 3 stations were selected for the current research: two in Zone 8 (Dagoretti and Embu) and one in Zone 7 (Makindu)

3.3. Weather Forecasting– Current Practice

Currently, monitoring of climatic/weather variations in Kenya is the mandate of the Kenya Meteorological Department. The Department is handicapped in the sense that all they have are sparsely distributed weather stations. They run 3 main types of stations that are currently managed by the Climatological Section of the Department (<http://www.meteo.go.ke>):





- (a) 700 rainfall stations; there were 2,000 of such in 1977; this figure drastically dropped to 1,653 by 1988 and down to 1,497 by 1990;
- (b) 62 temperature stations; and
- (c) 27 synoptic stations; these are used to observe and record all the surface meteorological data; rainfall, temperature, wind speed and direction, relative humidity, solar radiation, clouds, atmospheric pressure, sun shine hours, evaporation and visibility

The Agrometeorological Section (of the Kenya Meteorological Department) on the other hand manages 13 stations related to agriculture; data is remitted from these stations every 10 days. Apart from the normal meteorological observations, other observations by the Agrometeorological Section include: soil temperature, sunshine duration, radiation, pan evaporation and Potential Evapotranspiration. All this data is stored in semi-automated formats at the Department’s Head Quarters in Nairobi. The data is available to interested stakeholders at a fee and on request. The data used in this research was obtained from this Department.

The Meteorological Department uses the data collected to provide five main types of forecasts:

- (a) Daily forecast for main cities/towns in Kenya. This is provided in the format shown in figure 1.

Table 2. Format of Daily Forecasts in Kenya

Weather Forecast for 5 th September 2011				
Town	Morning	Afternoon	Sunrise	Sunset
Nairobi	 Cool cloudy/Light rains/Sunny intervals	 Sunny intervals/Showers	6:28	18:34
Mombasa	 Sunny intervals	 Sunny intervals	6:18	18:21
...

- (b) Four-Day forecast
- (c) Seven-Day Forecast
- (d) Monthly Forecast
- (e) Seasonal Forecast

The Four-Day, Seven-Day and Monthly forecasts are in form of downloadable reports summarizing the recent past, current and near-future weather patterns in conceptual terms. An example extract from such forecasts is shown in the figure below

Weather highlights

Review of weather for the last seven days period (25th to 31st August 2011)

- **During the review period (25th to 31st August 2011):**
- Wet weather conditions were maintained over the western highlands, Lake Basin and parts of Central Rift Valley. There was, however, a slight reduction of the rainfall activities over these areas during the second half of the review period. Occasional light rainfall was also experienced over the central highlands and Nairobi area during the review period. The coastal strip on the other hand experienced isolated rainfall activities during the second half of the review period. Elsewhere in the country, dry conditions persisted.
- Day-time (maximum) temperatures generally decreased over much of the country except over the north-eastern and the coastal strip.
- Minimum temperatures increased over the whole country except over few parts of western and central regions.

Forecast for the next seven days (2nd – 8th Sept 2011):

- The forecast for the next seven-day period (2nd to 8th of September 2011) indicates:
- Maintenance of wet weather activities over the western sector and central Rift Valley which may occasionally spread to the central highlands and Nairobi area;
- A slight reduction in rainfall activities over the coastal strip; and
- Relatively warmer temperatures are likely to be experienced over most parts of the country during the forecast period compared to the previous period.

Figure 1. Extract from a Seven-Day Forecasts report

The season forecasts are more detailed and they are also provided in Kiswahili language. The Department also works hand in hand with both the print and electronic media to disseminate the forecasts. All the above forecasts are freely accessible from the Department’s website.

3.4. Weather Forecasting – The Gaps

The following three main gaps were identified

3.4.1. Non-User Centered Weather Forecast Information

The usefulness of forecast information provided by the Kenya’s Meteorological Department to key stakeholders especially the farmers and policy formulators is rather wanting. It would for example be desirable if the Department could inform the relevant government ministries the actual implications (in operational quantifiable terms) of weather observations. It is not enough to report that; “60.4mm of rainfall was recorded in Kakamega”; instead, a report saying, “60.4mm of rainfall that was recorded in Kakamega raised the available water resource to above the normal for this area. It is predicted that the rains will continue for a week and this will lead to severe floods in the low-lying areas of the Kakamega”. The latter is more useful and can be used by policy makers to mount rescue operations. Proper use of EDI can fix this gap

3.4.2. Ineffective Information Dissemination

The channels that are used to disseminate the forecast information are ineffective; the farmers that need it most do not get it and those that do, cannot comprehend the information. A related study (by the authors) aims to fix this problem by use of natural language and text-to-speech tools to provide custom forecast information to farmers via mobile phones.

3.4.3. Poor Coverage by Weather Stations

The third gap is as a result of the small number of weather stations installed. The existing weather stations are far too few for the vast 582,650km² that is Kenya’s geographical area. This makes it difficult to get the micro-level weather indicators that are necessary for effective forecasts. This number may be increased through the adoption of cheaper and more automated wireless sensors based weather stations.

4. QUANTIFYING KENYA’S DROUGHTS/ FLOOD

4.1. Data Used

Daily precipitation data for years 1979 to 2009 for Dagoretti, Embu and Makindu weather stations was used. This translated into 3 X 365 X 31 records.

Table 3. Geo-Data of the Weather Stations Studied

Name	Dagoretti	Embu	Makindu
WMO #	63741	63720	63766
ICAO	HKNC	HKEM	HKMU
Year Opened	1954	1975	1904
Latitude	01 18S	00 30S	2 17S
Longitude	36 45E	37 27E	37 50E
Elevation	1798 m	1493m	1000m

Two Phases; (1) Data Cleaning; and (2) Data Analysis were then carried out as per the flow chart shown in below.

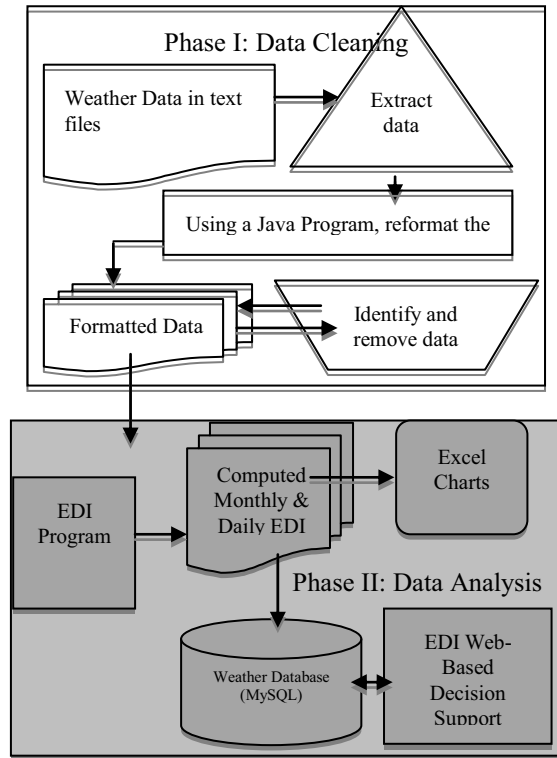


Figure 2. Data Cleaning and Analysis Flow Chart

Though data for other observations (temperature, humidity, pressure etc) and stations (Voi, Mombasa, etc) was available, only precipitation data for the 3 stations was used in this research. Raw data (in form of text files) was first processed manually to identify any data gaps and anomalies, then a Java program was written to covert the original format to more friendly formats acceptable by the EDI FORTRAN program (http://atmos.pknu.ac.kr/~intra2/down_src.php) that was used. This Fortran program uses equations (1), (2) and (3) to compute daily/monthly EDIs and outputs them in form of text files.

4.2. EDI Computation

Two versions of the EDI program were used:

4.2.1. Monthly EDI

This was used to calculate monthly EDI using monthly precipitation totals for 31 years (1979 to 2009). For each of the 3 weather stations, an Input File with the following format was processed using the EDI program.

Table 4. Monthly EDI Input File Format

Year	Month	Total Precipitation
1979	1	70.80
...
1979	12	68.80
1980	1	39.30
...
2009	12	121.10

Using modified (to replace days with months) equation (1), EDI computes the ‘normal’ monthly precipitation for each calendar month of the year based on the mean of precipitation of this

calendar month for all the 31 (of historical data) years. For example,

Normal precipitation for November is the mean of the precipitations for November 1979, November 1980, November 1980, ..., November 2009.

Obviously, a ‘normal’ precipitation for November in Embu may not be compared with the ‘normal’ precipitation for November for Makindu; the two are in different climatic zones. The EDI values solve this by use of standard deviation to compute universal real values that have similar meaning in all climatic zones. Therefore, an EDI value of -2.59 will represent ‘Extreme Drought’ irrespective of climatic zones. For each stations, an Output File was generated; each showing the Effective Drought Index (EDI) and Available Water Resource Index(AWRI) calculations for 30 years (1980 to 2009)

Table 5. Monthly EDI Output File Format

Date	Total Precipitation	AWRI	EDI
1980-01-15	70.80	230.0	-0.51
1980-04-15	105.9	229.8	-1.35
...
1980-12-15	33.5	460.8	1.05
1981-05-15	245.2	818.2	1.96
...
2009-12-15	121.10	297.5	-0.27

EDI applies a time-dependent reduction function in computing the monthly/daily water deficiency. This is to cater for runoff and evapotranspiration that progressively reduces soil moisture over time. In table 5for example,

- Available Water Resource Index for December 1980 was 460.8.
- A total of 14.9mm of precipitation was received in January 1981
- Simple Summation of monthly precipitation would have yielded 475.7
- But factoring in runoff and evapotranspiration EDI yielded 364.2

The value of EDI is determined by the climatic conditions of a given zone and period/season. For example;

- In March 1980, the AWRI for Dagoretti was 170.1. This translated to a Drought-Near Normal of -0.86
- In April 1980, the AWRI for same station was 229.8. This was equal to a moderate tending to severe drought (worse) of -1.35.

The reason for this is that April is generally a wet month; in a normal March-April-May rain season, rains will have been falling for a whole month and therefore April is normally a wet month.

In order to visualize drought in terms of classes described in [2], the following adapted classification was used:

Extremely Flood	EDI > 2
Severe Flood	1.5 > EDI < 1.99
Moderate Flood	1 > EDI < 1.49
Wet-Near Normal	0.01 < EDI > 0.99
Drought-Near Normal	-0.99 < EDI > 0.00
Moderate Drought	-1 < EDI > -1.49
Severe Drought	-1.5 < EDI > -1.99
Extreme Drought	EDI < -2

Colour-coding ranging from dark-red (representing ‘Extreme Drought’) and (dark-blue representing ‘Extreme-Floods’) was

then used to represent drought classes. The view was first presented in Excel before being migrated to the web based system

4.2.2. Daily EDI

Three weather stations were used; the Input File had the following format:

Table 6. Daily EDI Input File Format

Year	Month	Day	Total Precipitation
1979	1	1	0
1979	1	2	0
2009	12	31	0

Using equation (1) EDI computes the normal daily precipitation for each calendar day of the year based on the mean of precipitation of this calendar day for all the 31 (of historical data) years. Like in the case of monthly EDI, the ‘normal’ precipitation for March 28 is the mean of the precipitations for March 28 1979, March 28 1980, March 28 1980, ..., March 28 2009. Similar to Monthly EDI, an Output File for each of the stations with the following format was generated

Table 7. Daily EDI Output File Format

Date	Total Precipitation	AWRI	EDI
1980-01-01	0.0	117.0	-0.66
...
1981-01-31	0.0	208.1	0.96
...
2009-12-31	26.6	146.7	-0.20

5. RESULTS AND DISCUSSION

5.1. The 1983-1985 Drought

Drought was experienced in all the three regions with Dagoretti leading with an average of -1.06 (compared to -0.49 and -0.33 for Embu and Makindu respectively). Drought was worse in the November 1983 to November 1984 period. The graphs for Dagoretti and Embu have similar patterns; this is not surprising given that the two are classified under the same Climatic Zone (Zone 8); Makindu is classified under Zone 7. Interestingly, Embu rainfall had an abnormal spike during the October-November 1983(EDI of +0.36); normal rains were received while other stations had close to negative values.

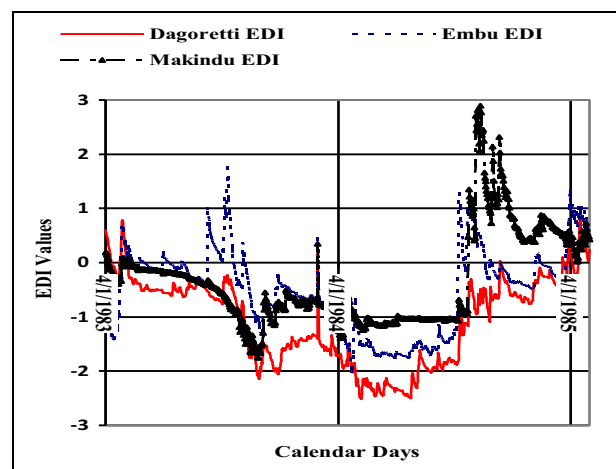


Figure 3. 83-85 Drought Onset/Termination/Severity

The 1997-1998 Floods

The October-November-December 1997 torrential rains triggered the floods. The March-April-May 1998 rains later worsened this. At some point, the floods were so severe; e.g., over 100mm of rainfall was recorded on 11 February 1998 in Makindu; February is generally a dry month so this put the EDI at a catastrophic value of +5.27. Similarly, a way above normal rainfall measuring 167.7mm was recorded in Dagoretti on 16 March 1998. Given that it had been raining and the ARWI was already at 303.3, this rainfall just turned things inside out and resulted in a dangerous EDI value of +4.85. As it was the case with the 1983-1985 drought, the flood pattern for Embu and Dagoretti are similar. In this case too, Embu had unique spike where the flood reduced during the March-May 1998 season.

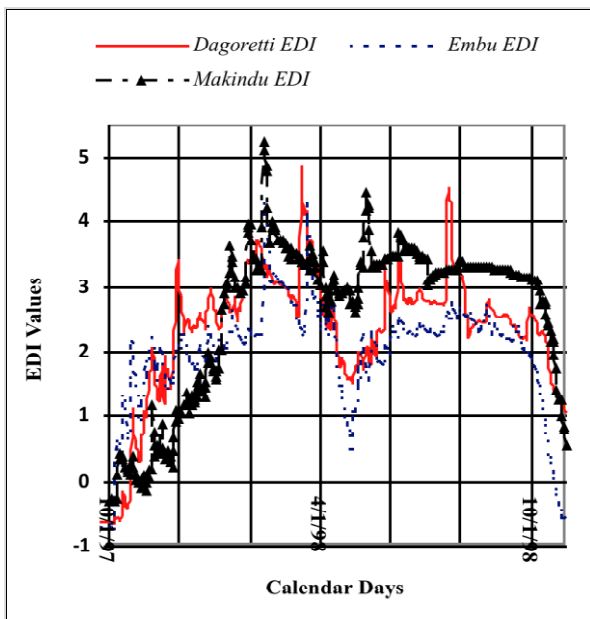


Figure 4. 97-98 Flood Onset/Termination/Severity

6. EDI WEB-BASED DECISION SUPPORT SYSTEM

6.1. System Overview

Using the daily precipitation, computed EDI/AWRI and Drought Classification, a web-based system was developed using PHP, MYSQL, JavaScripts and Apache to demonstrate how EDI can aid in detecting and preparing for droughts/floods. Jpgraph-3.5 was used to draw the dynamic (database driven) graphs. For example, presented with daily EDI values that are constantly rising (positive) and a 7 day weather prediction that shows steady increase in the amount of rainfall (not part of the system developed in this research), a decision maker can absolutely know when and how severe the coming floods will be. Figure 5 shows the main screen of the web-based system; it allows users to select a weather station (from either Dagoretti, Embu or Makindu), year and month. The user can also opt to choose all years and/or month.

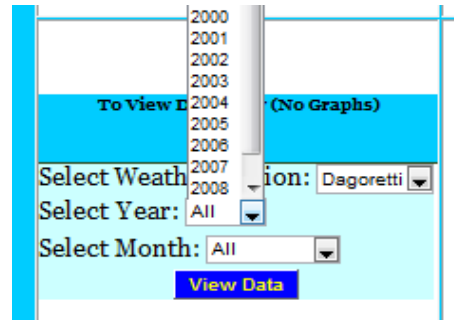


Figure 5. Main Screen - Selecting Parameters

6.2. Data Views

6.2.1 One Month, One Year, One Station Data

Once the user has made selections and clicks 'View Data' (for example, selecting Makindu, 2009 and May), the daily EDI values, AWRI Values and the Drought Classes are displayed as shown in figure 6. In this case, May 2009, was a drought (Severe and Moderate) month. On the right-hand(part) of the screen, hints, definitions and explanations of the data are provided to make it easier for the user to make interpretations/decisions.

Weather Readings for Makindu [May, 2009]				
Date	Precipitation	AWRI	EDI	Drought Class
1-May-2009	0	43.1	-1.59	Severe Drought
2-May-2009	11	53.1	-1.37	Moderate Drought
3-May-2009	0	50.6	-1.44	Moderate Drought
4-May-2009	0	49.1	-1.52	Severe Drought
5-May-2009	0	47.8	-1.51	Severe Drought
6-May-2009	2.1	48.9	-1.51	Severe Drought
7-May-2009	0	47.6	-1.54	Severe Drought
8-May-2009	0	46.6	-1.55	Severe Drought

Figure 6. One Station/Year/Month Data view Format

6.2.2. Multiple Years/Months

Views are also possible for a given month of all the years (say March 1980, March 1981,..., March 2009). These are displayed when a user selects one month (say March) and 'All' in the Year option. Similarly, selecting 'All' for month option and a given year (say 1998) will display daily values for January to December of the chosen year

6.2.3. Graphical View

From the view shown in figure 6, users can opt to view graphs of the EDI values by clicking on 'Click Here' button under 'View Graph'. These graphs are generated using *Jpgraph software* and they are useful in detecting trends that may lead to droughts/floods. By moving the mouse over bars, the user is able to see the actual value of any bar. For example, in figure 7, the value for the bar corresponding to July 21, 1998 is +4.28. Clicking on any bar also opens a new window with more interpretations of the EDI value.

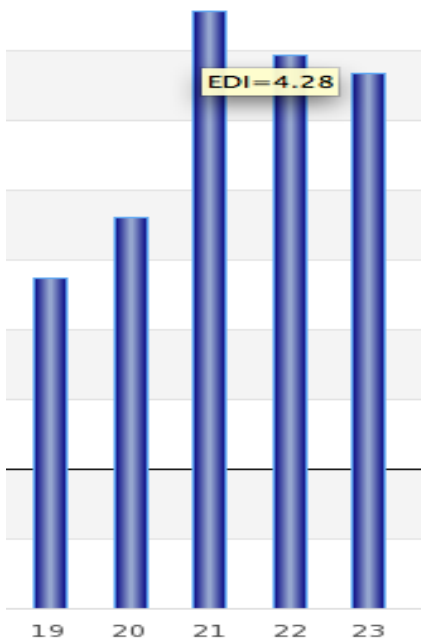


Figure 7. EDI Graph Extract for Embu, July 1998

The above graph extract shows the EDI values for 19th to 23rd July 2998 for Embu Station. It is easier to see a pattern in a graph than in a table of values. In this graph (where X-axis represents days of the month and ‘Y-axis represents the EDI values) for example, the floods were getting worse in Embu as the rains continued to pound. Plotting graphs for a whole year can also be useful in quantifying past droughts/floods.

6.3. On-Set/Termination and Severity of Drought

Using the EDI web-based system that we developed, it was possible to quantitatively and qualitatively identify that the drought that caused farmers massive losses in 2009 actually started on 29th October 2008 for both Embu and Dagoretti while it started on 23rd October 2008 in Makindu. The drought started worsening (below -1) on 15th March 2009 in Makindu, 22nd March in Embu and 10th April in Dagoretti. It subsided for 5 days (2nd to 6th November) in Embu and 5 days (18th to 22nd October) in Makindu. In general, the drought continued ravaging the country until end of 2009.

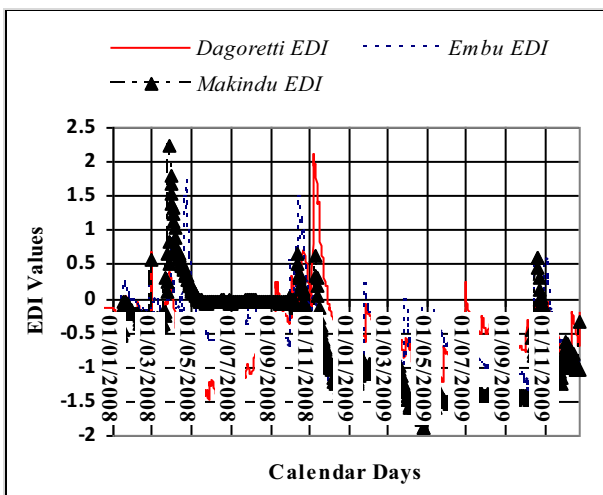


Figure 8. Onset of the 2009 Drought

With this kind of information available to stakeholders on a daily basis, solid decisions can be reached. Farmers can decide if to plant or not and even what to plant where, and when. This would save them losses that are incurred when a lot of investment is put into preparing farms that will not yield. The government can plan rescue measures in advance such as buying cattle from pastoralists while they (the cattle) are still healthy and import grains in advance. In the case of floods like the ones that occurred in 1998, reading EDI values together with weather forecast for March-April-May season would have saved human lives that were lost when Nairobi River broke its banks and carried away shanties in Mathare slums.

7. USING WSNS TO IMPROVE DROUGHT PREDICTIONS

As earlier explained in section 3.4.3, weather forecasts in Kenya is carried out by the Meteorological Department through very sparse network of weather stations. These are expensive professional weather stations that are located in only a given number of locations to represent very large geographical regions whose climate may not be homogenous. For example, Embu Weather Station represents the entire Embu County with an area of 2,818 km². The climate in Embu is so diverse that the upper areas support tea growing while the lower sides are close to desert. This sparse nature of the weather stations (and consequently the weather data) introduces some incredibility to the drought quantification discussed in this paper because the data used is too ‘course’ to have absolute meaning at the local (say a village) level. This visibility gap is currently being filled by use Wireless Sensor Networks(WSNs). WSNs are based on off-the-shelf low cost technology that provides the potential to reduce cost and allow sustainability in many developing countries where high cost of maintenance and repair are the main hindrance to projects success.

WSNs are made up of collections of relatively small, self-contained, micro-electro-mechanical devices. These tiny devices have sensors, computational processing ability, wireless receivers with transmitter technology and a power supply[10]. The sensor board (the Agricultural Sensor Board by Libellium – <http://www.libellium.com/>) that we are using for this research is capable of sensing temperature, relative humidity, atmospheric pressure, soil moisture, soil temperature, solar radiation among others. The board also comes with GPRS and GPS modules.

The sensor boards as installed currently work as peers and they individually send their readings to the MYSQL database (discussed earlier in this paper) in form of text messages. This configuration is being tested alongside an existing weather station at the University of Nairobi, Kenya. Our WSNs implementation includes temperature, relative humidity and atmospheric pressure sensors as well as an e-weather station (wind vane, anemometer and pluviometer) that is connected to the agricultural board to measure wind direction, wind speed and precipitation respectively.

8. CONCLUSIONS AND FURTHER WORK

We have successfully demonstrated that appropriate use of ICT tools (software in this case) can lead to weather forecasts that matter to all the stakeholders. By providing a software solution through which daily Effective Drought Index (EDI) can be monitored, timely decisions can be made and hence making it possible to put mitigation measures in place and consequently reducing negative impacts. In this paper, we managed to accurately quantify the 1983-1985 drought and the 1997-1998 floods in terms of when they struck, how severe they were and

how long they lasted. The research concentrated on only 3 stations in Kenya and data for 1979 to 2009. The tool presented here can be extended to cover other stations as well as the data for 2010-2011. The latter would be interesting because then it can be used today (August 2011) to aid in planning for any eventuality now that the March-April-May 2011 rain season failed in many parts of Kenya and the Horn of Africa Region is general.

The EDI web-based decision support system can further be improved by incorporating Google Maps similar to the ones used at the Drought Termination and Amelioration by National Oceanic and Atmospheric Administration at the National Climatic Data Center in the US (<http://lwf.ncdc.noaa.gov/temp-and-precip/drought/recovery.php>) Further improvements that are under way are in form of incorporating the use of mobile phones to send alerts to stakeholders as well as the use of Wireless Sensor Networks (WSNs) to improve the coverage and automation of weather data collection.

REFERENCES

- [1] The International Federation of Red Cross and Red Crescent Societies. World Disaster report 2009 – Focus on Urban Risk; www.ifrc.org/Global/Publications/disasters/WDR/WDR2009-full.pdf
- [2] Byun, H. And Wilhite, D.A., 1999. Objective Quantification Of Drought Severity And duration. *Journal of Climate*, 12(9), pp. 2747-2756.
- [3] Elsa, E., Moreira, Carlos, A., Coelho And Ana, A., Paulo, 2008. SPI-Based Drought Category Predication Using Loglinear Models. *Journal of Hydrology*, 354, pp. 116-130.
- [4] Palmer, W.C., 1965. Meteorological drought. US Department of Commerce; Weather Bureau, 45, pp. 1-58.
- [5] Huschke, R.E., 1959. Glossary of meteorology: Boston. Boston: American Meteorological Society.
- [6] Yevjevich, V., Hall, W.A. and Salas, J.D., 1977. Drought research needs, Conference on Drought Research Needs, December 12-15 1977, Colorado State University, Fort Collins, Colorado.
- [7] Rosenberg, N.J., 1979. Drought in the Great Plains-- Research on impacts and strategies, L. University Of Nebraska, ed. In: Workshop on Research in Great Plains Drought Management Strategies, March 26-28 1979, Water Resources Publications.
- [8] Mishra, A. K., and Singh, V. P. (2010), A review of drought concepts. *Journal of Hydrology* (In Press), doi:10.1016/j.jhydrol.2010.07.012
- [9] Kim, D., Hi-Ryong, B. And Ki-Seon, C., 2009. Evaluation, modification, and application of the Effective Drought Index to 200-Year drought climatology of Seoul, Korea. *Journal of Hydrology*, 378, pp. 1-12.
- [10] Eiko Yoneki and Jean Bacon, 2005; A survey of Wireless Sensor Network technologies: research trends and middleware's role; University of Cambridge, computer Laboratory

ICT USE IN SOUTH AFRICAN MICROENTERPRISES: AN ASSESSMENT OF LIVELIHOOD OUTCOMES

Frank Makoza and Wallace Chigona

Department of Information Systems
University of Cape Town
Private Bag X, Rondebosch 7701
Cape Town, South Africa.

Email: Wallace.Chigona@uct.ac.za

ABSTRACT

This paper reports on a study on the impact of using Information and Communication Technologies (ICT) on the livelihoods of microenterprises. The study used a qualitative approach and focused on the case of South Africa. Microenterprises play an important role in socio-economic development and in bridging the gap in the segments of the economy of South Africa. The study findings confirm that ICT use and support of institutions and organisations have a positive impact on the livelihoods of microenterprises. However, ICT use in microenterprises is curtailed by challenges beyond access and ownership of ICTs. Chief among these problems is lack of awareness of application of ICT in business activities and awareness of support services provided by business development organisations.

Keywords— ICT use, microenterprises, livelihoods

1. INTRODUCTION

Microenterprises play a vital role in socio-economic development in developing countries, particularly in areas of job creation, income generation and skills development among others [10, 16, 17]. In South Africa, microenterprises form a large part of the Small, Micro and Medium Enterprises (SMMEs) sector. Based on the National Small Business Amendment Bill Gazette of 2003 [25], microenterprises are defined as enterprises with less than ten employees, with a turnover of less than ZAR 0.2 Million (US\$26,500) and a net asset value of less than ZAR 0.1 Million (US\$13,300) [25]. The Act further categorises the SMMEs into medium, small, very small and micro, based on operating sectors, turnover, value of assets and number of employees. Examples of microenterprises in the South African context are hawkers, spaza shops, handcrafters and street vendors. Microenterprises face a myriad of challenges which may result in limited growth and failure [31].

There is a belief that the use of Information and Communication Technologies (ICT) may help microenterprises remain competitive and increase their survival chances [29]. There is growing empirical evidence

that the use of ICT in microenterprises may result in improved communication, reduction in operating costs and improved access to information and knowledge [23, 35]. Furthermore, use of ICT may lead to empowerment, growth and sustainability of microenterprises [18]. However, the use of ICTs amongst microenterprises, in most cases, is limited to non-sophisticated technologies [11]. Microenterprises do not use complex applications due to limited financial resources, lack of IT skills and lack of knowledge about the application of ICTs in their business activities [34]. A number of institutions do, however, support microenterprises to access and use the technology for that end.

The use of ICT, as is the case with any innovation, has potential consequences which could be intentional or unintentional. We used Sustainable Livelihood Approach (SLA) as theoretical underpinning to understand the impact of ICT use on microenterprises. Livelihoods are described as a means for living through capabilities of tangible and intangible assets [4]. The concept of livelihood may help us to understand the diversity of poverty, hence our choice of SLA for this study. Microenterprises are perceived to be a means for poverty alleviation in marginalised communities [8]. Understanding of poverty requires consideration of the diversity of livelihoods [3]. SLA takes into account wider issues for the livelihoods of households and communities, beyond technology and income poverty [3, 11, 17]. This study was guided by the research question: *How does the use of ICT affect the livelihood of microenterprises?*

We focused on South Africa which has a dual economy structure [1]. The first economy is well developed, integrated into the global supply chain and uses advanced technologies. The second economy is underdeveloped and communities in this economy experience inequalities and limited opportunities [20]. The majority of microenterprises operate in the second economy [14]. There are ongoing efforts by Government and development agencies to try to bridge the two economies, through interventions aimed at promoting SMMEs from the second economy to the first economy. The services offered by such agencies include internet access and computer support. It is, therefore, important to establish whether technology interventions

embedded in business support programs have an impact on the livelihood of microenterprises.

There is a paucity of studies on the consequences of ICT use on the livelihoods of microenterprises. The majority of studies on the impact of ICT use have focused on SMEs, but do not particularly focus on microenterprises [10, 18]. Due to the uniqueness of microenterprises, these findings may not always be applicable. Studies conducted in South Africa have focused on technology, productivity and organisation formation [14, 35]. There is still a limited understanding of ICT use in relation to social factors in the context of micro organisations. The main objective of this study was to analyse how microenterprises use ICT to utilise assets and structures, as well as the effects on livelihood outcomes. The findings of this study will serve to inform policy makers, especially those involved in business development interventions supporting microenterprises, on a wider scope of issues that affect the livelihoods of microenterprises.

2. LITERATURE REVIEW

2.1. SMMEs in the context of South Africa

As already stated, the South African economic landscape is segmented into two tiers, with different levels of economic opportunities. The majority of microenterprises operate in the second economy, where there is an abundance of unskilled labour that does not meet the requirements of the first economy [14].

The role of microenterprises in the economy is not well documented. As a result, there is less clarity on the extent to which microenterprises contribute to poverty alleviation and economic growth of the country [1]. That withstanding, there is growing anecdotal evidence about the role of microenterprises in economic development [10, 18]. They act as sources of employment, sources of income, seedbeds for skills development; for poverty alleviation, self-empowerment and sustainability of marginalised households and communities [14, 26].

In striving to make the economic ecosystem more productive, to support creation and growth of SMMEs and to achieve sustainable production, the government has introduced a range of policies, legislations, regulations and initiatives for supporting SMMEs. The initiatives aim at increasing the number of SMMEs graduating from the second economy to the first economy, as well as trying to address the high rates of failure of SMMEs (see Table 1) [31].

Table 1. South Africa SMMEs promotion initiatives

Initiative	Focus	Main objective(s)
Accelerated and Shared Growth for South Africa (AsgiSA)	National	Support SMMEs business start-up through capital financing.
Nsika Enterprise Promotion Agency	National	Support SMMEs businesses through non-financial services.
Small Enterprise Development Agency	National	Promote and support SMMEs through training, funding and business advice.
South Africa Micro-Finance Apex Fund (SAMAF)	National	Provide support for access to finance for microenterprises
Khula	National	Support SMMEs in accessing financial services.
Real Enterprise Development (RED) Door Program	Provincial	Small business formation support through funding and training.

Microenterprises are bedevilled by a wide range of problems which may result in limited growth of SMMEs and high failure rates [23, 31]. As summarised in Table 2, the challenges can be categorised into resources, capacity, regulations, support and health issues.

Table 2. Challenges besetting microenterprises

Category	Example of challenges	Citation
Resources	Less income, lack of collateral, lack of infrastructure and premises and poor working conditions for employees.	[26, 33, 34]
Capacity	Low literacy levels, lack of business skills, lack of managerial skills and lack of training opportunities.	[6, 21]
Regulations	Lack of compliance with regulations and laws and formalisation requirements.	[14, 19]
Support	Lack of awareness of support services, lack of access to business advice and acceptance in communities.	[21, 22]
Health	Effect of health conditions such as HIV/AIDS pandemic, care costs, benefit claims and burial fees.	[6, 12]

2.2. ICT use in microenterprises

This study views ICTs as a tool; conceptualised as labour substitution, production, productivity, information processing and social relations tool [27]. Due to limited

financial resources accessible to microenterprises, their use of ICT is often limited to less sophisticated technologies such as personal computers, laptops, radios, telephones, mobile phones, fax machines, photocopiers and television [10, 14]. Furthermore, in cases where microenterprises cannot afford to acquire their own ICTs, public or communal ICT facilities (e.g. telecenters, internet cafés and public telephones) are used [11, 23]. However, due to lack of separation between business and personal assets, ICTs in microenterprises are often used for both personal and business purposes [10, 14].

The benefits of using ICTs in microenterprises include increased productivity, better access to information and knowledge, reduced administrative costs and improved communications [10, 35]. Nevertheless, the use of ICT in microenterprises is often problematic. Microenterprises face challenges that hinder them from reaping the benefits of using ICTs [34]. Some of the challenges concern their capabilities, resources, access, operations, context and attitudes [34]. Capabilities challenges are caused by lack of ICT skills, planning and knowledge to apply ICTs to support the business activities of the microenterprises [28]. Microenterprises lack resources in the form of time, finance and information, which leads to failure in using ICTs [23]. Attitude (e.g. resistance to technology) also affects use of ICTs in microenterprises and may affect the confidence of microentrepreneurs to use ICTs. Microenterprises that have acquired ICTs may also face challenges in terms of support to implement and maintain ICTs [34].

3. THEORETICAL BACKGROUND TO THE STUDY

3.1. Sustainable Livelihood Approach (SLA)

SLA was developed as a result of the policy debate on the Brundtland Commission Report of 1987 [4]. Since then, the concept of livelihoods has evolved and SLA has been applied to numerous studies within the development discourse. A livelihood is defined as follows:

A livelihood comprises of assets (natural, physical, human, financial and social), activities, access to these (mediated by institutions and social relations) that together determine the living gained by individuals or households [13].

In other words, a livelihood is a means for living where capabilities and assets are utilised to enhance opportunities [4, 13]. A livelihood is considered sustainable when it can cope with hardships and can support a household to continue operating over a period of time [9].

SLA comprises of elements that are based on thinking about poverty reduction and can be used to holistically analyse activities within the livelihood. The elements are vulnerabilities, assets, structures and processes, strategies and outcomes. *Vulnerabilities* are external conditions that

may lead to hardships and undermine the potential of households. They encompass natural disasters, conflicts, limited employment opportunities, social exclusion and changes in prices of commodities [9]. *Assets* are the resources that households use to attain sustainable livelihoods; they are in the form of human, natural, financial, physical and social resources (see Table 3). Increased access to assets may lead to sustainable livelihoods for households or communities [13].

Table 3. Forms of Livelihood assets [9]

Type of Asset	Description
Human capital	Skills and knowledge that people have and use to achieve their livelihoods.
Social capital	Social relations that people have and influence their actions e.g. membership to organisations.
Financial capital	Items of value which people use to establish livelihood activities i.e. saving, cash and access to loans.
Natural capital	Natural resources used by households to achieve livelihood goals i.e. land, water, wildlife and biodiversity.
Physical capital	Resources created through economic production process i.e. infrastructure such as roads, power lines and telecommunications.

Structures are organisations and institutions that provide support for livelihoods to households and communities. Organisations could be international development agencies, government departments, Non-Governmental Organisations (NGOs) and Community Based Organisations (CBOs) and private sector institutions [9]. Institutions focus on supporting households and communities in areas of legislations and regulations affecting their livelihoods [3]. Organisations focus on supporting households and communities to think of strategies that reduce livelihood vulnerabilities. *Strategies* are activities carried out by both structures and households to achieve livelihoods *outcomes* [13]. Outcomes could be improved well-being, increased income, restored human dignity and reduced vulnerability [11].

ICT and information have the potential to play a significant role in attaining sustainable livelihoods. ICT supports communication of information and knowledge on livelihoods strategies [7]. Information within the SLA model plays both analytical and functional roles [11]. The roles are described as:

- *Analytical role:* how data is accessed, assessed and applied to understand livelihoods.
- *Functional role:* how information is used to initiate action for livelihoods.

Information and ICTs can be applied in vulnerability contextual analysis, and in the use of livelihood strengths and capabilities, which may be mediated through social relations, institutions and organisations [11]. Information for long-term decisions for livelihoods of microenterprises is usually communicated through formal means, by institutions and organisations. Information for short-term decisions, mediated through informal means such as social networks used for decisions about immediate needs for households or communities [12].

3.2. Justification for using SLA to assess the impact of ICT

Based on the context of the study, SLA was considered ideal because it encompasses a wider scope of poverty related challenges [5, 11]. Microenterprises are an option for alleviating poverty, leading to social and economic empowerment, especially in marginalised communities [28]. The core principles of SLA promote a wider perspective of analysis of livelihoods. Furthermore, the principles show a people-centred approach, holistic analysis, macro-micro links, sustainability, they are strength-based and show extensive stakeholder involvement [3, 9].

4. INTERVENTION UNDER STUDY

The study was conducted on an intervention for SMMEs development of the Western Cape Provincial Government called Real Enterprise Development (RED) Door. RED Door provided business support and advice to SMMEs in communities in the province. The services included [30]:

- Facilitating business start-ups for SMMEs;
- Providing access to non-financial resources and capacity;
- Providing training on business skills, as well as legal advice to SMMEs;
- Facilitating access to finance by service providers; and
- Providing information on market access, procurement and market linkages for SMMEs.

RED Door also provided free internet services to its clients [30]. The target groups for RED Door services are entrepreneurs who want to start or expand a small business.

5. METHODOLOGY

We employed a qualitative interpretive approach, with the aim of obtaining interpretations and explanations, from the respondents on the phenomena being investigated [24]. We employed purposeful and snowballing sampling techniques. Purposeful sampling was used to select respondents who

would provide the required data and who conformed to the criterion set by the researchers. A snowballing technique was applied to select new respondents, following recommendations by the participants.

Data was gathered through in-depth interviews, observations and documentary review. A total of 11 interviews were conducted, with respondents of whom four were a control sample (these were microenterprises that did not participate in programs or use the services of RED Door) and five were organisations that received support and used services provided by RED Door (we call these “*beneficiary organisations*”). Members of management from RED Door were also interviewed; one manager from head office and another from the business advice centre were interviewed. A follow-up interview was conducted with one manager.

Data analysis involved an iterative process, which began with analysing secondary data used, to understand the context of microenterprises. The transcribed primary data collected during field visits was prepared to ensure that data was complete, using field notes and observation notes. We employed thematic analysis [2] to analyse the interview transcripts. The data analysis process was deductive, guided by the SLA conceptual framework [32].

6. FINDINGS

6.1. Motivation, skills and financing of microenterprises

It was noted that some of the reasons microenterprises engaged in business activities were the need to develop skills and to generate income for their households. The owners of microenterprises perceived that, by operating their own businesses, they would generate more income than they would earn as wages from low-income jobs. This is exemplified in the following statement by one of the respondents:

... the only choice to make money is doing business ... I can make more money than working

Further, it was noted that microenterprises developed business skills from other microenterprises, who acted as role models for success. This was evident in the following statement by a respondent:

... but also the way you learn from others ... I learn from my brother

In some cases, microenterprises also benefited from business support organisations, where training on business skills was offered. Once they had acquired the skills and knowledge, microenterprises were confident to start their own businesses.

Consistent with the literature [28, 31, 33], we noted that capital was one of the major limiting factors for business start-up in microenterprises. Due to lack of resources for collateral from financial institutions, the majority of the microenterprises used their savings for capital. Microenterprises overcame the challenge of business capital, by developing business models that required small capital and by operating small scale business activities. Income generated from the small scale activities was re-invested in the business to expand business activities and achieve business growth.

6.2. Extent of ICT use in microenterprises

With regard to the extent of ICT use, the majority of microenterprises used simple ICTs. Cell phones were used in almost all the cases. Other ICTs that were used were PCs, telephones, the internet and electronic mail, televisions and radios. Public ICT facilities were also used. Usage of ICTs was on both business and personal purposes. The main uses of ICTs were for communication with customers, suppliers and close relations, as well as for information gathering. Table 4 summarises the ICTs used in microenterprises.

Table 4. Summary of ICTs used in microenterprises

ICT	Actual use
Computers	Preparation of documents e.g. business plans, leaflets and recording business transactions
Cell phones	Communicating with customers, suppliers and members of staff. Also used to communicate with family members.
Internet and email	Searching for information on tenders and communicating with customers.
Television	Entertaining customers, source of information and testing DVDs before they are sold.
Radio	Entertainment while conducting business and source of information for local news.

It was also observed that, in some cases, microenterprises did not use ICT for business, despite owning or having access to ICTs. Some of the factors that led to non-use of ICTs for business activities were lack of IT skills and lack of knowledge on how to apply ICTs in the business activities.

6.3. Outcomes of use of ICTs, assets and structures

The use of ICTs in microenterprises varied across the different organisations, depending on the nature of business activities. It emerged that ICTs were used in activities that indirectly facilitated better usage of assets and, in some cases, access to institutions and organisations. For example, microenterprises used telephones and cell phones to check availability of materials, prices of materials and services and interacting with business support organisations. ICTs

were mainly used in the acquisition and maintenance of social capital as an asset.

There is limited evidence to indicate that ICTs were directly used to acquire other forms of assets (i.e. human capital, physical capital, financial capital and natural capital). However, ICTs were significant in accessing information that was used for decision making regarding the use of assets, such as identifying information needs that influenced decisions on use of different forms of assets (human, financial, physical and natural). The information was also used in coordinating the activities for microenterprises. Further, ICTs indirectly affected the way resources were acquired and used. For instance, use of physical assets such as information on where to buy and local prices, were significant and such information, to some extent, was obtained using ICTs. Table 5 summarises how microenterprises used the different forms of livelihood assets.

Table 5. Summary on use of livelihood assets

Asset	Description on use of asset by microenterprises
Human capital	Microenterprises used prior business knowledge and skills; and methods of acquiring skills were also through trial and error and peer-learning
Financial capital	Resources for business start-up were through self-financing using savings and business expansion was through reinvesting of returns due to limited access to financial services.
Social capital	Microenterprises used social capital such as formal social groups and informal social groups. The support obtained from groups included business social support and support from family members and staff.
Physical capital	Microenterprises required information for operating their businesses. The information needs were market information, local prices, loan information, area information, business advice and supplier's information.
Natural capital	Majority of the respondents did not use natural resources and they were mainly involved in retail trading activities and faced challenges in sourcing materials and information on prices for materials.

The outcomes of use of ICTs, assets and interaction between microenterprises with institutions and organisations were mainly improved well-being and more income.

6.4. Vulnerabilities affecting microenterprises

The owners of microenterprises were facing limited opportunities for employment and were engaged in micro business activities to overcome this challenge. Another

challenge was fluctuations in prices of input materials. This affected the operations of microenterprises.

It was also noted that microenterprises experienced conditions that were difficult to deal with and were mainly internal to their livelihoods. Literature on SLA suggests two types of vulnerabilities:

- *Shocks*: short-term conditions leading to hardships in a livelihood i.e. epidemics, natural disasters and conflicts; and
- *Trends*: long-term conditions leading to hardships in a livelihood i.e. population decline, poor governance and technology changes.

Both these vulnerabilities are external to a livelihood [9]. However, the challenges identified in the analysis were mainly internal to the livelihood of the microenterprises, such as business operations challenges, lack of managerial skills and business performance challenges.

6.5. Use of ICT to interact with institutions and organisations

Apart from microenterprises that benefited from the RED Door services, the other microenterprises did not use ICT to interact with organisations providing support for SMMEs, mainly due to lack of awareness of such organisations. Microenterprises in this group, however, interacted with institutions regulating operations for microenterprises, for example, owners of the microenterprises interacting with local council officials when making payments for monthly operating fees. Communication with the officials was face to face.

7. DISCUSSION

Reflecting on the objective of the study on the effects of ICT use on the livelihood of microenterprises, the results in this study have confirmed that poverty is a multi-faceted phenomenon. The study has also confirmed that SLA can be used as a theoretical lens to understand the role of ICTs in a livelihood of microenterprises. This study confirms that microenterprises require information to address information gaps related to use of livelihood assets. Information needs influence the way the different forms of assets are used in microenterprises.

The microenterprises in this study benefited from using simple technologies, mainly cell phones and telephones, to obtain and share information. The information and knowledge were used to come up with strategies that led to the effective use of assets, leading to sustainable livelihoods, as suggested by Chapman and Slaymaker [7]. Although ICTs may be perceived to be an enabler that could increase the survival chances of microenterprises [16, 18], the microenterprises faced challenges that inhibited the

adoption and use of ICT for business purposes. The challenges included limited resources, capabilities and knowledge. These challenges led to non-use of ICTs, despite having access to and owning ICTs. This was consistent with a previous study [14] that demonstrated that provision of physical access to ICTs does not guarantee that the ICTs will be used. One way of addressing this problem is for business support institutions to provide training and support on how ICTs can support business needs for microenterprises as part of capacity building.

One of the challenges microenterprises faced was lack of awareness of support services. Duncombe and Heeks [12] expressed similar sentiments on lack of awareness of support services in microenterprises. Part of the reason could be a misalignment in the way that organisations and institutions communicate and the way that microenterprises communicate. Literature suggests that microenterprises rely more on social networks that are local, although the information from these sources may be of poor quality and unreliable [12, 23]. This could be done by using existing community structure, especially local and social networks, where microenterprises participate in promoting support services for SMMEs.

8. CONCLUSION

The study noted that microenterprises face a myriad of challenges, such as changes in prices for raw materials, operational issues, business performance issues and business competitiveness. The use of ICTs for the microenterprises in the study was relatively low in utilisation of livelihood assets and the ICTs were mainly used for communication and information gathering. In some cases, despite owning and accessing ICTs, microenterprises did not use ICTs for business purposes. Furthermore, the majority of microenterprises, especially those in the informal sector, did not use ICTs to interact with organisations and institutions providing support for SMMEs. Part of the reasons inhibiting the use of ICTs was lack of knowledge on application of ICTs in their business and lack of awareness of business support services. Consequently, some cases of microenterprises perceived organisations and institutions as not helpful and inaccessible. Overall, ICT use had a positive impact on livelihoods of microenterprises, mainly in a higher income and increased well-being.

REFERENCES

- [1] A. Berry, M. von Blottnitz, R. Cassim, A. Kesper, B. Rajaratnam, and D van Seventer, "The Economics of SMME in South Africa," Trade and Industry Strategies and Policies, 2002.
- [2] V. Braun and V. Clarke, "Using thematic analysis in Psychology," *Qualitative Research in Psychology*, vol. 3, pp. 77-101, 2006.
- [3] D. Carney, "Approaches for Sustainable Livelihood for the rural poor," ODI Poverty Briefing, 1999.
- [4] R. Chambers and G. Conway, "Sustainable rural livelihood: practical concepts for the 21st century," Institute of Development Studies, University of Essex, UK, Discussion Paper 296, 1999.
- [5] V. Chandra, L. Moorty, J. Nganou, B. Rajaratnam and K. Schaefer. "Growth and Employment in SA: Evidence from the Small, Medium and Micro Enterprise Firm Survey," Trade and Industrial Policy Strategies (TIPS), South Africa, Informal Discussion Paper on aspects of the South African Economy, 2001.
- [6] L. Chao, M. Pauly, H. Szrek, N. Pereira, F. Bundred, C. Cross, and J. Gow, "Poor health kills small business: illness and microenterprises in South Africa," *Health Affairs*, vol. 26 no. 2, pp. 474-482, 2007.
- [7] R. Chapman and T. Slaymaker, "ICTs and Rural Development: Review of the Literature, Current Interventions and Opportunities for Action," Overseas Development Initiative, Working Paper 192, 2002.
- [8] H. Chew, P. Ilavarasan and M. Levy "The economic impact of Information and Communication Technologies (ICT) on Microenterprises in development context". *Electronic Journal on Information Systems in Developing Countries*, vol. 44 no. 4, pp. 1-19, 2010.
- [9] DFID, "Sustainable Livelihood Approaches Guidance Sheet," Department for International Development: London, 1999.
- [10] J. Donner, "The use of Mobile Phones by Microentrepreneurs in Kigali, Rwanda: Changes to Social and Business Networks," *Information Technologies and International Development*, vol. 3 no. 2, pp. 3-19, 2006.
- [11] R. Duncombe. "Using the Livelihood Framework to analyze ICT applications for Poverty Reduction through Microenterprise," *Information Technologies and International Development*, vol. 3 no. 3, pp. 81-100, 2006.
- [12] R. Duncombe and R. Heeks, "Information and Communication Technologies (ICTs), poverty reduction and Micro, Small and Medium-scale Enterprises (MSMEs)" IDPM, University of Manchester, UK, 2005.
- [13] F. Ellis, "Rural Livelihood and Diversity in developing countries", Oxford Press, 2000.
- [14] S. Esselaar, C. Stork, A. Ndawalana, and M. Deen-Swarray, "ICT usage and its impact on profitability of SME in 13 African Countries", *Information Technologies and International Development*, vol. 4 no. 1, pp. 87-100, 2007.
- [15] F. Fraser, W. Grant, P. Mwanza and V. Naidoo, "Impact of HIV/AIDS on Small and Medium Enterprises in South Africa," *South African Journal of Economics*, vol. 70, no. 7, pp. 1217-1234, 2005.
- [16] T. Good, and S. Qureshi, "Investigating the effects of microenterprise access and use of ICT through a Capability lens: Implications for Global Development", Proceeding of the Second Annual SIG GlobeDev Workshop, Phoenix, USA, 2009.
- [17] Information Economy Report, "ICTs, enterprises and poverty alleviation," United Nations Conference on Trade and Development, 2010.
- [18] M. Kamal and S. Qureshi, "An Approach to IT Adoption in Micro-enterprises: Insights into Development," Proceeding of the Fourth Midwest United States Association for Information Systems Conference, Madison, USA, 2009.
- [19] M. Kyobe, "Factors influencing SME compliance with government on use of IT," *Journal of Global Information Management*, vol. 17 no. 2, pp. 30-59, 2009.
- [20] A. Ligthelm, "Size estimate of the informal sector in South Africa," *Southern African Business Review*, Vol. 10 no. 2, pp. 33-40, 2006.
- [21] J. Luiz "Small business development, entrepreneurship and expanding the business sector in a developing economy: The case of South Africa," *The Journal of Applied Business Research*, vol. 18 no. 2, pp. 53-69, 2002.
- [22] Z. Mitrovic and A. Bytheway, "Awareness of e-Government related Small Business development services in Cape Town" *Electronic Journal of Information Systems in Developing Countries*, vol. 39 no. 4, pp. 1-14, 2009.
- [23] E. Moyi, "Networks, information and small enterprises: New technologies and the ambiguity of empowerment," *Information Technology for Development*, vol. 10, pp. 221-232, 2003.
- [24] M. Myers, "Qualitative research in Business and Management," Sage Publications: London, 2009.

- [25] National Small business Amendment Bill, “Republic of South Africa: Minister of Trade and Industry, Government Gazette,” ISBN 062133889, 2003.
- [26] R. O’Neil and L. Viljoen, “Support for female entrepreneurs in South Africa: improvement or decline?,” *Journal of Family Ecology and Consumer Sciences*, vol. 19, pp. 37-44, 2001.
- [27] W. Orlikowski and C. Iacono, “Research Commentary: Desperately seeking the IT in IT Research-A call to theorizing the IT Artifact,” *Information Systems Research*, vol. 12 no. 2, pp. 121-134, 2001.
- [28] S. Parkinson and R. Ramirez, “Using a Sustainable Livelihoods Approach to assessing the impact of ICTs in Development,” *Journal of Community Informatics*, vol. 2 no. 3, pp. 1-14, 2006.
- [29] S. Qureshi, “How does information technology effect development? Integrating theory and practice into a process model,” *Proceeding of Americas Conference on Information Systems*, Omaha, NE, USA, 2005.
- [30] RED Door Business Plan, “The RED Door Business Plan,” Western Cape Town, Department of Economic Development and Tourism, 2004.
- [31] C. Rogerson, “Tracking SMME development in South Africa: Issues of finance, training and the regulatory environment,” *Urban Forum*, vol. 19, pp. 61-81, 2008.
- [32] D. Thomas, “A general inductive approach for analyzing qualitative evaluation data,” *American Journal of Evaluation*, vol. 27 no. 2, pp. 237-246, 2003.
- [33] J. Visagie, “SMME challenges in reconstruction South Africa,” *Management Decision*, vol. 35 no. 9, pp. 660-667, 1997.
- [34] P. Wolcott, M. Kamal, and S. Qureshi, “Meeting the challenges of ICT adoption by micro-enterprises,” *Journal of Enterprise Information Management*, vol. 21 no. 6, pp. 616-632, 2008.
- [35] S. Wolf, “Determinants and Impact of ICT use for African SMEs: Implications for Rural South African,” *Annual Forum, Trade and Industrial Policy Strategies*, 2001.

SM² : SOLAR MONITORING SYSTEM IN MALAWI

Mayamiko Nkoloma^A, Marco Zennaro^B, Antoine Bagula^C

A. Department of Electrical Engineering

Malawi Polytechnic

Blantyre, Malawi

B. ARPL

The Abdus Salam International Centre for Theoretical Physics

Trieste, Italy

C. ISAT Laboratory

University of Cape Town

Cape Town, South Africa

Emails: mnkoloma@poly.ac.mw, mzennaro@ictp.it, bagula@cs.uct.ac.za

ABSTRACT

This paper describes recent work on the development of a wireless based remote monitoring system for renewable energy plants in Malawi. The main goal was to develop a cost effective data acquisition system that continuously presents remote energy yields and performance measures. A test bed comprising of a solar photovoltaic (PV) power plant has been set up at Malawi Primary School and a central management system at Malawi Polytechnic. The project output gives direct access to generated electric power at the rural site through the use of wireless sensor boards and text message (SMS) transmission over cellular network. The SMS recipient at the central site houses an intelligent management system based on FrontlineSMS for hosting SMSs and publishing remote measurement trends over the Internet. Preliminary experimental results reveal that the performance of renewable energy systems in remote rural sites can be evaluated efficiently at low cost.

Keywords – SMS, Wireless Sensors, Solar Power, Remote Monitoring

1. INTRODUCTION

The potential provided by solar power combined with the dangers raised by greenhouse gas emissions leading to climate change have paved the way for the adoption of the solar technology and subsequent investments in solar installations as a less polluting power source alternative. It has been recently reported [1] that if humanity could capture one tenth of one percent of the solar energy striking the earth, the World would have access to six times as much energy as we consume in all forms today. The African continent receives an average of 6kWh of solar energy per square meter every day [2]. Yet, as currently exploited, solar power is still an untapped resource representing only a minuscule fraction of the planets power generation capacity.

This work is supported by grant numbers 2047362 and 2677 from the South African National Research Foundations, Siemens Telecommunications and Telkom SA Limited.

Solar technology infrastructures provide the potential for social and economic advances in the rural areas of the developing world which generally suffer from a lack of appropriate and reliable electrical grids. However, the monitoring of their installation is an important parameter upon which wide deployment of such infrastructures depend. Some of the advantages of solar installation monitoring include (1) letting businesses and home owners to get a real-time readout of their solar panels with the associated economic benefit of making the best trade-off between switching between electrical and solar supply (2) rapid problem identification and preemptive resilience to failure allowing qualified service technicians to quickly fix the problem (3) self-repairing of the solar system through automated software when possible.

1.1. Related work

A variety of commercial monitoring systems are available for plants ranging from small-scale residential rooftop to large commercial renewable energy system (RES). These plants typically consist of a local electronic device (data logger) that connects to the energy system and records data over time and thereafter relays it on to the monitoring service provider's central data centre. These commercial tools comprehensively display transient real-time and historical graphical trends of RES plants, usually in Web based format over the Internet. In addition, some advanced applications offer system alarms and notifications via email or SMS during operation failure or when specific conditions are met. Companies offering remote monitoring products and services include SMA Solar Technologies [3], inAccess Networks [4], Fat Spaniel Technologies [5], Morningstar Corporation [6], SolarMax [7] and others.

Developed by the New York University, the SIMbalink project aim is to provide sustainable electrification solutions for rural areas. SIMbaLink is based on an extremely low cost real time solar monitoring system that reduces the

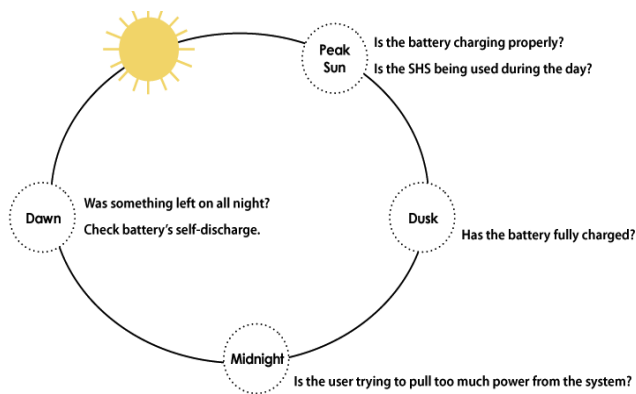


Figure 1. The SimbaLink System

maintenance costs and the time to repair. The system reveals important information about battery's state of charge and daily energy use. The data is transmitted over GSM cellular network to a regional technician to allow remote system diagnostics. However in their innovative solution, readings are only taken four times per day, as displayed in Figure 1. Consequently, this does not presents real time trends to enable critical performance analysis and timely detection of solar plant problem.

1.2 Contributions and Outline

A research project aiming at developing advancement in manufacturing, installation and maintenance support of new small scale solar and hydro electrical energy generating equipment in Malawi is being conducted by the faculty of Engineering at the Malawi Polytechnic. The project includes the installation of solar plants in rural primary schools and health centres. However, long distance and a poor road network between sites make it more challenging for the team to perform tests and monitor the performance of the plants. As a first step towards efficient management of the energy generating equipment, this paper reports on a solar system monitoring application developed as part of the whole project. The main features of the application include (1) solar power consumption monitoring using sensors measuring panel voltage and current capture (2) information dissemination using FrontlineSMS [9] and (3) data publishing using Web services based on PHP and associated graphing tools. As a new innovative solution that demonstrates a low cost mechanism for RES using the existing mobile network infrastructure, the proposed application present the following key benefits:

- Access to PV system performance from anywhere through the use of Internet.
- Reports of power output and energy production trends.
- Verification of system operation.
- Collection of data for service and maintenance planning.
- Use of open devices which lower the cost and enable the replicability of the solution.

The choice of using the GSM network was dictated by the lack of other solutions in the Malawi area. From the power consumption point of view, using a

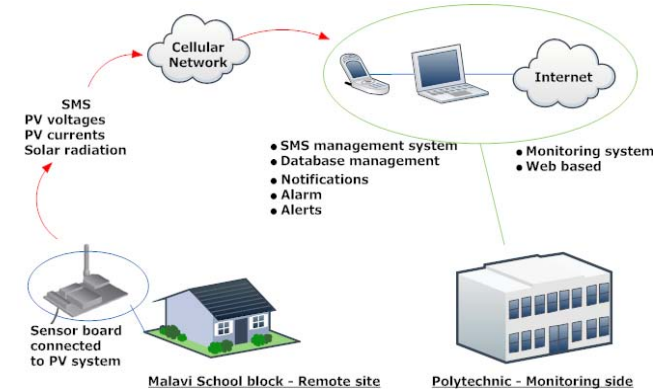


Figure 2. System Architecture: Solar PV with Wireless Sensor and Central Management Servers

GSM module requires more energy than using low-power protocols such as Zigbee. From the connectivity point of view, GSM networks cover most of the country and connectivity costs are limited given the competition of two operators (AirTel Malawi Limited and Telekom Networks Malawi). The SIMbaLink project follows an approach which is similar to ours as it is based on Arduino devices and builds upon open source solutions. However, our solution makes use of FrontlineSMS to enable users to access the status of the system from remote.

The remainder of this paper is organized as follows. Section 2 describes the System Architecture while Section 3 presents some results from Malawi Primary school testbed. Section 4 contains our conclusions and examines the way forward.

2. SYSTEM ARCHITECTURE

A general view of the system is shown in Figure 2. It is composed of three elements: the remote site where the solar PV system is installed, the wireless sensor data capturing boards and the server side at the Malawi Polytechnic where system management is hosted.

2.1 Photovoltaic System

The solar system at the Malawi Primary school is made up of three key components; solar modules, charge controller and battery bank. The solar array is composed of two LORENTZ LA75-12S 75W PV modules [10] each module having 32 monocrystalline silicon cells. The manufacturer claims that the cells yield higher voltage making the module provide sufficient voltage as that realized with traditional 36 cell modules. In addition to that, each PV module is capable of yielding short circuit current (ISC) of 5.4A and open circuit voltage (VOC) of 21V. The StecaTarom 12A charge controller is the brain of the system which controls the amount of power going into and coming out of two 102Ah Deltec batteries. As

such, it prevents the solar module from overcharging and the load from over discharging the batteries, thereby maximizing the battery life. Consequently, all components of the solar system are connected through the charge controller. A 600W TES inverter is installed to power a single AC socket and lastly the DC load consists of eighteen 11W bulbs.

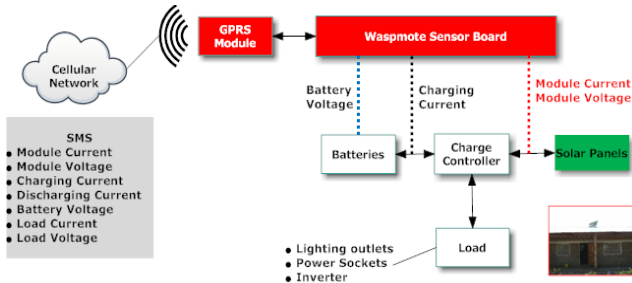


Figure 3. Detailed system architecture for remote data capturing and transmission

2.2 Wireless Sensor

Solar system monitoring is achieved through the use of the Wasp mote by Libelium [11]. The main advantages of the Wasp mote are 1) its modular architecture allows developers to choose a wide range of available modules depending on the application, 2) low-cost resulting from an open source software and hardware platform design, 3) an easy programming environment where the development phase and the adaptation to different needs takes little time and 4) the provision for a wide range of wireless applications. Using different modules, Wasp mote can connect to low-power wireless networks based on 802.15.4 and Zigbee. In our case, we used the GSM module to send data to the Malawi Polytechnic. Equipped with a SIM card, this module allows to send and receive SMS and even to connect to the Internet via GPRS when available. The Wasp mote has 7 accessible analog inputs, which can be utilized in capturing solar system performance parameters via analog sensors. Each input is directly connected to the microcontroller which uses a 10 bit successive approximation analog to digital converter (ADC). The reference voltage value for the inputs is 0V (GND) and the maximum value is 3.3V which corresponds to the microcontroller’s general power voltage. Consequently, the board represents integer values ranging between 0 and 1023 which corresponds to actual input range of 0V and 3.3V. Figure 3 presents detailed system architecture for remote data capturing and transmission.

To accomplish voltage and current measurements, the Wasp mote is equipped with two external circuits: Phidget 1117 voltage sensor [12] and Phidget i-snail-VC 100 current sensor [13]. The voltage sensor allows measurements up to 30 Volts, while readings of up to 100 Amperes can be achieved with the current sensor. Figure 4 shows the Wasp mote with the GSM module during voltage sensor calibrations. Voltage reading is attained by

tapping solar PV module voltage directly to a voltage sensor, which is connected in parallel with the charge controller, as depicted in Figure 5.



Figure 4. Sensor Calibration Setup

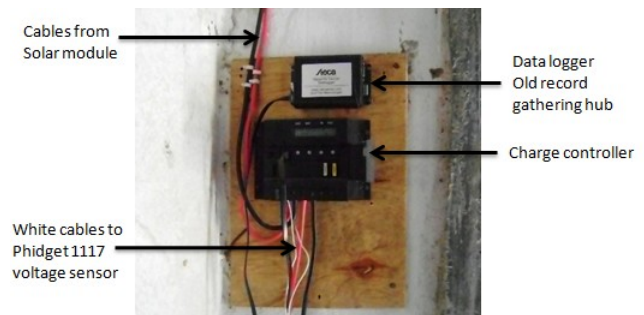


Figure 5. Wire connections to voltage sensor

The Wasp mote board is programmed to calculate input voltage given a voltage reading at its analog input using the following method:

$$\text{Solar system voltage (in volts)} = \frac{\text{Read voltage} \times 200 \times 0.06}{-30}$$

2.3 Central SMS Management System

Cost effectiveness of this project is entrusted on the rate at which the measurements are conducted, as this determines the total number of SMSs to be transmitted in a particular period of investigation. Experimentally, a 30 minute-measurement interval is being opted for and 8 readings are logged locally in Wasp mote’s SD card before transmitting. Ultimately, this results with a requirement of 6 SMS transmissions in a day. Currently, the cost to send an SMS within the same network is about 0.07USD (10 Malawi Kwacha), and a bulk sum of 180 SMSs can be sent for 12.60USD, which is the total running cost of monitoring the system in a month. In addition to comprehensively securing and transmitting solar system parameters over cellular network, from Malawi Primary School, the project mission attainment also depends on the central communication hub proficiency in data management, concise relying and presentation of performance trends on a Web portal. Consequently, it is essential to identify other system components that need to be integrated and propel towards actualization of the goals. Furthermore, due to limited project time frame, it was necessary to amalgamate with core tools that have been experimented and proved to be

feasible for these kinds of projects and also that are highly adopted in robust Internet based infrastructure. Resultantly, en route to the ambition, diagrams comprising of main system building blocks are presented in Figures 6 and 7 that were opted as a blueprint in the designing of a central SMS management system at the Malawi Polytechnic.

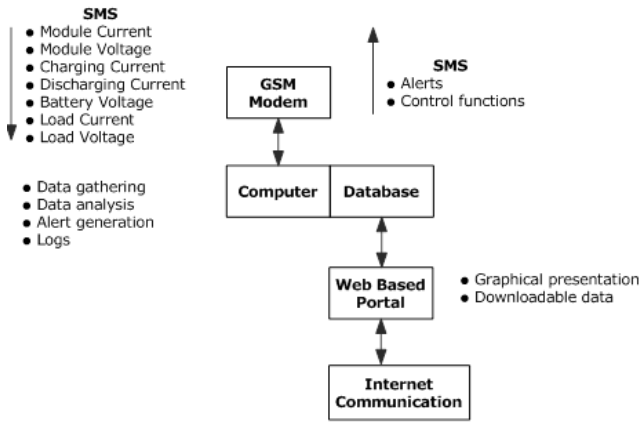


Figure 6. Remote site sensing mechanism

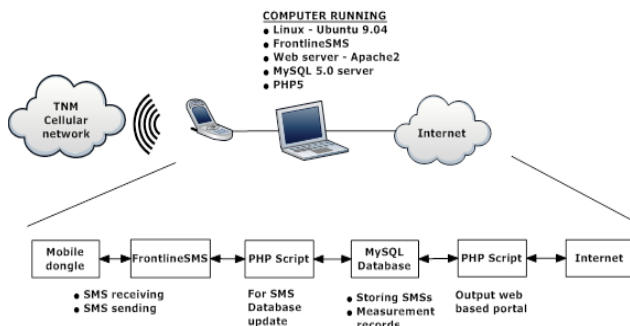


Figure 7. Remote site sensing mechanism

Through consecutive series of research studies, the following are typical elements used in constructing the Linux based central communication hub.

- Mobile phone in this case a Huawei mobile broadband dongle E160.
- Computer with the following packages:
 - LAMP bundle, which is a combination of Linux, Apache, MySQL and PHP
 - FrontlineSMS set up as an SMS gateway.

FrontlineSMS is an award winning open source software that turns a laptop and a mobile phone into a central communication hub. It is mainly used by the non-profit sector and non governmental organisations (NGOs) to reach specific groups of individuals within a target community. Its functionalities have found usage in sending information on health, security alerts, job information, and market prices from monitors, surveys and other data collection sources. Specifically in 2007,

the software was internationally adopted to assist local NGO groups carry out citizen monitoring of the Nigerian Presidential elections. Furthermore in 2009, in conjunction with the Ushahidi mapping platform, the application was used to track essential medicine stock outs in several East African countries. On the other hand, in Kenya and Tanzania, FrontlineSMS is being implemented as part of social business to keep in-touch with farmers who have bought, or expressed interest in buying Kickstart pumps.

In this project, FrontlineSMS abilities are harnessed in obtaining a solution for capturing SMSs from wireless sensor nodes that capture performance parameters of RES. Furthermore, in this scenario, FrontlineSMS works hand in hand with other scripts so as to house all received SMSs in a database for further analysis. A backend PHP script for populating a MySQL database is linked to the SMS gateway and once an SMS is received, the script gets triggered and acquires two strings which are ushered by FrontlineSMS's external command triggering functionality. In this case, the strings accommodate sender number and message content. However, apart from mere character passing to the database, the script also checks for particular keywords which signify a text message with solar readings from remote site. In addition, as the information encloses logged performance parameters that are different and also captured at different times, the script unscrambles the content and then updates MySQL table fields in a logical manner. On the other hand, the frontend PHP script connects to the database and retrieves parameter readings for web based presentation. Plotting is achieved with the use of JpGraph, an object-oriented graph creating library for PHP5. The library is completely written in PHP and can be incorporated in any PHP script. It supports several plot types; spider plots, pie charts (both 2D and 3D), scatter plots, line plots and bar plots just to mention a few.

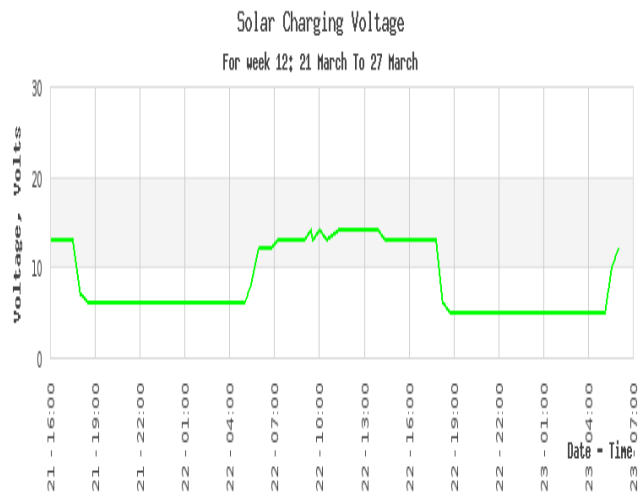


Figure 8. Solar Charging voltage

3. PRELIMINARY EXPERIMENTAL RESULTS

We conducted a number of experiments with the objective of monitoring the performance of the PV installation in terms of solar charging voltage, load voltage and Wasmote internal temperature. Figures 8, 9 and 10 show a graphical representation of these three parameters during a particular test period. Solar charging voltage had a constant trend as depicted in a snapshot of a two day period shown in Figure 8. Referring to the same result, it is clear that voltage follows a day/night pattern, as solar panels provide voltage during the day and provide little voltage during the night. In other words, during this season of the year voltage grows starting at 5 am and reaches its maximum at 7 am. In the evening, it starts to drop at 6 pm and by 7 pm the panels provide small voltages.

The Wasmote internal temperature is shown in Figure 9. This element has been monitored to check that the board is not exposed to excessive heat at the testbed site.

Apart from this, Figure 10 shows a trend analysis of the load voltage that is essential in supplying power to the eighteen 11W bulbs and a single AC socket of the school block.

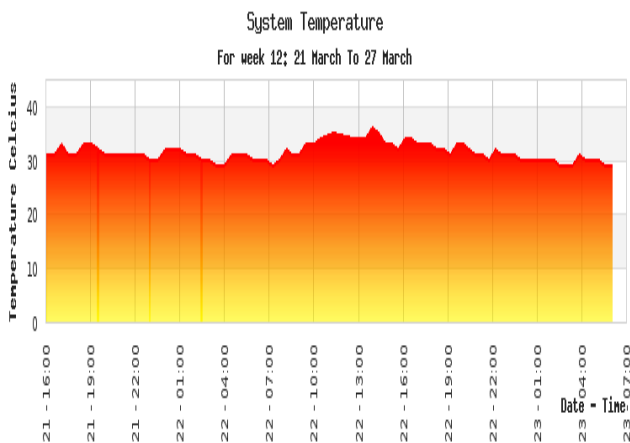


Figure 9. System Temperature

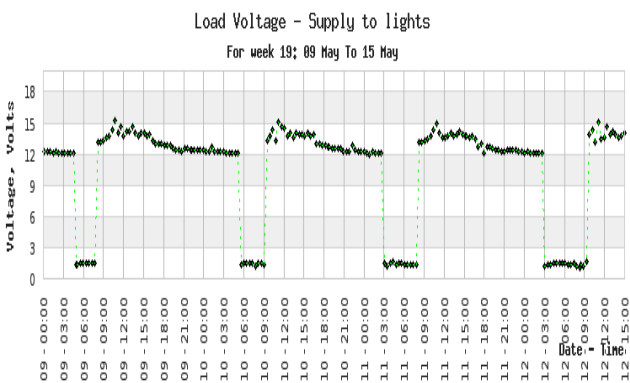


Figure 10. Load voltage (a)

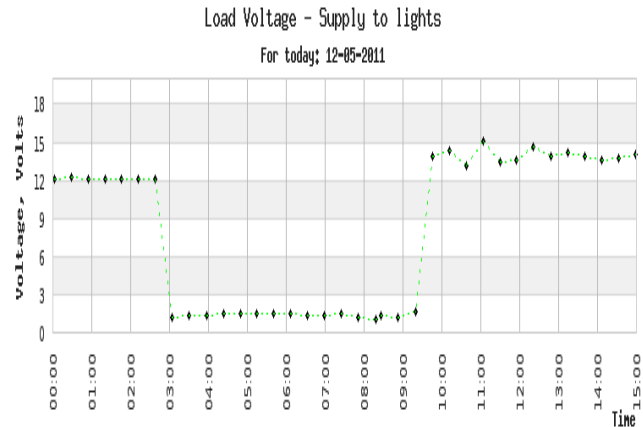


Figure 11. Load voltage (b)

For simplicity, referring to Figure 11, during the night the system obtains a constant voltage of about 12 V, which is supplied from the batteries. However due to activities that are running at the school during this period of investigation, the battery voltage is not sufficient enough to supply power throughout the period of little solar energy. Consequently, at around 3 am, the battery voltage drops and results in the switching off of the load power and commencement of battery charging. This effect is evidently observed as there is a complete power outage until around 10 am during which the school block obtains power directly from the solar panels after full charging of the batteries. This output indicates that the system was under designed and there is a need of adding one or more batteries to meet the required energy demands of the rural community school. On the project website it is possible to visualize these results on weekly, monthly and yearly graphs.

4. CONCLUSIONS AND FUTURE WORK

A sensor-based monitoring application built around a solar photovoltaic power plant at Malawi primary School has been presented in this paper as part of a bigger project led by Malawi Polytechnic. The application uses a Wasmote sensor board from Libelium where two phidget sensors have been grafted to allow voltage and current readings. The proposed application combines built-in Wasmote SMS capabilities and the widely known FrontlineSMS software to achieve information dissemination from the PV system site to the monitoring site and also from this site to the PV system's supervisors. This is augmented by web publishing using PHP and associated graphing tools. The proposed monitoring system is currently running and has proved to lower management cost as timely information reaches the group at the Polytechnic right in front of their work stations. This can assist in alerting technical team of remote circumstances and also ease system study time for the researchers. The designed system monitoring website enables users to select specific monitoring times to suit the analysis at hand. The preliminary results presented in this paper logically agree with what is expected as the trend for solar module voltage during the day and night.

There is room for future work to extend the capabilities of

the proposed system in different directions. One way is to expand it to effectively measure more performance parameters such as current consumption of different users of the PV system. The proposed system could also be extended to allow a smooth switchover between electrical and solar power supply depending on time-of-the-day power needs. Using the attribute of the GSM communication channel to allow easy system replication to other remote rural RES plants is another avenue for future research work.

REFERENCES

- [1] "Smaller, cheaper, faster: Does Moore's law apply to solar cells?": <http://goo.gl/qcUcm>
- [2] "Africa: Time to go solar": <http://goo.gl/w0QXt>
- [3] SMA Solar Technologies Monitoring Systems: <http://goo.gl/Bc7bq>
- [4] inAccess Networks Site Controllers: <http://goo.gl/kfwe3>
- [5] Fat Spaniel Solutions: <http://goo.gl/mSfl8>
- [6] Morningstar Products: <http://goo.gl/hCJ5Y>
- [7] SolarMax MaxVisio: <http://goo.gl/OaZYg>
- [8] "SIMbaLink: Towards a Sustainable and Feasible Solar Rural Electrification System", Nahana Schelling, Meredith J. Hasson, Sara Leeun Huong, Ariel Nevarez, Wei-Chih Lu, Matt Tierney, Lakshminarayanan Subramanian, and Harald Schutzeichel, Proceedings of the International Conference on Communication Technologies and Development (ICTD), 2010
- [9] FrontlineSMS: <http://www.frontlinesms.com>
- [10] LORENTZ LA75-12S: <http://goo.gl/bBImt>
- [11] Waspnote by Libelium: <http://www.libelium.com/products/waspnote>
- [12] Phidget Voltage Sensor 1117: <http://goo.gl/vLM1p>
- [13] Phidget i-snail-VC 100: <http://goo.gl/qmGsn>

SESSION 2

CONNECTING RURAL REGIONS

- S2.1 Proposal of a Wired Rural Area Network with Optical Submarine Cables
- S2.2 Development of an ICT road map for eServices in rural areas
- S2.3 Investigating implementation of communication networks for advanced metering infrastructure in South Africa

PROPOSAL OF A WIRED RURAL AREA NETWORK WITH OPTICAL SUBMARINE CABLES

Yoshitoshi Murata[†], Hiroshi Mano^{††}, Hitoshi Morioka^{††}

[†] Faculty of Software and Information Science, Iwate Prefectural University
Takizawa-mura, Iwate, 020-0913 Japan

^{††} Root inc.
Gotanda, Shinagawa-ku, Tokyo, 141-0031 Japan

ABSTRACT

The lack of access to fast Internet services is a serious digital divide for future networks. Most areas outside fast Internet service coverage are rural areas. Many kinds of wireless systems have been proposed for rural areas because of their low establishment cost per residence. Where there are several residences in a small area, a wireless system is effective for establishing a network at a low cost. Our investigation of residence plots in rural areas around Morioka city, Japan, revealed 15 residences on average at intervals of 50–200 m along roads. For such areas, there are no suitable wireless systems. In this paper, we propose a wired rural area network system that uses an optical submarine cable instead of a wireless system. We name it OSC-RAN. One of the goals for the OSC-RAN is to reduce the total cost, which includes both establishment and maintenance costs. We conducted the OSC-RAN field trial and provided trial services for about six months. We verified that residents and/or home appliance installation workers could establish the network, and they longed for the Internet.

Keywords— Digital divide, rural area, wired network, fast Internet, optical submarine cable

1. INTRODUCTION

The lack of access to fast Internet services is a serious digital divide for future networks. Most areas outside fast Internet service coverage are rural areas because such areas are sparsely populated and the cost of establishing a network is higher than in urban or suburban areas. This high establishment cost is a disincentive to investment in rural areas.

A network system that suits rural areas consists of an entrance network and a local area network (LAN). Some wireless systems that consist of WiFi units and directional antennas have been proposed [1, 2] as a low cost entrance network. On the other hand, NTT has proposed the wireless IP access system for a LAN [3, 4]. This system is suitable for areas where it is difficult to use optical cables and the

density of residences is high. Zhang and Wolff proposed a wireless mesh network that uses WiFi combined with a directional antenna [5, 6]. Liu et al. also proposed a wireless mesh network using WiFi combined with a three-sector antenna instead of a directional antenna [7]. These network systems have low establishment costs. When the number of hops is high, it will be difficult to maintain transmission reliability. Some trees and tall grasses will block the lines of sight between wireless access points. In Zhang's study, the hop number was three. They analyzed traffic capacities, but did not perform any experiments. Liu et al. did not perform a long-term experiment, but only a short-term one, and the hop number in their network was also three.

When trouble occurs in a rural area, it takes a long time for running to perform repairs. Therefore, reliability and maintenance ability are very important for a rural area network. It is desirable for residents to be able to repair a rural network by themselves when trouble occurs.

In this study, we investigated residence plots in 28 rural areas around Morioka city in Japan. These areas are outside the fast Internet service coverage. In all of the examined rural areas, we found residences clustered in specified places on a road or at a crossroads. The interval between residences is from 50 m to 200 m and the number of residences per area is 15 on average. These areas are unsuited to the fixed wireless access system since there are a few residences in the area covered by one wireless access point. In the case of a WiFi mesh network, the number of hops needed to cover all residences is too high.

The passive optical network (PON) is usually used for a wired LAN [8, 9]. Since a PON uses an optical splitter that does not convert optical energy into electrical energy, or vice versa, and a PON is a point-to-multipoint optical communication system, it is very economical and ecological for urban areas, especially for condominiums. However, a PON is not economical for sparsely populated areas such as the rural areas that we examined because many optical fibers would have to be deployed.

Therefore, we propose a wired rural area network with the relay units in residences connected step by step by optical submarine cables. We name it OSC-RAN. One of the goals for the OSC-RAN is to reduce the total cost, which includes

*This work was supported by SCOPE from MIAC, Japan

both establishment cost and maintenance cost, by having residents and/or home appliance installation workers establish networks by themselves. We set up an OSC-RAN ourselves, conducted a field trial for six months during winter and spring. Most of rural areas are also outside the terrestrial digital television (TV) broadcasting service, and residences perhaps demand local TV broadcasting to get local advertisements. Hence, we provided the fast Internet access and the local TV broadcasting. We verified that both residents and/or home appliance installation workers could establish a network, and they longed for the Internet access service. The trial OSC-RAN was not encountered any major problems except for a cable being cut by a snowplow.

The residence plots we investigated are described in Section 2. The basic concepts, requirements, and structure of the OSC-RAN that suits these areas are considered in Section 3. The OSC-RAN field trial is described in Section 4. Conclusions are presented in Section 5.



(a) Example of residences clustered near a crossroads.



(b) Example of residences clustered along a road.

Figure 1. Examples of rural areas around Morioka, Japan

2. INVESTIGATED RESIDENCE PLOTS

A rural area network must be designed according to the residence plots. Because it is difficult to judge the difference between a residence and a shed remotely, we visited the investigated areas and measured the number of residences and the intervals between them. We checked for the presence or absence of gutters and ditches containing cables. To measure the intervals between residences we used a distance meter of a car. From our findings, we classified rural areas into five types.

Type I: residences clustered at a crossroads; intervals between residences: 50–100 m (Figure 1 (a)).

Type II: residences clustered along one section of a road; intervals between residences: 50–100 m (Figure 1 (b)).

Type III: residences at a crossroads; intervals between residences: roughly 200 m.

Type IV: residences along one section of a road; intervals between residences: roughly 200 m.

Type V: residences along one section of a road; intervals between residences: longer than 500m.

The number of each type is shown in Figure 2, and the number of areas versus the number of residences is shown in Figure 3. Type II was the most common type. There are 5–14 such residences in 57% of areas, 15 on average.

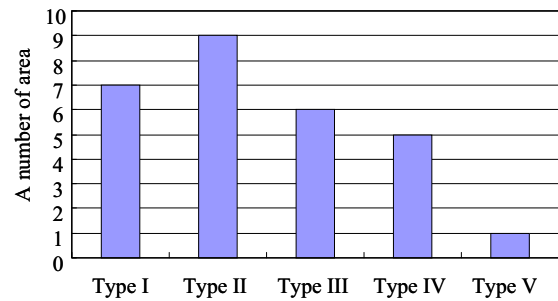


Figure 2. A number of areas vs. rural types

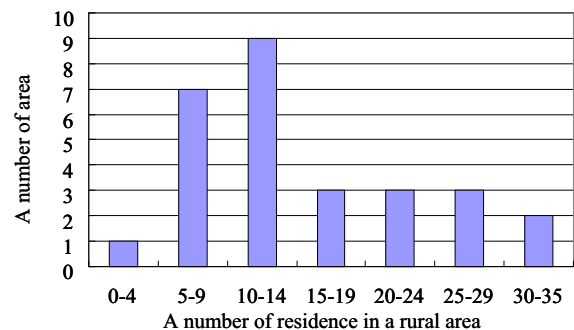


Figure 3. A number of areas vs. a number of residences

3. OSC-RAN STRUCTURE

For our OSC-RAN, we chose a loop topology network for the following reasons: residences line the roads on both sides, the number of residences is not too high, the intervals between residences are not too great for residents to deploy cables by themselves, and self-establishment and self-maintenance should lead to a lower total network cost. The requirements are high reliability, easy establishment work, easy maintenance, short repair time, and low total cost. However, in meeting these requirements, there are some problems:

- Overhead wiring is difficult for residents without a bucket car.
- Optical submarine cables excel in tolerance to shock and water-resistance, but besides being expensive, they are made to order and sold by the drum. It is difficult to obtain cables for replacement quickly.
- Optical fibers can be connected by several methods; the main two methods are mechanical splicing and fusion splicing, but both of these require special knowhow.

We propose a network structure to solve above problems.

3.1. Network topology and relay scheme

We chose the loop topology to maintain reliability. When a cable is cut, the end terminals at both ends of the cable can connect the entrance network with the live-side of the cable. There are two kinds of unit: the control unit and the relay unit. These are connected as shown in Figure 4. The control unit, connects the entrance network to the OSC-RAN, consists of one intelligent Ethernet switch (IES) and three media converters (MCs). The MC converts optical signals into electrical signals and vice versa. The IES is equipped with the spanning tree protocol (STP). When the IES detects a network loop, it lets one port for one MC be live. If it detects a cable cut off, it lets both ports for both MCs be live, and all end terminals continue to connect to the entrance network.

The relay unit contains only an optical switch. It relays data packets from one path to another path and splits off data packets to a personal computer (PC) or a router in a residence via 100BASE-TX. The optical switch also exchanges optical and electrical signals..

3.2. Wiring method and cable

We chose to use not an aerial wiring method, in which cables are suspended between utility poles, but to lay cables on the ground or in gutters for easier deployment. We chose optical submarine cable for a wired cable to keep high reliability considering that cables are on the ground. Optical submarine cables are produced in various different lengths in manufacturing plants and kept in stock. From the field trial described in the next section, we found that cables longer than 150 m are too heavy for residents to deploy, so we utilized lengths of 50 m, 100 m, and 150 m. Each end of

the cable has a factory-fitted SC connector. Cables from these three types were chosen and connected via an optical splice box to adjust the intervals between residences. Cut cable can be repaired by replaced it using a cable from stock.

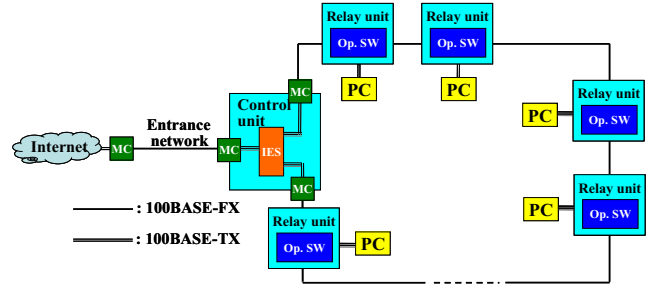


Figure 4. Relay scheme of OSC-RAN

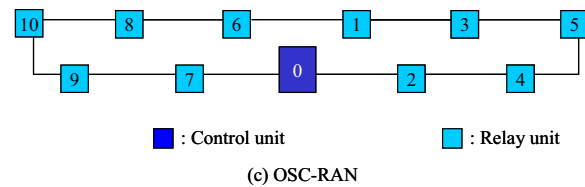
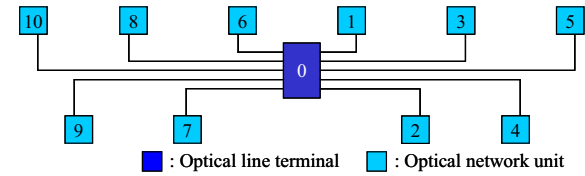
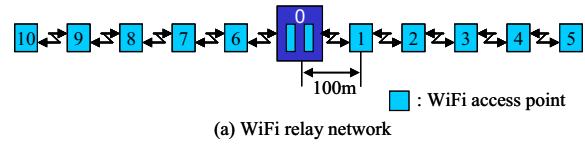


Figure 5. Evaluation models of each network system

Table 1. Evaluations results

	WiFi relay network	PON	OSC-RAN
Equipment cost	Low	High?	High
Construction cost	High	High?	Low
Maintenance cost	High	High?	Low
Reliability	Low	High	High

3.3. Comparison with existing network systems

We compare and evaluate the OSC-RAN with existing network systems, namely the WiFi relay network and the PON. Since most equipment costs are not openly available, and equipment prices and labor costs differ greatly from country to country, it is difficult to evaluate their costs quantitatively. In this paper, we hold characteristics evaluations. The network system structures for this

evaluation are showed in Figure 5. The number of residences is 11, and 5 residences line a road on both sides of the central residence, and the distance between residents is 100 m. Evaluation results are shown in Table 1.

If it were possible to relay more than 5 hops while keeping high reliability in the WiFi relay network, then this network would be cheap and suitable for rural areas. However, it is very difficult to achieve and keep high reliability in the case of more than 3 hops. Even if 5-hop relaying could be achieved, the network reliability would be low and frequent network maintenance would be required.

Since the PON is provided by NTT¹, the network cost and the optical line terminal price are not openly available, and its construction and maintenance costs must be many times higher than those of the OSC-RAN.

If the OSC-RAN were in commercial use, the prices of a control unit and a relay unit would be the same level as WiFi access point prices. The total cable length is about 2.2 km in this evaluation model (c); shorter than that of PON (3 km). However, as existing optical submarine cable is very expensive: about \$10 per meter, the total purchase cost of optical submarine cables is about \$22,000. When a sub-total of unit price is \$3,000, a total cost is \$25,000 in the model (c). If its equipment lifetime is 60 months, the share of the expenses for each residence per month is about \$38. We believe that the total cost of the OSC-RAN is lower than that of PON. If the price of an optical submarine cable were to drop one second in a future, their share would be also about one second. On the other hand, the price of ordinary optical cable is very cheap: about \$0.5 per meter. If aerial wiring is applicable, then ordinary cable should be used.

4. FIELD TRIAL

We conducted the field trial and provided a trial service from December 2009 to June 2010.

4.1. Field trial network

A trial location and a network route are shown in Figure 6. The number of residences in this location is 10. This trial location has four kinds of terrain: a concrete gutter, a creek, forests, and fields. Before the trial, we examined the shock characteristics of two kinds of cable in our university laboratory: an armored cable, shown in Figure 7 (a), and a regular round cable, shown in Figure 7 (b). Both of them can endure the shock of a car passing over it. On the basis of our examination results and their prices, we chose to use the round cable in this field trial.

Most of rural areas are also outside the terrestrial digital TV broadcasting service, and residences perhaps demand local broadcasting services to get local advertisements. Terrestrial televisions channels can not be received well in the trial are, too. Hence, we provided the fast Internet access and the local TV broadcasting. The network configuration to provide these services is shown in Figure 8. We set a TV broadcasting station in Kitakami Cable Television (CATV) office. Since the digital TV

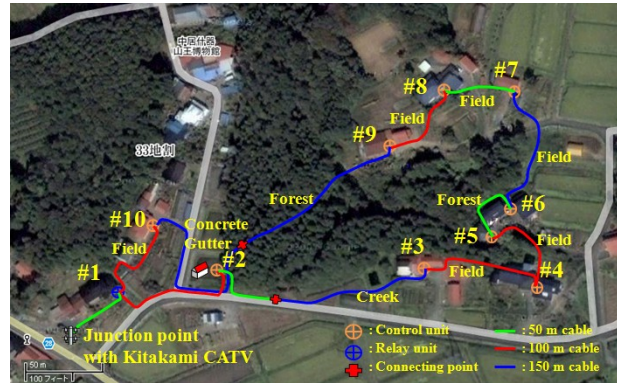


Figure 6. Field trial location

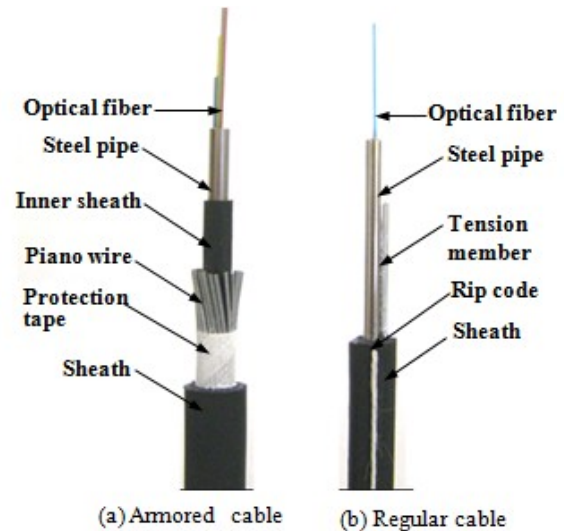


Figure 7. Structures of optical submarine cables

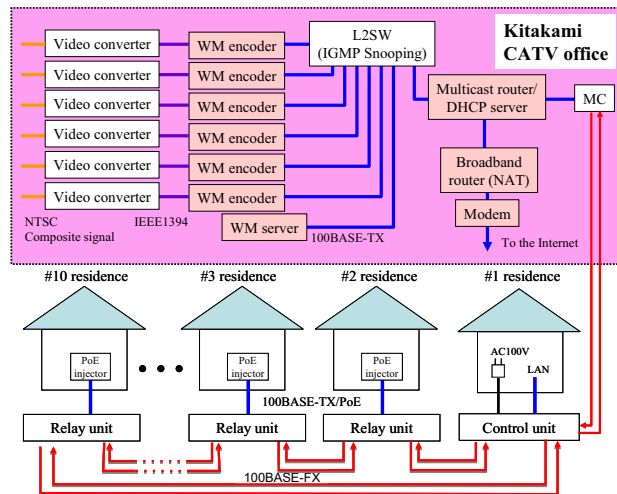


Figure 8. Network configuration

broadcasting was prohibited for retransmission in a low, we retransmitted analog TV channels with the IP broadcasting. Since six channels were broadcasted in a neighboring area, we broadcasted same channels, too. NTSC composite signal from TV receivers was converted to digital video signal,

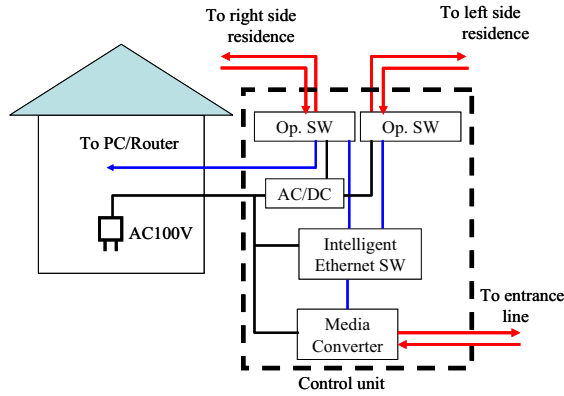


Figure 9. Structure of the control unit

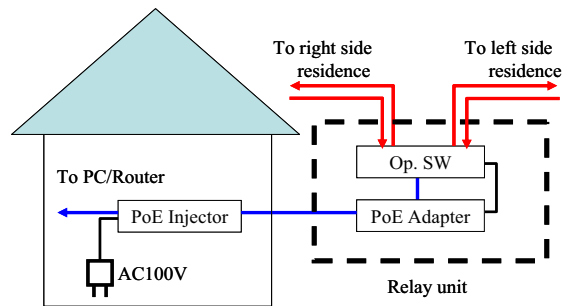


Figure 10. Structure of the relay unit

and encoded to Windows Media format. We used L2SW which supports the IGMP (Internet Group Management Protocol) snooping to broadcast just demanded channels.

Encoded TV signals were broadcasted through the multicast router, converted to electrical signal at MC. The MC were connected the control unit, which was set in #1 residence.

The structure of the used control unit is shown in Figure 9. The control unit has three MCs in Figure 4. But, we did not have MCs, which could not connect an optical switch in the following relay unit this time. Hence, we used an optical switch, which used in the relay unit, instead of an MC. One of the optical switches is connected to a PC in the residence where the control unit was installed.

The structure of the used relay unit is shown in Figure 10. The optical switch consists of a layer-2 switch (L2SW) and two MCs. The L2SW has a packet multiplexing function and supports the CSMA/CD (carrier sense multiple access with collision detection) protocol. This optical switch retrieves the MAC (media access control) address of the connected PC and sends packets to the PC according to their destination address. Since many residences do not have outdoor power sockets available, we used power over Ethernet (PoE) to supply electricity to the optical switch in the relay unit from an indoor wall socket. The user's PC connects to an optical switch through a PoE injector and a PoE adapter.

4.2. Wiring work

The total times for various kinds of work are listed in Table 2. Before wiring cables, we spent two days investigating the route and the intervals between residences. Since all the people involved in this field trial were present on the first

Table 2. A number of hours and persons worked for kinds of work

Kinds of Works	Working date	Working period	Working hours	A number of workers					Subtotal hours
				Kitakami CATV	Root Inc.	OCC Inc.	Univ.	Inst. workers	
Pre-investigation	02/12/2009	11:30 - 13:00	1.5	2	2	1	2	0	15.5
Pre-investigation	09/12/2009	13:00 - 15:30	2.5	0	0	0	2	0	
Cable wiring	11/12/2009	10:30 - 17:00	5.5	1	0	1	4	0	56.0
Cable wiring	13/12/2009	12:20 - 14:00	1.5	0	0	0	1	0	
Cable wiring	17/12/2009	11:30 - 16:30	4	2	0	0	2	0	
Cable wiring	18/12/2009	10:30 - 17:00	5.5	0	0	1	0	0	
Ether cable putting	11/12/2009	16:00 - 17:00	2	1	0	0	0	0	8.0
Ether cable putting	17/12/2009	13:00 - 16:00	3	0	0	0	0	2	
Units installing	18/12/2009	10:30 - 17:00	5.5	0	2	0	0	0	11.0
Confirmation of system	18/12/2009	10:30 - 17:00	5.5	0	0	0	2	0	43.0
Confirmation of system	23/12/2009	10:30 - 17:30	5.5	0	1	1	2	0	
Confirmation of system	25/12/2009	10:30 - 15:30	5	0	0	0	2	0	
Explanatory meeting	16/01/2010	10:00 - 17:30	7	1	0	0	4	0	35.0
Maintenance	19/01/2010	10:50 - 17:30	6.5	0	0	1	1	0	28.5
Maintenance	02/02/2010	14:00 - 16:30	2.5	0	1	0	2	2	
Maintenance	03/02/2010	10:00 - 15:00	4	0	2	0	0	0	

¹Nippon Telegraph and Telephone Corp.

day, this wasted too much time. In retrospect, we think that two people would have been sufficient for this investigation. In fact, one of us (Yoshitoshi Murata) and an assistant measured the intervals of all route again on the second day. We evaluated the residents living in this area and found that measurements required to finish this work took about two hours.

In total, seven persons worked to deploy cables. They consisted of one employee of the cable manufacturer OCC, two employees of a CATV (cable television) company and four university persons. The wiring team consisted of the following roles, as shown in Figure 11.

- Sender: reeled out a cable
- Assistant: corrected any twisting of the cable
- Drawer: drew the cable

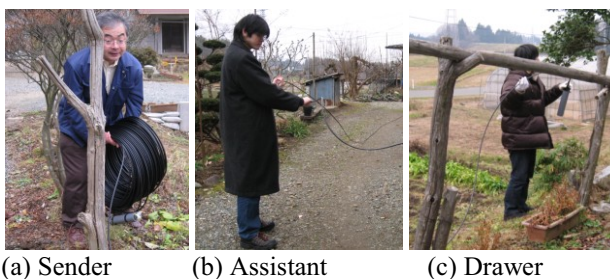


Figure 11. Wiring work

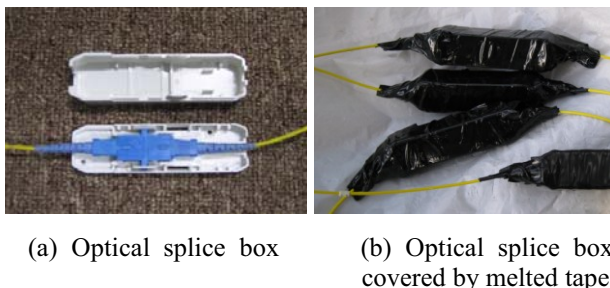


Figure 12. Cable connection

It took about 30 minutes to lay a 50-m cable. We used an optical splice box (Figure 12 (a)) to connect fibers and wrapped it with melted tape (Figure 12 (b)). Each connection took about 20 minutes. Since one cable has four fibers (two are spares), it took about 80 minutes. If there had been a waterproof-type splice box for the optical submarine cable, the working time would have been much shorter.

The control unit was installed in the first residence, and relay units were installed in the other residences.

We used an optical pulse tester and a visible light source to confirm the connection of fibers. In this field trial, a visible light source was very useful. The pulse tester was actually unnecessary since the cable lengths were not so long. We fed the Ethernet cable from the relay unit set up outside of the residence to a PC through an air-conditioning duct. We hired two home appliance installation workers for this work to avoid working troubles.

In total, it took 133.5 hours from the start of the pre-investigation to the end of network confirmation. The reasons for so much time being taken were a shortage of experience and poor planning. If experienced workers work with residents, we think that 50 hours should be sufficient.

We started to wire cables on December 2, 2009, and finished establishing the network on December 25. The reason for taking more than 20 days was that we duties to perform in our university, so the time available for this work was short. At this time of year, the daylight hours are short and the journey to the field site from our university took about 1 hour.

4.3. Network trouble and maintenance

All residences were connected to the Internet on 25/12/2009. However, because of a delay in detecting two problems, the local loop was established on 03/02/2010. One problem was that an SC connector extension fiber in the submarine cable was cut between the 1st and 10th residences. During this trouble, every PC continued to connect to the Internet, if there were not any other troubles in cables. The other problem was that STP of an IES in the control unit was invalid. Therefore, if a cable had been cut somewhere before 03/02/2010, PCs in residences between the 10th residence and the break would have been cut off from the Internet.

We provided a trial service until 05/05/2010. All the network problems that occurred during this period are listed in Table 3. We had seven problems besides the two mentioned above. Three of them were related to the OSC-RAN, and four of them were related to TV broadcasting equipments and so on in Kitakami CATV office. At first, we explain about problems related to an OSC-RAN. The first problem was that a resident pulled the electrical socket off a PoE injector. We asked the resident not to pull the electric socket off again, changed to using a different wall socket in another place, and attached a warning tag. The second problem was that a submarine cable was broken by a domestic (personal) snowplow. Because of the large quantity of snow, we judged that it was dangerous to replace the broken cable with a spare one. Therefore, we repaired the cable with mechanical splices and a splice box. Repairing during the snow season is a topic for further study.

During these two problems, all PCs in residences after the break point were cut off from the Internet, so we checked all cables using a visible light source again. As a result of this reconfirmation, we discovered the SC connector extension cable cut off a submarine cable between the 1st and 10th residences (the fourth problem). After replacing the broken fiber with a spare fiber, all PCs were cut off from the Internet. We guessed that a control unit had a problem and discovered that the STP of the IES had become invalid (the fifth problem). After re-setting the STP, we noticed the sixth problem, pictures sometimes stopped in lower residences. We measured the throughput for each residence, and discovered the auto-negotiation function of the control unit did not work well. We fixed the transmission rate 100 mbps this time.

Table 3. List of problems

Blue colored issues are related to OSC-RAN, and no colored issues are related to broadband distribution equipments.

No	Date of awareness	Details	Reasons	Repair measures
1	05/01/2010	No Internet and TV access at #9 and #10 res.	The electric socket of a PoE injector was pulled off.	- Requested resident not to pull off the socket. - Attached a warning tag. - Used a different wall socket.
2	16/01/2010	No Internet and TV access at residences #7 - 10	The cable between residences #6 and #7 was cut by a snowplow.	Repaired the broken cable with a mechanical splice and a splice box.
3	19/01/2010	No TV access at any residences	When TV broadcasting comes to an end and a video converter stops to transmit signal, Windows media encoders maybe stop.	A WM encoder PC was rebooted every early morning automatically.
4	19/01/2010	A SC connector extension cable between #10 and #1 was broken.	Same as for the left issue	Replaced the broken fibers with spare fibers.
5	19/01/2010	No Internet and TV access at any residences except for #1	STP of the control unit was invalid.	Set the STP to valid.
6	19/01/2010	Pictures sometimes stopped in lower residences.	Auto-negotiation function of the control unit did not work well.	Transmission speed of the optical switch in the control unit was fixed 100Mbps
7	25/01/2010	No Internet and TV access at any residences	The modem to connect the Internet stopped.	The modem was restarted.
8	01/03/2010	No TV access at any residences	The Windows media server stopped.	The Windows media server was rebooted.
9	01/04/2010	No Internet access at any residences	The Internet router of Kitakami CATV stopped.	The Internet router was restarted.

It took too much time to discover the above problems. The reasons were a shortage of preparation and absence of a check list for this network before the start of service provision.

Four problems occurred in Kitakami CATV office. We provided the Internet access and the TV broadcasting, but did not keep watch on these services. After receiving complaints, we discovered and repaired problems. Hence, we wasted too much time. Since a resident usually becomes a member of commercial services, these types of problems must be solved by operators of each service.

4.4. Transmission characteristics

We measured the throughput and transmission delay in the case of transmission control protocol (TCP). The measured data is shown in Figure 13. The transmission delay (round trip time (RTT)) became cumulatively bigger when the number of relays became higher. For TCP, the number of relays became higher and the throughput became lower. An

example of this characteristic is shown in Figure 13. However, the lowest throughput was more than 50 Mbps.

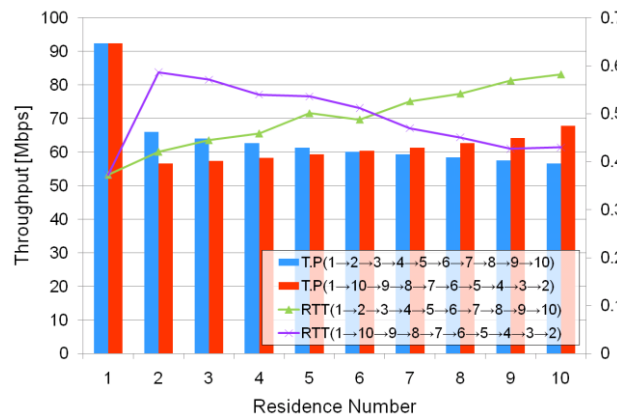


Figure 13. Throughput and transmission delay

4.5. Questionnaires to residents

We sent out questionnaires to residents about services we provided. Questionnaire items and results are as follows.

Q1: How many times did you access to the Internet?

A1; Almost every day: 2 residences
4-5 days/weeks: 2
2-3 days/week: 3
Few days/week: 0
No days/week: 2
No answer: 1

Q2: How many times did you access to the TV service?

A2; Almost every day: 0 residences
4-5 days/weeks: 0
2-3 days/week: 3
Few days/week: 4
No days/week: 2
No answer: 1

Q3: How did you evaluate the quality of pictures in the TV service?

A3; Very good: 1 residence
Better: 1
Fair: 3
Worse: 1
Bad: 1
No TV access: 2
No answer: 1

The Internet access service was used frequently in seven residences. These families include some young or middle-aged persons. All of the people in the remaining three residences are elderly and not interested in information and communications technology (ICT). We think that most people except for very old people will want to use the Internet. On the other hand, fewer people accessed the TV service than the Internet access. This is probably because that they already have satellite TV receivers with much larger screens than the display provided with our service, and demands of getting local advertisements is not so strong.

5. CONCLUSION

As the result of investigating residence plots in rural areas around Morioka city, Japan, we found that residences are clustered at specific portions of roads and crossroads. The length between residences is roughly from 50 m to 200 m. The number of residences per cluster is 15 on average. Since most wireless systems for rural areas are unsuitable for the investigated areas, we set up the OSC-RAN with the relay units in residences connected by optical submarine cable step by step. One of goals for the OSC-RAN is to reduce the total cost, which includes both establishment and maintenance costs, by getting residents and/or home appliance installation workers to establish networks by themselves. Through a field trial, we confirmed that residents and/or home appliance installation workers could

indeed establish and maintain a network, except during the snow season, and they longed for the Internet. We hope that the OSC-RAN will decrease digital-divide areas outside future networks and services, and more people will be able to access to the Internet.

REFERENCES

- [1] Daniele Trincherio, Alessandro Galardini, Riccardo Stefanelli, "Comparative analysis of extended geographical wireless networks based on Diversity transmission systems," ITU-T Kaleidoscope event 2010, Poster session P.2, December, 2010.
- [2] Riccardo Stefanelli, Alessandro Galardini, Daniele Trincherio, "Reliability and Scalability Analysis of Low Cost Long Distance IP-Based Wireless Networks," ITU-T Kaleidoscope event 2009, S4.4, August, 2009.
- [3] K. Nidaira, T. Shirouzu, M. Baba and K. Inoue, "Wireless IP access system for broadband access service," International Conference on Communications IEEE, Vol. 6 WC15-1, pp. 3434-3438, June, 2004.
- [4] T. Shimizu, et. al, "A Study of Advanced Broadband Wireless Access System to Address Digital Divide Issues," IEICE Technical Report, RCS2006-42, pp. 43-48, June, 2006.
- [5] Mingliu Zhang, Richard S. Wolff, "Crossing the Digital Divide: Cost-Effective Broadband Wireless Access for Rural and Remote Areas," IEEE Communication Magazine, pp. 99-105, February, 2004.
- [6] Mingliu Zhang, Richard S. Wolff, "Using Multi-hop for Broadband Fixed Wireless Access in Rural Areas," The 16th International Conference on Wireless Communications, July, 2004.
- [7] Xiaohan Liu, Hiraku Okada, Kenichi Mase, "Performance of Wireless Mesh Networks with Three Sector Antenna," 2010 Sixth International Conference on Mobile Ad-hoc and Sensor Networks, pp. 146-153, December, 2010.
- [8] K. Ochiai, T. Tatsuta, T. Tanaka, O. Yoshihara, N. Oota, and N. Miki, "Development of a Gigabit Ethernet Passive Optical Network (GE-PON) System," NTT Technical Review, Vol. 3, No. 5, pp. 51-56, May, 2005.
- [9] Junichi Kani, Ken-Ichi Suzuki, "Standardization Trends of Next-generation 10 Gigabit-class Passive Optical Network Systems," NTT Technical Review, Vol. 7 No. 11, November, 2009

DEVELOPMENT OF AN ICT ROAD MAP FOR ESERVICES IN RURAL AREAS

Nobert Rangarirai Jere, Mamello Thinyane and Alfredo Terzoli

Computer Science Department, University of Fort Hare, Telkom Centre of Excellence in ICTD,
Alice, 5700, South Africa:

Emails: njere@ufh.ac.za; mthinyane@gmail.com; a.terzoli@ru.ac.za

ABSTRACT

ICTs, driven by the convergence of computers, telecommunications and traditional media, are crucial for the knowledge-based economy of the future. The rapid technological changes have resulted in different ideas being suggested for the expected ICT applications. As a result, different e-Service applications have been developed as a way to foster ICT developments. However, ICT applications deployed at the moment may not be able to sustain the rural communities in maybe 10 years or more to come. The paper considers the past, analyzes the present and conduct surveys to gain insight into the future. Based on all of this information, the research tries to provide an ICT road map for what is to come. What kind of applications can we develop now to cater for the technological changes, so that the ICT applications developed today would still be compatible with those developed in years to come? The Siyakhula Living Lab (SLL) is used as the case study in this paper and some interviews and literature review are done to get different ideas on the future of ICTs.

Keywords—e-Services, ICTD, future of ICTs, future of Technology, ICT road map

1. INTRODUCTION

Different individuals and organizations have anticipated the future of technology and some of the proposed ideas are: For example, some digital giants speak about cloud computing and looks forward to a combined platform for TV, radio and the web [1, 2]. Some talk of the power of mobile applications and how they can enable meaningful human relationships. Others suggest that wireless capabilities in the future of technology will improve drastically [2]. While some anticipate that the evolution of the web will fuel small business formation, operations and innovation, especially as technology becomes cheaper and social networking and virtual worlds become more popular [3, 4]. These projections have a significant impact on ICT applications and should be considered in implementing ICT projects. We appreciate that all these future expectations have an effect on the ICT platforms and the fact that there is a lot of work been done to improve the service delivery through ICTs in developing countries. However, there many challenges that makes it difficult for the future projections to be achieved in all parts of Africa.

2. RESEARCH OBJECTIVES

Having considered the past, present and future, it is clear that there are various changes in terms of technology. There are several projections for the future. Of course, no one knows what exactly the future is, but we better try to be prepared for the future. ICT applications developed today should be able to be useful for the next 5-10 years, thus the future should be considered. This paper moves the concept of the wait and sees approach to getting prepared for what is to come. This could enable the sustainability of ICT services in rural communities.

The authors attempt to ensure sustainability of ICT applications by laying out a foundation where applications could be based. This is achieved through detailed analysis and experiment on the different architectures which accommodate some of the proposed future applications as technology changes. Some of the tangible and research objectives include:

- Coming up with the blue print projections for ICT4D in Africa
- Choosing the best projections which works in African rural ICT4Ds
- Coming up with the best architectures which could accommodate the technological projections
- Deciding on specific applications that can be developed on the architecture developed and a business model
- Developing an ICTD technological road map for eServices in rural areas

Through literature review the authors appreciate that there are different ICT road maps which are available in different parts of the world for specific areas. Some of the ideas on the existing ICT road maps will be used, but in this research the focus is to target the eServices through ICTs for rural communities.

3. RESEARCH METHODOLOGY

The methodology we have used to anticipate the future of ICTs is built on a documentary analysis of literature concerning the issues of e-Service and administration, by examining various web sites of government and cities including case studies in order to have a broad understanding of contents and services available at the time of the study. The case study approach allows the development of an in-depth empirical inquiry of the subject. And also allows the authors to identify the different e-Services that are currently available and different views of

people working in particular ICT areas. Data for this study include a variety of personal experiences of the co-authors, interviews of selected individuals working in ICTs and the rural community members. The research methodologies include:

- Literature review
- questionnaires
- Architecture development
- Engage rural community users

4. CASE STUDY SLL

The e-Services provision explained in this paper are done within the context of the Siyakhula Living Lab¹ referred in this paper as (SLL) undertaken in Dwesa [5]. The name Siyakhula, means that we are “growing together”. The University of Fort Hare (UFH) and Rhodes University (RU) run it jointly, both Universities are located in the Eastern Cape Province of South Africa. The mission of the Siyakhula Living Lab is to develop and field-test the prototype of a multi-functional, distributed community communication platform for deployment in marginalized and semi-marginalized communities in South Africa [5, 6]. SLL aims to develop the marginalized community by equipping people in the area with the necessary technological skills to be able to support projects deployed. It shows how marginalized communities that are very difficult to reach, may in future be joined with the greater South African and African communities to the economic, social and cultural benefit of all [7].

The initial objectives of the ICT4D intervention in Dwesa were to develop a prototype of an eCommerce² platform for the arts and crafts entrepreneurs in the community, and also for the possible exploration of micro-tourism potential in the area [6]. The introduction of the eCommerce aspect to the economic activities in Dwesa was aimed at activating the community towards greater involvement in economic activities in the region, but also at opening up the market base to incorporate wider international customers [8]. An eCommerce portal was developed in direct interaction with the local arts and crafts entrepreneurs to integrate their specific needs and requirements into the platform. The portal was developed around a metaphor of a mall (i.e. an eMall) in which the different sellers have a store that they manage and that they are responsible for [9]. Besides the deployment of the eCommerce portal in Dwesa, a number of other projects have been undertaken within the context of facilitating the implementation of the ICT4D intervention:

A key component of a successful ICT infrastructure in Dwesa is the telecommunication network. A project was undertaken to setup a WiMAX based, local loop to connect the different points of presence in Dwesa [9]. At the moment, the schools are the key points of Internet access as

they are some of the few places within the community that have electricity. Another project related to the networking in Dwesa was exploring the back-haul connectivity options to the Internet and their associated costs and benefits [10]. The resultant back-haul connectivity is provided via a VSAT satellite link [10]. There is a current project addressing the issue of ensuring robustness and effectiveness of the network, through exploring options for redundancy, remote network monitoring and management, and fault-tolerance mechanisms within the network.

The sustainability of the project is a key objective that has been explored from a financial, technical, social and cultural point of view. As a result, research is under way to develop cost-sharing systems and models for the community wherein the different members can contribute in various ways to the upkeep of the deployed infrastructure [10]. From the technical sustainability point of view, researchers are involved with developing a help-desk system to be utilized in capturing and disseminating the problem solving knowledge that has been accumulated in Dwesa. Other research is associated with studying and implementing the culture specific requirements as far as the Human Computer Interface (HCI) components of the deployed systems are concerned.

The key benefit of deploying the computing network in Dwesa is the ability to provide different value-added, relevant services for the people in Dwesa [10]. Currently there are projects under way exploring the development and implementation of eHealth services, eGovernment services, and eJudiciary service for the community. An internal Voice over Internet Protocol (VoIP) communication network has been set up in the Dwesa. Further investigation is underway of the development of a framework for user-driven telephony (VoIP) services. This infrastructure enables the rapid-service development and deployment on the telephony infrastructure, by exposing interfaces that are simple, intuitive and geared towards the (generally) technically illiterate communities [10]. The initial deployment of the services in Dwesa was centralized and predominantly web-based. Some of the above mentioned service portals are accessed primarily through a web interface to a server deployed in one of the schools in Dwesa [9]. This initial deployment paved a way for a subsequent web-services based evolution.

4.1. Summary of the SLL eServices

eCommerce: this service was developed to modify the traditional telecenter model. Its main aims are to expand the financial revenue spectrum by exposing local business men and women products to the rest of the world (since they only depended on local revenues) and in the process exposing the same business people to the use of ICTs for business purposes [11]

eGovernment: Dwesa is located at 47 Km outside of Willowvale. In order for community Dwesa to access basic public services (e.g. applying for an identity document),

¹ www.dwesa.org

² www.dwesa.com

they have to travel the distance. Disadvantaged rural communities in South Africa have suffered from little or no public services provision. The eGovernment system was developed to remedy such issues [12].

eJudiciary: Dwesa legal issues have been always discussed at the chief's house under a tree, and once the issues have been resolved, the people are dismissed. A few months down the line, a little is remembered concerning the matter. The eJudiciary offers a remedy to that by making available a way of safeguarding vital legal data, and guaranteed persistence and availability of data. It also makes legal information available to the public [13].

eHealth: The successful development and implementation of this project provides the community of Dwesa with an e-Health portal that allows them to access a medical ontology that is part of this project [14]. Ontology is a tool that can be used to facilitate communication between people, organization and software. This ontology is based on Xhosa traditional medicine. Adding to that, this project seeks to provide the community with a portal that allows them to browse health information from the Internet and also from the Department of Health [14]. The success of all these projects is from the support we get from the Dwesa community and ICT training programs which are conducted in this area [14].

4.2. Current problems on the deployed SLL application

The following problems were noted through the authors' observations and informal interviews with some community members in Dwesa:

- Each application is a standalone
- Community members feel the applications are too many to master all of them as there is need to know different URLs for different applications
- Training on ICTs and deployed application on a daily basis
- The network is still managed with the universities
- Expense incurred by the universities for the internet accessibility Dwesa community
- The other 3 schools relies on one school where the Vsat is deployed for internet
- No access to computers when schools are closed
- The teachers are in full control of the resources creating some conflicts with the community members
- The future was not considered during the deployment of the eService e.g all the applications are computer web based none runs on mobile phones.

4.3. Common Challenges in rural communities

In any developing country the characteristics and challenges faced by the people staying in rural areas are almost the same [15]. Some of the key characteristics are listed below as explained by Pitke [15, 16]:

- Lack of basic facilities such as water, proper roads and reliable electrical supply;
- Lack of technically skilled people;
- Environmental obstructions such as hills and valley; making the construction of telecommunications networks very costly;
- Bad climatic conditions that damage equipment.

According to a report by the Rural WINS by Groenveld, the following 4 barriers should be addressed by Broadband ICT data-intensive communication applications and services for rural areas [17]:

- **Distance barriers** are the general factor in intermediate and remote rural areas that influences increased costs of business and entrepreneurs' endeavour, transport and cultural activities and negatively affect the quality of rural life [17]. It plays its role in an access of rural inhabitants to cultural and shopping centres, administrative and governmental structures, educational facilities, social and health services.
- **Economic barriers** in access of inhabitants of rural areas to wider business and labour markets (suppliers, customers, opportunities). Producers, when not using local input material, have to import inputs and due to small consumer markets have to export products out of the area, increasing thus the costs burden for their products and services [17]. These barriers can be addressed by a combination of increased awareness about rural areas, eTransactions (eBusiness, eBanking and eLogistics) and by increased development of production
- **Information barriers** – currently the amenities of many rural areas are "invisible" to the "outside world" (inhabitants of other areas, urban centres or citizens of other states – rural tourism, local products [17]. To overcome information barriers in this sense means to implement ICT to enable a full bi-directional access of rural inhabitants to information via data and voice services e.g. Internet and at the same time to increase the awareness of the world outside of the rural area, of its amenities and opportunities for business and tourism, cultural traditions and recreational facilities [17]. The expected outcome is that more business and tourist visitors will come to rural areas to invest and/or spend their money [17].
- **Social barriers** of rural inhabitants to information, education facilities, health and social services etc. An application or service, to be of interest to the RURAL WINS project, should be intensive in data communication and should help to overcome some of these barriers. A consumer, to benefit from an application or service is expected to be computer literate [17]. An appropriate training in computer literacy and application's use is a necessary precondition for a successful introduction of a service or application into practical use.

4.4. Efforts to improve rural communities

Besides the SLL explained in this paper, there are many government, Non Governmental and individual initiatives which are done to improve ICTs in rural communities. Such projects motivate us in writing this paper. The Department

of Science and Technology is currently implementing a national ICT research development and innovation (RDI) strategy which will target rural and marginalised communities. This development was started in 2010. Science and Technology Minister Naledi Pandor said that the strategy seeks to ensure the development of high-end skills to enable, build and strengthen the innovation chain and the capacity of South Africa to perform competitive research in ICT [18].

"Among these is the Digital Doorway, which is a robust computer facility, designed to provide access to computing resources to these disadvantaged communities," the minister said [18]. Another project from the ICT RDI implementation programme, that seeks to enhance access to ICT in rural areas, is a large scale technology demonstrator pilot project, which seeks to deploy affordable broadband connectivity infrastructure.

In the same year 2010, R28 million has been approved to partially fund internet access networks for rural higher education campuses [19]. Higher Education South Africa and the Tertiary Education Network of South Africa (TENET) had requested to further extend points of presence on the existing network to strengthen universities' research and teaching capabilities [19]. Higher Education and Training Minister Blade Nzimande said that funds had been approved because the capacity of universities to conduct research was of great importance as it would allow each university to have all its campuses connected at sufficiently high bandwidths [19]. "This enables shared production and distribution of teaching and learning materials, deployment of centralised administrative systems and processes for the efficient management of multi-campus institutions, access to high performance scientific computing facilities and other educational and research resources via the existing backbone and equitable internet access to other research and education networks globally," Nzimande said.

The Department of Communications has launched a new Information and Communication Technology (ICT) programme to address the country's ICT skills shortage [20]. The National e-Skills Dialogue Initiative (Ne-SDI) which will be implemented by the Meraka e-Skills Institute (Me-SI) will produce workplace-ready ICT graduates. Speaking at the launch on Monday, the department's Acting Director-General Greda Grabe said the Ne-SDI will provide and promote leadership in the area of e-skills development in the country [20]. "It will directly impact on the quality of teaching at the institute as well as the workplace readiness of students." The programme focuses on different categories of e-skills such as ICT practitioner skills, ICT user skills, e-business skills and e-literacy, with a special focus on the social appropriation of ICTs.

All these are government efforts to empower rural areas with ICT applications. A lot of work is done in deploying ICTs to the majority of the areas throughout the entire country. However, we feel there is need to be prepared for

the technological changes and the future of ICT projects has to be planned for. This enables the sustainability of ICTs in rural community. A general platform to cater for the future changes is proposed and several ICT applications are proposed for rural ICTs to cater for the future technological projections.

5. ICT ROADMAP

Based on the future technological projections, general problems in rural areas, different ICT developments and the current status of the e-Services within the SLL and the technological changes, this paper proposes an ICT road map for African rural areas. The road map is proposed as a solution to prepare for the future of ICTs. Road mapping is nowadays recognized as an important strategic planning tool to forecast both the critical development needs and the steps required to reach major advances in an area; and thus provides a valuable tool for decision making.

The main benefit of technology road mapping is that it provides information to make better investment decisions by identifying the critical technologies and technology gaps, as well as identifying the ways to leverage R&D investments for opening new frontiers in the areas [21]. However, a common definition for road mapping or developing a "roadmap" does not exist. Furthermore, the observation of roadmaps that have been so far created indicates that there is considerable diversity among practitioners as to what constitutes a roadmap and the road mapping techniques employed [21]. According to some authors, road mapping is just *good planning* [22]. For Robert Galvin in [23]: "A 'roadmap' is an extended look at the future of a chosen field of inquiry composed from the collective knowledge and imagination of the brightest drivers of change in that field." Another definition from Vähäniitty, et al. [24]: "Road mapping is a popular metaphor for planning and portraying the use of scientific and technological resources, elements and their structural relationships over a period of time. The process of road mapping identifies, evaluates and selects strategic alternatives that can be used to achieve desired objectives, and the resulting roadmaps summarise and communicate the results of key business decisions".

The ICT road map proposed should accommodate the different challenges faced by rural communities. According to Heeks, in an attempt to move from ICT 1.0 to ICT 2.0, the road map provides some of the examples of services which could be provided for rural communities at minimal cost or no cost [25]. This involves working with the rural community members and putting the users at the centre of the road map development. The focus is to propose ICT solutions which could improve the sustainability of ICT solutions in rural areas. The key outputs of the proposed road map are:

- Basic architecture for ICT services in rural areas
- User involvement for ICT road map
- Technical and business models to allow ICT sustainability

- Some examples of services for the community based on the users' views

There is need of a business model concept that everybody understands: one that facilitates description and discussion. The challenge is that the concept must be simple, relevant, and intuitively understandable, while not oversimplifying the complexities of how enterprises function [26]. We believe the technical and business models can best be described through nine basic building blocks that show the logic of how a company intends to make money as said by Alexandra [26]. The nine blocks cover the four main areas of a business: customers, offer, infrastructure, and financial viability.

5.1. The architecture

The ICT road map suggested is summarized on a basic ICT architecture that explains the rural community users as the key stakeholders of the proposed roadmap plan. The ideas of the ICT architecture are as proposed by the New Zealand Government on eGovernment delivery services [27]

- user access;
- user services and guidance;
- service enabling tools;
- accessibility and connection tools;
- technical and business delivery systems

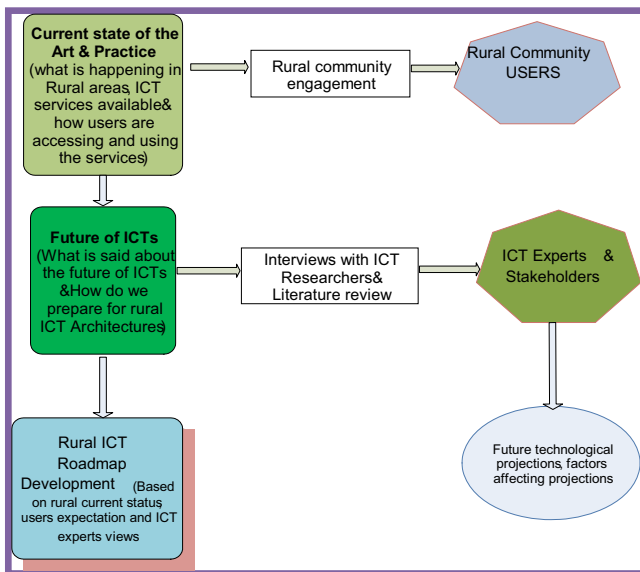


Figure 1. ICT Methodological Approach architecture – User centric: Authors' Perspective

The architecture is part of the ICT road map, which should highlight how the future is going to be prepared for. It considers the technological projections, challenges in rural areas, how services are accessed and provide a basic structure to cater for that. The community members within the SLL are the key users for the ICT road map architectural development.

The architecture is based on different entities such as:

- Rural Users

- Expected ICT services
- Future Technological projections
- User Access
- Different ICT factors in rural areas

There is need to have an understanding on: the current state of the art,

- Then an overview of what is happening in rural communities and
- The changes happening in ICTs
- The ICT architecture considers the future of technology and ICTs through an analysis of different technological projections.
- Different experiments on how to deploy ICT services and identifying the projections which are suitable for rural areas is also part of the approach. The approach means that the rural users remain at the centre of each activity.

The road map explained in this paper is unique in the sense that it caters for ICT service provision in rural areas. The methodological approach used in developing the road map involves the community engagement to get the views of the rural users. This is slightly different from most ICT road maps which are developed without involving the actual users. It also answers how the users are going to access the services and how they will benefit from the services offered to them.

The overall ICT roadmap proposed should be as follows:

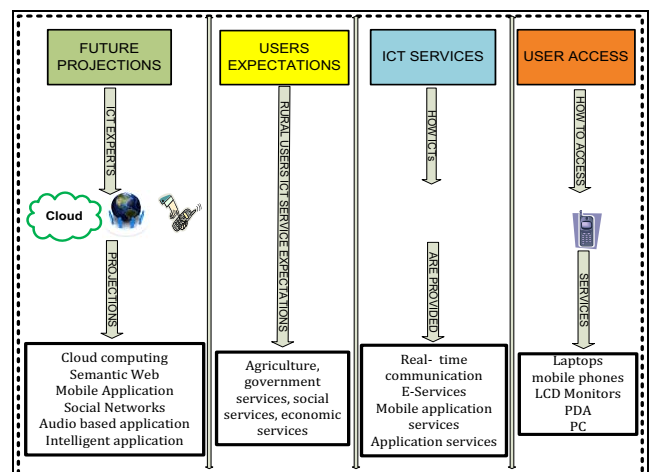


Figure 2. ICT Road Map Components: Authors' Perspective

6. CONCLUSION

Rural ICTs face a lot of challenges and technology is changing drastically. Most of the deployed ICTs fail to sustain rural communities and as a result most of the ICT projects fail. Therefore, we feel proposing ICT road map for the SLL and generic enough to cater for the majority of rural communities could improve sustainability of ICTs and the future of such applications. The ICT road map thus, recommends provision of affordable access to a modern, advanced telecommunications infrastructure and widespread access to end-user equipment;

- ICT skills and competences amongst the population or, at least, the training opportunities through which to acquire them;
- The development and supply of a wide range of ICT applications and services, accessible to both public and private, which meet user needs (citizens, enterprises and public bodies);
- Stimulation of demand (e.g. through awareness-raising programmes) to encourage take-up of services by businesses and citizens; and
- The development of technical and business models applicable for rural communities.

This paper highlights the proposed idea to improve the sustainability of rural ICTs by preparing for the future through an ICT road map. And also the paper gives some key components of the proposed road map. ICT education and training involving the rural community members being the main activities for the success of the road map plan.

REFERENCES

- [1] R. MacManus. 10 "Future Web Trends". 2009. Available: http://www.readwriteweb.com/archives/10_more_future_web_trends.php
- [2] M. Arrington, "Cloud Computing the Future of Microsoft". 2009. Available at: <http://techcrunch.com/2009/09/24/microsoft-ballmer-interview-exclusive-techcrunch-bing-mobile-azur/>
- [3] Collective Impact and Blusik Technologies. "Technology Future Quotes". The catalyst of great results. 2009. Available: <http://www.collectiveimpact.com/page.php?num=95>
- [4] R. Holway. TechmarketView. 2008 Available: <http://hotviews.blogspot.com/2008/02/i-think-there-is-world-market-for-maybe.html>
- [5] Dwesa Project, "Siyakhula Dwesa eCommerce Platform Project". 2008. Available: <http://www.dwesa.org>
- [6] L. Dalvit, M. Thinyane, H. Muyingi and A. Terzoli, "The Deployment of an e-Commerce Platform and Related Projects in a Rural Area in South Africa". International Journal of Computing and ICT Research: 2007 vol. 1, pg. 9-18
- [7] P. Tarwireyi, A. Terzoli, H. Muyingi, "Adapter-based revenue management system for the exploration of non-conventional billing options in new markets for telecommunications". SATNAC conference Wild Coast, Eastern Cape Province, South Africa, 2008. Available: www.satnac.org.za/proceedings/2008/management.htm
- [8] N. R Jere, M. Thinyane, A. Terzoli, "Development of a Reward Based Program for an e-Commerce Platform for a Marginalized Area". WORLDCOMP'09. The 2009 World Congress in Computer Science, Computer Engineering and Applied Computing, Las Vegas, USA.
- [9] R.R. Wertlen, "An Overview of ICT Innovation for Developmental Projects in Marginalised Rural Areas". 2007. Available: <http://ekhayaict.com/eKhayaICT4D.pdf>
- [10] M. Thinyane, A. Terzoli, P. Clayton "eServices Provisioning in a Community Development Context Through a JADE MAS Platform" 3rd International Conference on Information and Communication Technologies and Development, 2009 Carnegie Mellon University in Qatar Education City, Doha, Qatar
- [11] S.G. Njenje, "Implementation of a virtual shopping mall for Dwesa, a rural area in the Eastern Cape", South Africa, M. Sc dissertation, Department Of Computer Science: University of Fort Hare, 2008
- [12] B.T Jakachira, "Implementing an integrated e-Government functionality for a marginalized community in the Eastern Cape, South Africa". University of Fort Hare Department of Computer Science, Alice 2009.
- [13] M.S Scott. "Investigation and Development of an e-Judiciary Service for a Citizen-Oriented Judiciary System in Rural Community". University of Fort Hare Department of Computer Science, Alice 2010.
- [14] B. Hlungulu, M.Thinyane, A.Terzoli. "Building an e-health component for a multipurpose communication centre for a marginalized community using FOSS". Proceedings of the 12th Annual Conference on World Wide Web Applications, Durban, 21-23 September 2010.
- [15] M. Pitke. "The Internet in Developing Countries: Issues and Alternatives". Tata Institute of Fundamental Research, Available: <http://www.isoc.org/inet95/proceedings/PAPER/050>. 2007
- [16] J. Wire. "Meeting the Challenge: Delivering Digital Access in Rural Africa". 2008 Available: <http://50x15.amd.com/en-us/docs/InvencoICIPWhitePapera0407.pdf>.
- [17] P. Groenveld. "Roadmap for Rural ICTs". Preliminary Version,, Deliverable D5.1, RURAL WINS, November 2002 IST 12 adapted from, "The Roadmapping Creation Process," Presentation at the Technology Roadmap Workshop, Washington, DC, October 29, 2002.

- [18] BuaNews, (2010a). ICT Programme to Empower Rural Communities. Available at: <http://allafrica.com/stories/201005250056.html>
- [19] BuaNews. (2010b). Rural Varsities Get Better Connected in South Africa in R28 million funding deal. <http://www.balancingact-africa.com/news/en/issue-no-528/money/rural-varsities-get/en>
- [20] BuaNews. (2010b). ICT programme to address SA's tech skills shortage. http://www.skillsportal.co.za/page/training/training_companies/Information_Communications_Technology_ICT_Training/979313-ICT-programme-to-address-SAs-tech-skills-shortage
- [21] M. Luis, H. Afsarmanesh, Camarinha-Matos, "A roadmapping methodology for strategic research". New University of Lisbon, University of Amsterdam, 2006
- [22] R.E. Albright, "A roadmapping perspective: Science-driven technologies", 2002. Available: http://www.albrightstrategy.com/papers/A_Roadmapping_Perspective-Albright-09-26-02.pdf
- [23] D.R. MacKenzie, S. Donald, M. Harrington, R. Heil, T.J Helms, D. Lund, D. "Methods in science roadmapping: How to plan research priorities", University of Maryland, 2004. Available: www.escop.msstate.edu/archive/roadmap-methods.doc
- [24] J. Vähäniitty, C. Lassenius, K. Rautiainen, "An Approach to Product Roadmapping in Small Software Product Businesses". 2006. Available: <http://www.soberit.hut.fi/sems/QConn-7/QConn-7>
- [25] R.Heeks. The ICT4D 2.0 Manifesto: "Where Next for ICTs and International Development"? Published by Development Informatics Group University of Manchester, UK. 2009
- [26] A. Osterwalder, Y. Pigneur. "12 business model generation". Self Published.ISBN: 978-2-8399-0580-0. 2009.
- [27] Networking government in New Zealand Design of the architecture "Design of the architecture". 2003. Available:http://www.e.govt.nz/plone/archive/services/e-services/service-arch-200303/listing_archives.html

INVESTIGATING IMPLEMENTATION OF COMMUNICATION NETWORKS FOR ADVANCED METERING INFRASTRUCTURE IN SOUTH AFRICA

Monontši Nthontho* | SP Chowdhury* | Simon Winberg*

monontsi.nthontho@uct.ac.za | sp.chowdhury@uct.ac.za | simon.winberg@uct.ac.za

*Department of Electrical Engineering, University of Cape Town, South Africa

ABSTRACT

Advanced metering infrastructure (AMI) is a relatively new field in South Africa. The first standard to govern the envisaged AMI implementation was released in 2008 in NRS 049 document. This paper reports on the investigation of supporting communication networks for AMI. There are several approaches that the South African utilities (Eskom and Municipalities) can follow to implement AMI communication networks. Two broad options are constructing their own private network or connecting with existing network service providers. The communication networks can either use wired or wireless media technologies. They can use mobile broadband networks for a wireless wide area network. Moreover, they can build on and use their legacy optic fibre and PLC networks used for supervisory control and data acquisition (SCADA) application. This paper investigates both wired and wireless technologies that can be considered. Furthermore, it discusses different factors such as bandwidth capacity that would influence the approach chosen.

Keywords— *Advanced metering infrastructure, smart grid, implementation, communication networks*

1. INTRODUCTION

Modern utilities are faced with increasing energy demands. South African power utility, Eskom faces greater challenges due to growing economic activities in the country [1]. While electricity demand is increasing, reserve margins are continuously diminishing. In some cases, demand surpasses installed generation capacity [2]. The utility needs to implement energy management programs on the grid to ensure efficient use of electricity by customers. The future grid is achieved by incorporating advanced IT, communication networks, sensors and smart meters into the conventional power grid. The resulting system is advanced metering infrastructure (AMI). AMI is a subset of smart grid programs whose focus is to provide a two-way communication between the utility and its customers. AMI is thus the main component of a smart grid that interlinks metering data management systems (MDMS) with smart meters in consumers' households or businesses. This interlinking is achieved through communication networks

as discussed further in this paper. Communication networks are thus the backbone of smart grids. They provide a two-way integrated communication between the utility management and sensing devices, grid self-healing capability devices and demand side management (DSM) systems [3]. Smart grid and AIM concepts are broad. They include, but are not limited to, the devices mentioned above. For this reason the scope of this paper is refined below.

1.1. Scope of investigation

Communication networks for AMI are the focus of this paper. The paper discusses investigations on communication network solutions for South African implementation of AMI. In 2008, South Africa defined a standard for AMI in the National Regulatory Services (NRS) 049 standard. This is highlighted in subsection 1.2 below. Several alternative networking solutions are proposed in the standard. Subsection 1.3 summarizes relevant options. This will lead to an experimental investigation that was conducted to determine minimum utilization (bandwidth requirements) of a communication network implemented for AMI. Choice of the communication network type (wireless or wired), capacity of the network, topologies to implement, hence the capital investment required depends on utilization of the network. Knowing data rates that AMI applications require the network infrastructure to have can assist good planning for the infrastructure. It enables the utility to plan for current data traffic profiles as well as future profiles determined by envisaged smart grid applications. NRS 049 and the above-mentioned data traffic profiles that represent network utilisation, serve as guidelines for design and planning of the AMI network. Thus, an overview of the NRS 049 is given below.

1.2. Overview of AMI as defined in NRS 049

NRS specifications are documented by NRS Project Management Agency in collaboration with South African Bureau of Standards (SABS) for Electricity Suppliers Liaison Committee [4]. Accommodating the NRS 049 standard in power distribution networks is the first step for South Africa's realisation of AIM. The standard proposes two architectures for the AMI systems for use in South African markets [5]. Figure 1 below illustrates one such

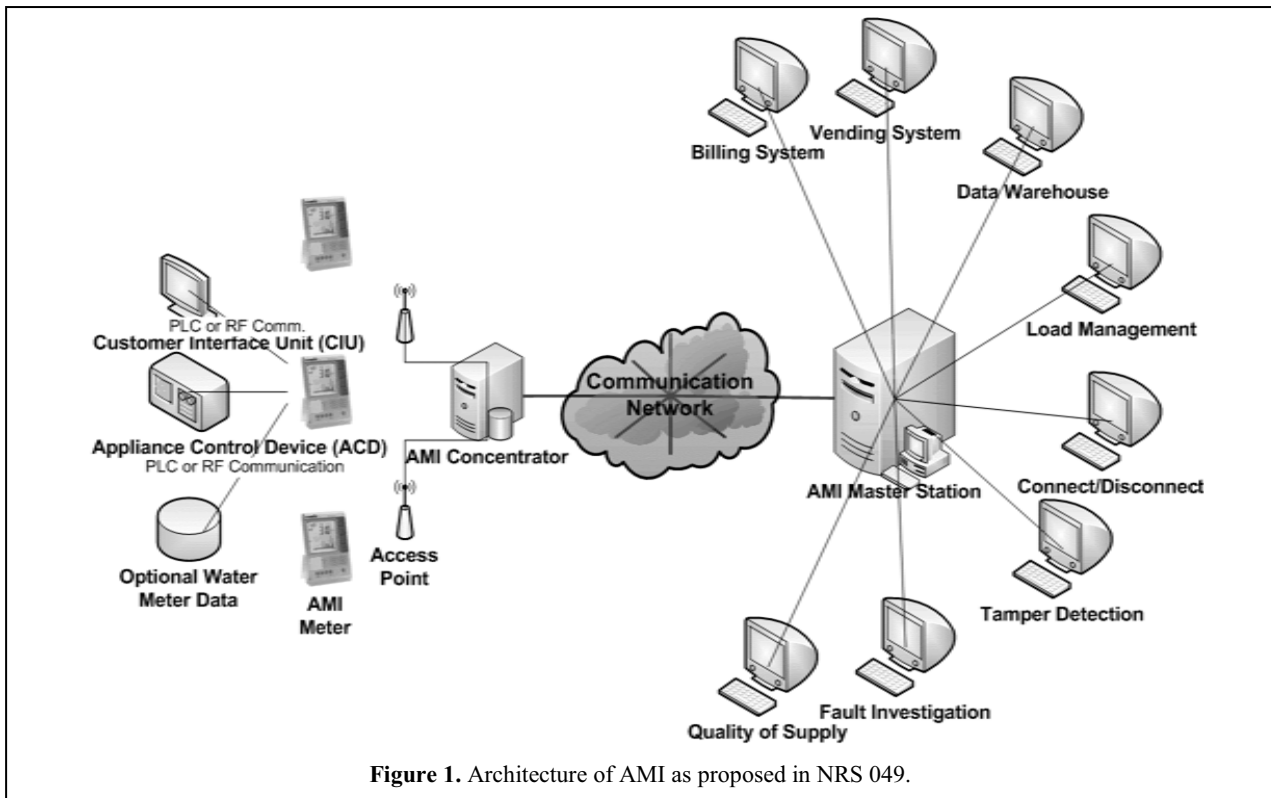


Figure 1. Architecture of AMI as proposed in NRS 049.

architecture of the system. The system consists of the AMI master station, communication networks (represented with the cloud in Figure 1), the AMI concentrator, AMI smart meters distributed in households and businesses, and an optional customer's mobile interface. The AMI meter has functional support components. These are: the customer interface unit (CIU), appliance control devices (ACDs) and an optional water meter [5] [6] [7]. The second model is similar; but it excludes the concentrator. In this model, AMI meters connect directly to the master station via communication networks. The required communication networks integrate each of the major components of the architecture. The components are summarised below.

The Master Station (MDMS) - houses all remote control features of the network. It incorporates storage for metering data information. It allows the control personnel to connect and disconnect loads, manage and monitor load, detect tamper on the equipment, investigate faults, and monitor quality of supply. It also incorporates other applications such as billing system etc. as show on the right side of Figure 1 above.

The AMI meter - The main function of the meter is to register power consumption information of the customer. It also stores time of use (ToU) billing information which can either be retrieved by the master station or by the customer interface unit (CIU).

Customer interface unit (CIU) - enables the customer to see power consumption data and billing information. It displays alarms and load control history.

Appliance Control Devices (ACD) - devices switch controllable appliances on and off according to the time-of-use pattern which will depend on the ToU billing system.

The communication network - is the backbone medium of transmitting information between meters and the master station (wide area network - WAN). Between the AMI meter and the CIU and AMI meter and the ACD (i.e. for home area network - HAN), the model proposed in Figure 1 suggests power-line carrier (PLC) and radio frequency (RF) communication networks [5]. Nevertheless, the focus here is on WAN (indicated by the communication network cloud in Figure 1 above) and local area network (LAN) of smart meters as discussed in subsection 1.3 below.

1.3. WAN and LAN for AMI

AMI WAN media can be wireless or wired network or a combination of both media. The wireless implementation can use mobile broadband network technologies. These are Global System for Mobile (GSM) data services such as: General Packet Radio Services (GPRS) and Enhanced Data Rates for GSM Evolution (EDGE). The newer technology, Worldwide Interoperability for Microwave Access (WiMAX) is also a candidate. Moreover, Wireless Fidelity (Wi-Fi also known as IEEE 802.11 standards) is another wireless technology that can be deployed for the LAN to connect smart meters in households to concentrators that link to the main WAN.

There are five options for wired networks that can be used for WAN or LAN, namely: 1) IP/TCP Ethernet for LAN, 2) Dial-Up using public switched telephone networks (PSTN), 3) digital subscriber line (DSL), 4) synchronous optical

network (SONET) for WAN and 5) existing electrical wiring can itself be used to form part of the network via PLC modulation [8].

Eskom can approach AMI WAN implementation based on mobile broadband networks in two ways. The first option involves building its own mobile broadband network infrastructure. The second option is connecting to broadband internet service providers (ISPs) or directly to network service providers. Figure 2 below illustrates a simplified architecture of an EDGE/3G/4G mobile broadband based AMI WAN. The network consists of three sections: the smart meters, the core mobile broadband network and the utility head-office network. The base station controllers (BSC) and mobile switching centre devices include gateways, IP agents and other systems present in a conventional mobile network. An alteration is in the air interface. In a conventional mobile phone network, the air interface devices would be subscribers' mobile stations. However, in the proposed architecture, the figure shows that the mobile stations have been replaced by smart meters. The smart meters would be connected to the base transceiver stations (BTS) via air interface. The BTSs serving a certain location are connected to a BSC which acts as a concentrator illustrated in the proposed architecture in Figure 1. The BSC is connected to other BSCs, to a mobile switching centre (MSC), a broadband gateway and an IP agent. Another gateway support node (GSN) can be added (not shown in the diagram) to enable the network to interwork with fixed-line networks such as PTSN.

There are several advantages of using an IP based mobile broadband network. Interoperability of equipment from different manufacturers will be achieved [9]. Security applied in existing mobile broadband applications such as banking, military and government will be inherited. Furthermore, being IP based, it allows for the expected huge number of nodes to be identifiable within the network.

Nevertheless, cost of implementing a wireless mobile broadband network, its reliability and performance for servicing critical AMI applications are issues to ponder. A wired network may be a better solution.

For wired networks, there are various options that include PLC modulation, Ethernet, and optic fibre. Eskom already has legacy PLC and optic fibre networks used for control purposes. The legacy networks can also be considered for AMI implementation [3]. However, PLC is restrictive in the range of future AMI services that it can accommodate. For instance, substation surveillance camera traffic and all ultra-broadband services all demand significant bandwidth capacity which PLC cannot provide [10]. Further disadvantages of PLC include interoperability problems and high sensitivity to interference.

For LANs connecting meters in households, a private IP-based Ethernet network is a potential alternative for a wired infrastructure. Figure 3 below illustrates the architecture of an AMI Ethernet network on the customers' home side. Again, the LAN can be connected to an ISP network or to the utility's legacy communication networks: optic fibre and on the PLC infrastructure. A tree topology is suitable for the LAN part of the network because of geographically dispersed structure of households. As shown in Figure 3, the proposed network adheres to the hierarchical structure of the conventional power networks. A power grid can be viewed as a hierarchy with three layers: a power network, the management side, and the customers' side. The proposed AMI network adopts this hierarchy. There is the utility management side where the MDMS is found (head-office network in Figures 2 and 3), the network side of the AMI which incorporates equipment placed at substations and finally, there is the customers' side which has a LAN connecting meters. This modularity makes it possible for the utility of choose to connect with ISPs and own only the two ends instead of instigating their own private backhaul network.

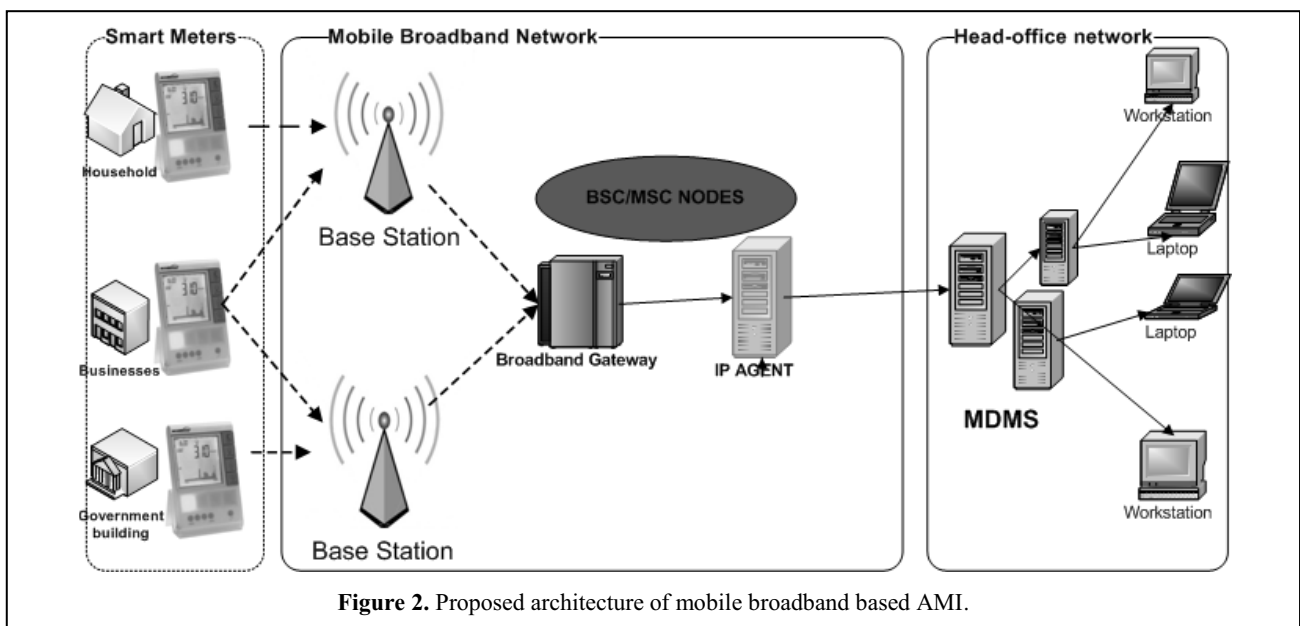


Figure 2. Proposed architecture of mobile broadband based AMI.

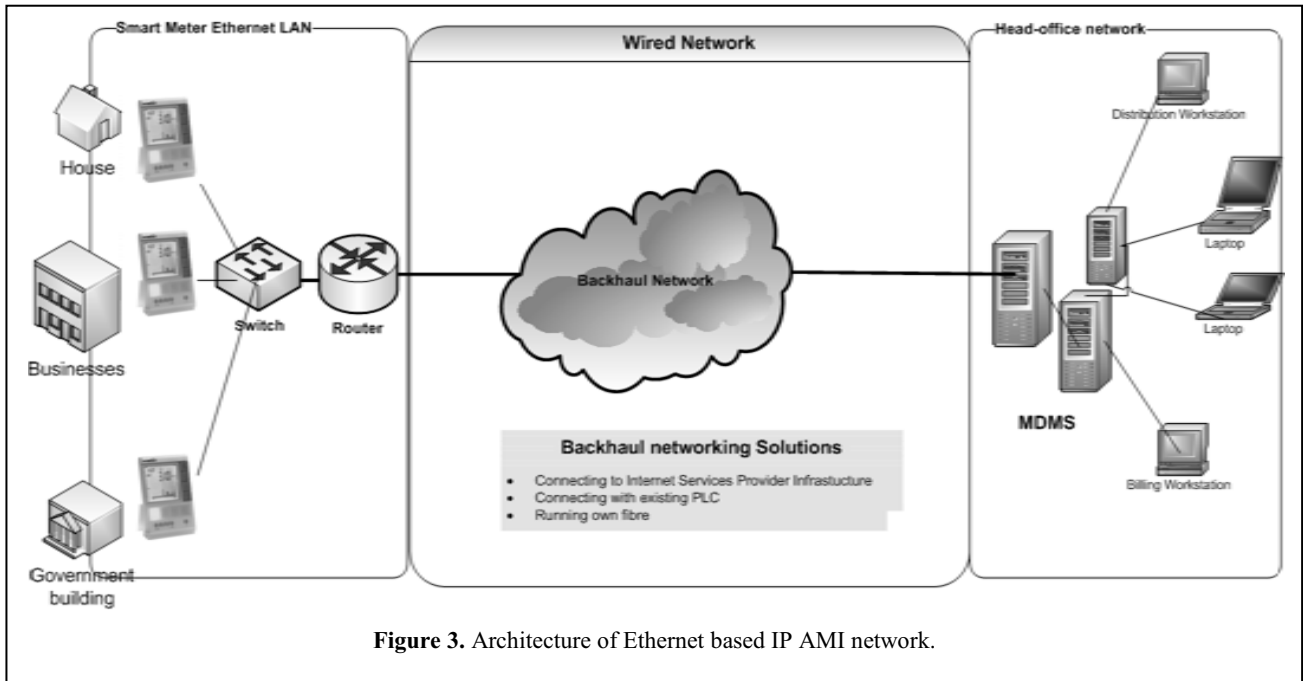


Figure 3. Architecture of Ethernet based IP AMI network.

If the utility chooses to deploy a private network, a suitable and economic architecture would have the components of the communication network overlaid on the existing power network equipment. That is, Ethernet routers are placed at substations, switches or hubs are housed at transformer pillar boxes and communication network cables follow the power network as much as possible.

Wenpeng Luan et al defined three categories of AMI traffic. It is traffic from AMI meters, grid control and monitoring and advanced applications [3]. He further argued that the initial step in planning AMI communication network is understanding traffic profiles of AMI applications in order to decide on the capacity required [3]. While this work attempted to devise a bottom-up approach to planning the network, the network utilisation of 30% used was not based on any experimental data. Utilisation of the network plays important role in deciding on the capacity of the network. Hence it is the focus of the study reported on in this paper.

According to Jacky Dahany of SmartGridNews.com a smart meter sends data of approximately 400MB per year [11]. Based on this information, data rate of a smart meter can be calculated as follows.

$$\text{Data Rate (Bytes/sec)} = \frac{MB}{d \times h \times s} \times 1\,000\,000$$

Where d is number of days per year (365 days), h is the number of hours per day (24hours), s is number of seconds per hour (3600seconds) and MB is data in Megabytes sent in a year. Applying this formula with 400MB of data send per year, each meter sends 12.68 Bytes per second.

A simulation was performed to model an AMI network. The simulation was constructed to study an AMI network for Cape Town and its surrounding cities in the Western Cape. This simulation is described further below.

2. MATERIALS AND METHODS

The simulation performed involved building an Ethernet IP LANs (subnets) on OPNET 14.0 network simulator. The subnets represented smart meter installations in each city around Cape Town. Utilisation of the communication links was determined and then used to estimate bandwidth requirements of AMI. Population statistics of each of the cities were used to estimate the number of households that would have smart meters installed.

The aim of the simulation is to study point-to-point utilisation of the AMI Ethernet network. Since the network needs to cater for future expansion, packet size sent per second was doubled to 24 Bytes instead of 12.68 Bytes as calculated in 1.3 above. The AMI network built assumes a tree topology. Optic fibre and DS0 cables are used for the backbone network (WAN) which spans longer distances. The LAN connection is established through twisted pair cables as these subnets are plus or minus 100 meters long. Figure 4 illustrates a snapshot from OPNET of a simple AMI communication network covering the City of Cape Town and its surrounding towns. The smart meter LANs have been abstracted by the octagonal subnet nodes in the figure. In this simulation, each town (subnet) has thirteen smart meters connected to a switch and the switch is connected to the router in that specific town. The router of each town is connected to Cape Town router with DS0. Optic fibre cable is used for a link between Atlantis and Cape Town. Cape Town router also connects a switch and a server. After the network model was build, a function in OPNET that allows individual statics of the simulation to be collected was used to generate average network link utilisation graphs as displayed in Figure 5. Furthermore, calculated utilisation of these links was then determined and compared to the results in the graphs. These results are discussed below.

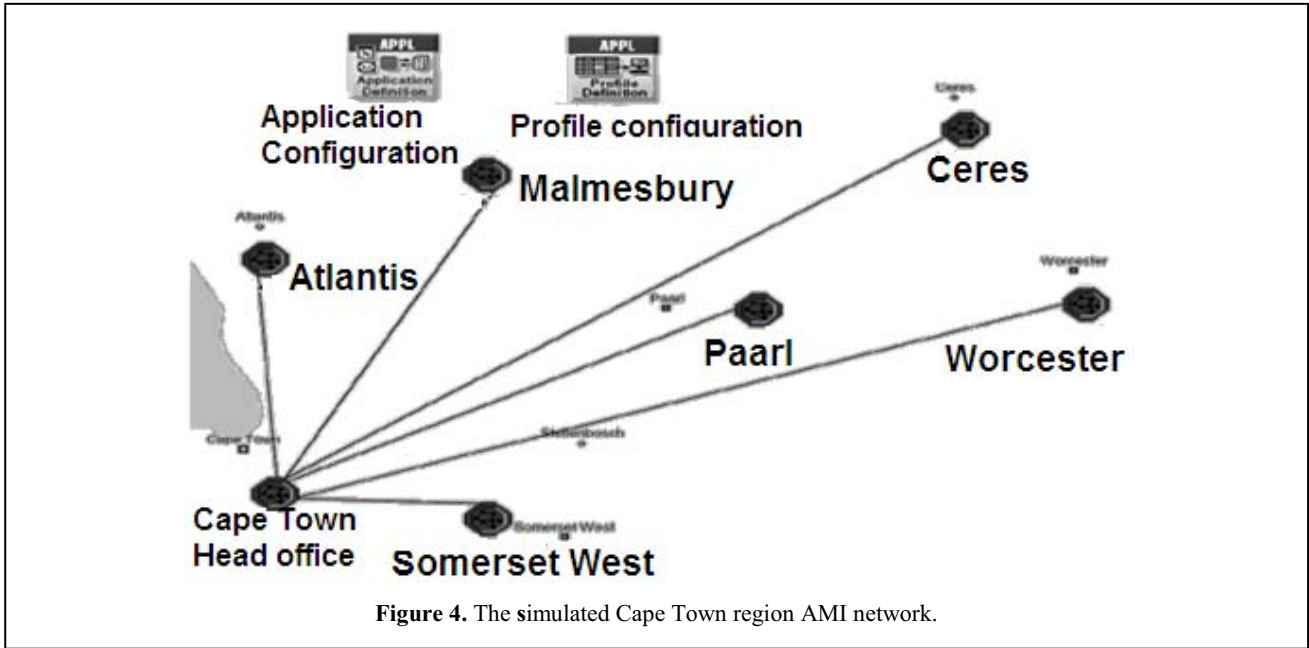


Figure 4. The simulated Cape Town region AMI network.

3. RESULTS

The 24 bytes (which form the size of one packet transmitted by a smart meter) gives 192 bits (8bits *24bytes/1bytes = 192bits). The 192 bits is data transmitted by only one smart meter per packet. For 13 meters in a subnet, the result is 2496 bits (13* 192bits = 2496 bits).

If 64 Kilobits represent 100% of the DS0 cable capacity, then 2496 bits represent 3.9% utilisation as shown by the second graph on Figure 5 below. (2496bits * 100%/64000bits =3.9%). If one smart meter sends 192bits/s, this reveals that one DS0 cable is capable of creating a hop from

a subnet with a maximum of $(64000/192) = 333$ smart meters to the head office with a delay objective of seconds.

Furthermore, using population statistics for Ceres, bandwidth requirements for the Ceres subnet can be estimated. Ceres has an estimated population of 41596 [12]. Assuming each household has an average of 4 people. There are $(41596/4) = 10399$ households in Ceres. If all of these houses have smart meters, the link between Ceres and Cape Town must have a capacity of not less than 1941kilobits/s ($10399/(192bits *1000) =1949\text{Kilobits/s}$) which is a 0.24Mbytes/sec link. Table 1 shows bandwidth requirements for other subnets around Cape Town (calculated as in Ceres above).

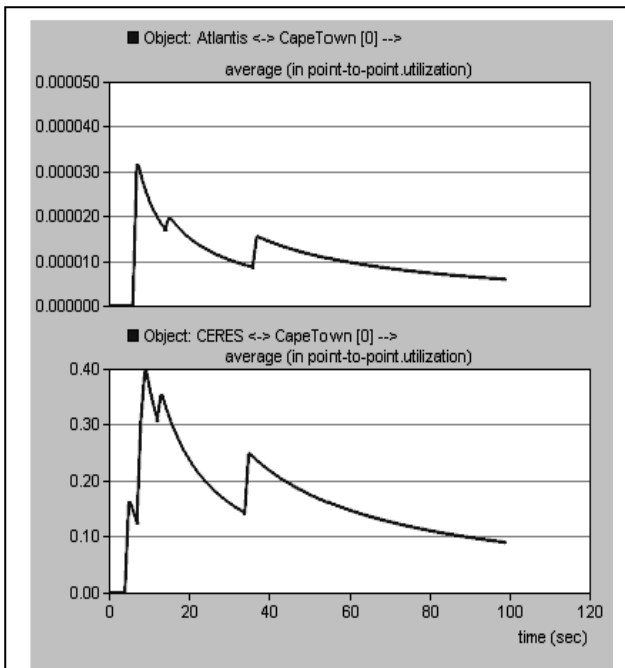


Figure 5. Average point to point utilisation of links.

Table 1. Bandwidth estimations for AMI in Cape Town region.

Town	Population	Houses/ Smart Meters	Bandwidth Mbytes/sec
Ceres	41 596	10 399	0.24
Atlantis	60266	15067	0.34
Cape Town	3433441	858360	19.6
Worcester	127597	31 899	0.7
Paarl	61660	15415	0.4
Somerset West	60000	15000	0.3
Malmesbury	34991	8747	0.2

Having the bandwidth specifications estimated, other factors that need to be considered are discussed in Section 4 below.

4. DISCUSSIONS

Bandwidth requirement is one factor to consider in deciding on the appropriate networking technology that utilities can implement for AMI. Estimations of bandwidth requirement

enable utilities to decide whether or not it will be economically viable to deploy their own communication network or to subcontract their AMI communication activities to existing communication network providers.

From the end nodes' point of view (smart meters) smart metering applications bring a unique characteristic in the networks as they require uplink bandwidth to be greater than downlink. There will be more data flowing from the meters to the MDMS. On the other hand, meter firmware upgrade, remote connect/disconnect messages, critical peak pricing alerts and load control commands data will be flowing sporadically [3].

4.1. Possible AMI networking approaches in SA context

As mentioned in subsection 1.3, the two broad networking technologies to decide from are either wired or wireless. A combination of the two can also be build. There are pros and cons for both approaches. Table 2 summarises features that would need to be taken into consideration.

In terms of data rate requirements, Table 1 showed that AMI applications are not as demanding as other broadband applications such as web browsing. For example, a dial-up connection from a PSTN with a capacity of 56 Kbits/sec is considered slow by most internet users. However, this connection can sufficiently link $(56000/192) = 291$ smart meters to an ISP or WAN.

Table 2. Network media characteristics.

Characteristic	Network Media	
	Wired	Wireless
Installation	Difficult	Easier
Cost	Less	More
Reliability	High	Reasonably high
Performance	Very good	Good
Security	Very good	Good
Mobility	Limited	Outstanding
Speed	Faster	Slower

The mobile broadband networks include GPRS, Third generation (3G) networks with theoretical data rate of 2Mbits/sec (0.25Mbytes/sec) and Fourth generation (4G) standards with theoretical data rates of 70Mbits/sec (8.75 Mbytes/sec) [13]. GPRS network is based on GSM. However, it has lower data rate. A better solution is its evolution which is EDGE. EDGE has a data rate of 384 Kbit/sec. There are many 3G networks that Eskom can consider for wireless metropolitan area networks of AMI. These include different variants of code division multiple access (CDMA) based on both interim standard 95 (IS-95) and GSM. Moreover, there is also high-speed downlink/uplink packet access (HSDPA/HSUPA) and evolution data optimised (EVDO). If high data rates that will cater for future expansion and services are prioritised, 4G networks can be considered. These are IEEE 802.16 standards. The most well known standard is WIMAX [14].

On the other hand, switched network (fixed-line) telephone loops such as variants of digital subscriber line (DSL) wired networks can also be considered for LANs of smart meters. These have data rates ranging from 1.55 Mbits/sec to 3 Mbits/sec [15]. Backhaul network infrastructures that may be considered are synchronous optical network (SONET) and synchronous digital hierarchy (SDH) technologies. However, according to [8], these technologies cannot economically support jittering changes in bandwidth requirements, high resilience, transporting IP traffic, network automation and meeting the requirements of flexible connectivity.

Nevertheless, speed, mobility and performance in Table 2 are not very important deciding factors. The important factors to consider are ease of installation, cost, security and resilience. If Eskom decides to build a private network for AMI, complexity of implementing (installing) the network has a direct effect on the cost hence possible return on capital invested on the network infrastructure. Therefore, in the interest of economic viability, the most important decision that utilities need to make is whether to construct their own network or to utilise network infrastructure of existing communication network service providers.

There are several factors to consider in deciding whether or not to construct a private AMI network. We have highlighted different types of communication networks that Eskom can consider in building a private AMI communication network. That is one approach. The ultimate goal is to come up with the most profitable option offering customer satisfaction, a secure network for grid automation, etc.

The second approach is liaising with existing operators. According to [16], Telkom (a state-owned corporation) and Neotel (privately owned entity) are the major operators in the fixed-line networking services in South Africa. The mobile broadband operators are Vodacom, MTN, Cell C and Telkom's 8ta [16]. Another state-owned operator, Broadband Infracore, also operates inter-city bandwidth. Swiftnet, another state-owned entity and subsidiary of Telkom, operates national wireless data network and wireless data telecommunications services. Telkom offers broadband, data, packaged voice and internet services that are necessary for AMI applications [17].

4.2. Regulatory environment of communications in SA

It is important that Eskom's decision considers the regulatory environment that governs communication network providers. The Electronics Communications Act has revolutionised the communication networks industry in South Africa. It has opened doors to new companies; intensifying competition in the process, driving rapid growth in mobile telephony and connectivity of broadband [17] [16]. The competition and deployment of Seacom marine cable have let to lower broadband prices. Eskom can take advantage of these lower prices by connecting with existing service providers instead of investing in new infrastructure.

4.3. Conclusions and Recommendations

There are several networking approaches that can be considered from both wired and wireless networking media technologies. Most networking technologies are applicable because of less demanding data traffic profiles of AMI. However, stringent security, resilience and reliability requirements of the AMI applications cannot be achieved by current technologies such as dial-up systems offered by PSTN and PLC modulation. More robust technologies are needed.

Eskom should liaise with Telkom and its subsidiaries. Since Eskom, Telkom, Broadband Infraco and Swiftnet are all state-owned entities, they are more likely to have similar mandate of servicing customers. Moreover, these are the biggest communication network services providers in the country. Subcontracting AMI communication network activities to these companies will benefit Eskom as it will harness skills and knowledge vested into these entities. Moreover, competition from privately owned companies (as highlighted in [17]) is forcing Telkom to look for new markets to enter. AMI application can therefore be a lucrative investment with less or no competition.

Issues discussed in this paper are based on a South African context. However, most aspects such as types of network media and whether to build own network or use ISPs can guide utilities in other countries.

4.4. Future work

Smart metering applications are relatively new in South Africa. The NRS 049 standard has only defined an abstract view of advanced metering infrastructure implementation in South Africa. Therefore, more work is still due to decide on communication networks standards that can be deployed for HANs, LANs and WANs of the AMI system. Architectures of these standards need to be defined. Furthermore, the question that still remains is whether or not it is more profitable for utilities to deploy their own communication networks for AMI. Whether they choose to instigate their own network or they decide to connect with existing ISPs or communication network providers, more research needs to be done to define appropriate business models that can be adopted globally. If Eskom pioneers a private network, the simulation has shown that a decent network (depending on the capacity) may not be fully utilised. Thus, they may be forced to enter into new markets to offer network services. A research into how such models can be regulated still needs to be done.

REFERENCES

- [1] Brand South Africa. (2011, May) SouthAfrica.info. [Online]. www.southafrica.info/business/economy/infrastructure/energy.htm
- [2] S. Chowdhury, S.P. Chowdhury, C.T.Gaunt, A. Van Deventer, "Management of Emergency Reserves Dispatching in Electricity Networks," in *Power System Technology (POWERCON), 2010 International Conference on*, Johannesburg, 2010, pp. 1-5.
- [3] Wenpeng Luan, Duncan Sharp, and Sol Lancashire, "Smart grid communication network capacity planning for power utilities," in *Transmission and Distribution Conference and Exposition, 2010 IEEE PES*, Burnaby, 2010, pp. 1-4.
- [4] NRS. (2011, May) NRS online. [Online]. www.nrs.eskom.co.za
- [5] Henri Groenewald, "NRS049 – ADVANCED METERING INFRASTRUCTURE (AMI) FOR RESIDENTIAL AND COMMERCIAL CUSTOMERS," ESKOM, Johannesburg, Presentation 2009.
- [6] Xu Ren Wu-China Electric Power Research Institute TIAN Shiming, "Key Technology Research of CHINA Advanced Metering Infrastructure," in *IEEE 2010 International Conference on Power System Technology (POWERCON)*, vol. SR26, Hangzhou, 2010, pp. 1-2.
- [7] Government of Ontario IT standards, "Advanced Metering Infrastructure," Government of Ontario, Ontario, IT Standard GO-ITS51, 2007.
- [8] Z.M. Fadlullah et al., "Toward intelligent machine-to-machine communications in smart grid," *Communications Magazine, IEEE*, vol. 49, no. 4, pp. 60-65, April 2011.
- [9] Victor C. M. Leung, Jun Wang, "A survey of technical requirements and consumer application standards for IP-based smart grid AMI network," in *Information Networking (ICOIN), 2011 International Conference on*, 2011, pp. 114-119.
- [10] Claude THURLE, "SMART GRID AND AUTOMATIC METER MANAGEMENT: DREAM OR REALITY?," in *19th International Conference on Electricity Distribution*, 2007.
- [11] Bill Sweet. (2009, October) IEEE Inside Technology Spectrum. [Online]. <http://spectrum.ieee.org/energywise/energy/renewables/the-smart-meter-avalanche>
- [12] Mongabay. (2009) Mongabay.com. [Online]. <http://population.mongabay.com/population/south-africa>
- [13] Val T, Fraise P, Mercier J. J, Fourty N, "Comparative analysis of new high data rate wireless communication technologies "From Wi-Fi to WiMAX"," in *Autonomic and Autonomous Systems and International Conference on Networking and Services, 2005*, 2005, p. 66.
- [14] Ghosh A, Sankaran C, Fleming P, Hshieh F, Benes S Fan Wang, "Mobile WIMAX systems: performance and evolution," *Communications Magazine, IEEE*, vol. 46, no. 10, pp. 41-49, October 2008.
- [15] Sistanizadeh K, Kerpez K J, "High bit rate digital communications over telephone loops," *IEEE Transactions on Communications*, vol. 43, no. 6, pp. 2038 - 2049, June 1995.

- [16] Brand South Africa. (2011, May) SouthAfrica.info. [Online].
www.southafrica.info/business/economy/infrastructure/telecoms.htm
- [17] Telkom SA Limited. (2011, May) Telkom Investor Relations. [Online].
<https://secure1.telkom.co.za/ir/sustainability/industry-overview/regulatory-and-competitive-landscape.jsp>

SESSION 3

REFLECTIONS ON A FULLY NETWORKED SOCIETY

- S3.1 Invited paper: Cooperative Wi-Fi-Sharing: Encouraging Fair Play
- S3.2 Making things socialize in the Internet – Does it help our lives?
- S3.3 Net-Centric World: Lifestyle of the 21st Century
- S3.4 Reflexive Standardization of Network Technology

COOPERATIVE WI-FI-SHARING: ENCOURAGING FAIR PLAY

Hanno Wirtz, René Hummen, Nicolai Viol, Tobias Heer, Mónica Alejandra Lora Girón and Klaus Wehrle

Chair of Communication and Distributed Systems
RWTH Aachen University
{wirtz, hummen, viol, heer, lora, wehrle}@cs.rwth-aachen.de

ABSTRACT

Cooperation enables single devices or applications to establish systems that exceed the capabilities of single entities. A prime example for cooperation are Wi-Fi-sharing networks, in which multiple parties cooperatively share their resources, such as wireless access points and Internet uplinks, to form a large-scale Wi-Fi network that offers access to mobile users. Mobile users benefit from this network by gaining free network access at every access point of the network. However, such cooperation needs to be established in the first place by providing incentives to users to join the network. Furthermore, in an established network, users need incentives to behave cooperatively when using the network. Frameworks to provide incentives and to regulate user behavior in the presence of malicious parties can exist at multiple levels: The technical level inside the given network, a contractual level that regulates the operation of the network and the legislative level that establishes general rules for the operation of Wi-Fi-sharing networks. In this paper, we analyze requirements and mechanisms to establish such frameworks at each level and discuss possible solutions and existing examples.

Keywords— Cooperative networking, Ubiquitous networking, Wi-Fi-sharing

1. INTRODUCTION

Cooperation is a compelling concept to surpass the technical and conceptual restrictions of a single device or application. It can enable the creation of new networks and services that a single user or device could not establish by itself. Building on this notion, cooperation is a fundamental principle at different levels in many of today's network approaches. For example, message forwarding and data aggregation in multi-hop wireless networks such as wireless mesh or sensor networks are built entirely on the principle of cooperation between the participating nodes. Peer-to-peer (P2P) networks, on the other hand, use cooperation not between system entities but between users at the application level. The benefits of cooperation in these scenarios include the establishment of communication beyond the communication range of an individual radio and access to otherwise unavailable resources, such as storage space and information.

As in other systems comprised of strangers, cooperative users are tempted to defect, i.e. to behave selfishly and follow

their own interests. As a result, their behavior may stand in stark contrast to the interests of the other cooperating parties and may degrade the overall cooperative system. In addition, evolution theory has shown that, cooperation cannot sustain without any support through regulations [1]. The prime example for defective behavior in P2P networks is "Free-riding", where users exploit the resources of the network without providing resources back to the network. Still, users need to show preliminary trust when providing a service without a guarantee for benign behavior of other users, in order to enable an initial network creation. In the further course of network operation, mechanisms are required for users to check whether this initial investment of trust is justified by the cooperative behavior of other users, with the common goal of ensuring a sustainable cooperation between benign users. In consequence, the risk of defecting users requires *incentives* for individuals to cooperate and a framework that allows to detect defection. In case a user shows *defective behavior* in the network, e.g. stops cooperating, or misuses a given resource, such a framework establishes rules for penalizing or excluding defecting users.

In this paper, we analyze Wi-Fi-sharing networks as a case study for the implementation of different aspects of cooperation. Cooperative Wi-Fi-sharing networks and their services rely on the contribution of Wi-Fi resources and Internet uplinks by participating users. However, the operation of such networks also affects external parties such as Internet Service Providers (ISPs) that operate the wired uplink for the shared wireless network. Hence, analyzing Wi-Fi-sharing networks with regard to cooperation requires to look at mechanisms for a regulatory framework of cooperation at multiple levels. We identify three hierarchical levels, as shown in Figure 1, on which frameworks for user cooperation can be established: i) the technical level, ii) the contractual level, iii) and the legislative level. As Wi-Fi-sharing networks account for the interests of the different participating parties as well as the respective judicial framework, single networks typically differ on the technical and contractual level. As such, no standardized scheme for Wi-Fi-sharing networks exists as of now. A push towards standardized Wi-Fi-sharing could be provided by ISPs that propose network modes and technical mechanisms for standardization. Standardized frameworks and mechanisms would provide additional support for the acceptance and interoperability of Wi-Fi-sharing networks.

Figure 1 illustrates the scope of the respective frameworks on each level as well as the diversity of frameworks. The *technical level* enables monitoring and controlling of user behavior based on network-centric mechanisms. The choice of specific mechanisms to be implemented depends on the agreements at the contractual and legislative level. The *contractual level* predominantly enables agreements between the parties participating in a cooperative Wi-Fi-sharing network. Defecting users of a network that are identified by technical means can then be punished or excluded based on contractual agreements within the specific network. The *legislative level*, on the other hand, governs regulation in a network-independent way. Laws passed by the legislative power form the standardized basis for the relationship between participants of the Wi-Fi-sharing network and external parties. This hierarchy allows the top-down definition of rules and the subsequent control and punishment of participant behavior in cooperative Wi-Fi-sharing networks in a bottom-up fashion.

The remainder of this paper is structured as follows: In Section 2, we introduce the concepts of cooperative Wi-Fi-sharing networks and the affected parties in this specific network scenario. Section 3 gives an overview over technical agreements and presents cooperation strategies and actual implementations. Furthermore, we discuss quality metrics and measures supporting user incentives. We focus on the community scope and the effect of contracts and agreements within a network in Section 4. Section 5 discusses the role of legislation and shows how laws regulate user behavior on a nation-wide scope with respect to cooperation. Finally, we conclude this paper and present an outlook on cooperative Wi-Fi networks in Section 6.

2. WI-FI-SHARING NETWORKS AND COOPERATION

The availability of cheap wireless router hardware and a free wireless frequency band have created numerous user-driven initiatives to create cooperative networks as a cost-efficient alternative to provider-driven Wi-Fi networks. Examples for such Wi-Fi-sharing networks are the Freifunk [2] and Funkfeuer [3] initiatives in Berlin and Vienna, respectively, as well as the roofnet project in Cambridge [4, 5] and its spinoff Meraki [6]. The fundamental principle of these networks is cooperation by means of resource sharing. Participants of the Wi-Fi-sharing network provide network access to other participants as *micro operators* at their own AP while in turn receiving wireless access at other residential APs when they are mobile. At the technical level, Wi-Fi-sharing networks often build on existing standards such as the IEEE 802.11 infrastructure or ad-hoc mode. However, additional proprietary mechanisms may be added at the link layer and above according to the provider's or the community's design. Likewise, the terms and conditions of the network typically differ between networks depending on their respective provider model. As such, each Wi-Fi-sharing network establishes its own set of *internal standards* on the technical and contractual level that only apply to this very network. Lastly, the

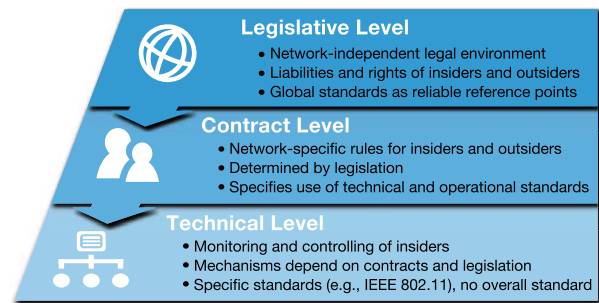


Figure 1. Frameworks to regulate user behavior and define network operation on different hierarchical levels.

different legislative frameworks in different countries hinder a global standardization of how Wi-Fi-sharing networks can be realized.

Due to the ongoing interest in these cooperative Wi-Fi-sharing networks, their network size has in many cases exceeded a level on which trust relations between every pair of users are possible. To handle such networks, a legal basis, in the form of a pico peering agreement [7], has been established to regulate the service provision by micro operators. This agreement states that users provide free data transit and open access while including a liability exclusion for the micro operator. However, while this agreement defines the legal aspects of service provision, corresponding technical measures to enforce proper user behavior and to detect misbehavior in such open networks must be provided.

Building on community concepts, companies such as FON [8] and Wippies [9] offer *commercial Wi-Fi-sharing products*. These companies offer customized IEEE 802.11 access points (APs) to community members in order for them to give access to their residential broadband connection. However, provider-driven Wi-Fi networks not only allow access for community members, but typically offer additional tariffs at which non-members can rent access to the network. These tariffs can either be time- or volume-based; the resource usage thus has to be measured by trustworthy, standardized tools to avoid disputes over fairness or fraud. Network providers thus become a stakeholder in the network and have to ensure that legal aspects and user contracts are fulfilled within the cooperative network.

Finally, a number of Wi-Fi-sharing networks exist as ongoing research-driven design concepts or architectural prototypes. Mobile ACcess [10] is an example for such a concept. It enables multiple parties such as private users, companies, universities, and municipalities to provide a unified cooperative network at company APs, on campus, or in public places. It has similar properties as commercial Wi-Fi-sharing networks, but does not depend on the presence of a single central network provider. In previous work, we introduced a general framework for securely providing such a network by means of PISA [11] and PISA-SA [12] and discussed challenges and applications for municipal Wi-Fi-sharing networks in [13]. In this paper, we focus on the as-

pects of cooperation within the scope of Wi-Fi-sharing networks and the parties affected by such networks.

2.1. Parties Affected by Wi-Fi-Sharing Networks

In cooperative Wi-Fi-sharing networks, users provide their broadband connections to the network and willingly contribute to the overall cooperative system. In addition, commercial networks typically involve a network provider managing and controlling the overall network. In the network, these *insiders* act out of their respective interests. However, as individual interests might interfere with the interests of the cooperative system, a balance between the strive for maximal personal gain and a fair use of the network needs to be ensured. As actions and repercussions of insiders occur in and affect the Wi-Fi-sharing network at hand, a regulation of insider actions is therefore achieved best on a per-network level. This then allows to enforce insider behavior by means of technical mechanisms within the network and membership contracts.

ISPs and other parties that are affected by but do not directly take part in the network can be considered as *outsiders* of a cooperative Wi-Fi-sharing network. ISPs only implicitly become stakeholders in the network as residential user Internet uplinks are provided by them. Hence, ISPs are providers of the backbone of a cooperative network, possibly even without being aware of this fact. Thus, explicit regulations are required between the ISP and the user as well as between the two economic entities, the ISP and an eventual provider of the Wi-Fi-sharing network. These regulations have to regulate insider behavior to not harm possibly unknowing outsiders and need to clarify the position of outsiders with regard to insider behavior.

3. TECHNICAL AGREEMENTS

Wi-Fi-sharing communities rely on cooperation on the network level. To balance the contribution and consumption of shared resources between community members, they must adhere to a common set of rules and conventions. If these conventions are not met, some members provide far more than they receive, others may be subject to malicious actions of others. As a first rule, users that expect other users to provide wireless Internet access need to open their own access point to other community members. Second, users should use the provided Internet access respectfully and within the bounds of the law because otherwise the access point owner may appear as the originator of the illegitimate action. Both of these rules are difficult to enforce because illegitimate user actions cannot be foreseen and may only prove illicit or selfish in hindsight.

Since no a-priori trust and no proof of legitimacy between users exist, an initial investment of trust is needed to bootstrap the system. However, users need to be able to check whether their trust in other users is justified. In this section, we discuss incentives and mechanisms to motivate, check

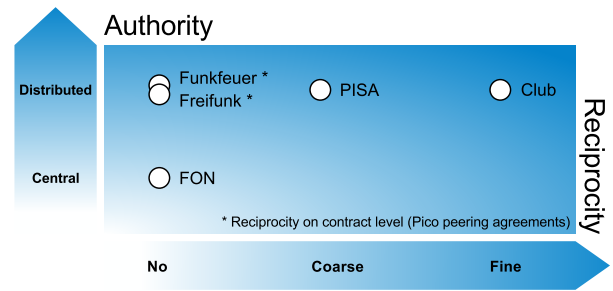


Figure 2. Use of reciprocity schemes and distribution of authority in different cooperative Wi-Fi-sharing systems.

and enforce user behavior with regard to both aspects mentioned above. We focus on the concepts of these mechanisms as their actual implementation is network-dependent.

3.1. Incentives to Contribute

The most powerful incentive for a user to contribute is a personal benefit that outweighs his invested resources. Hence, in a balanced system, each user experiences a personal gain or payoff from cooperating. The assumption that the other users will provide this gain is the basis for an initial cooperation. Consequently, if cooperation fails and the payoff decreases, this assumption proves unjustified and the trust in the cooperation of other users diminishes. In such a situation, the lacking of trust will cause discontinuation of the cooperation.

Reciprocity, the act of rewarding a positive action with another positive action, explicitly encourages cooperation and discourages defective behavior as cooperating users directly experience personal benefits. Fitzek et al. compare different cooperation strategies [14]. They show that reciprocal strategies prove more lucrative than non-reciprocal strategies in the long run, although pure defective behavior appears more profitable in the first place. Hence, in cooperative Wi-Fi-sharing networks, reciprocity schemes foster the trust in cooperation and the mutual interest in providing Wi-Fi resources and Internet uplinks.

On a technical level, initial trust can be rewarded through reciprocity in three ways: i) Through the implementation of a coarse-grained reciprocity scheme that allows for personal payoffs but lacks mechanisms to adapt to differences in resource provision or consumption, ii) by operating a fine-grained reciprocity scheme that allows for rewarding resource provision to other users with providing the same amount of this resource back to the user at foreign APs and iii) by incorporating multiple aspects of quality of the provided services as a metric to further differentiate between the contribution to the network made by different users. Figure 2 shows different actual networks and the implemented type of reciprocity scheme in this network. In the following, we discuss each of these technical concepts for achieving reciprocity as the basis for sustainable cooperation in a network.

3.1.1. Coarse-Grained Reciprocity

In a cooperative Wi-Fi-sharing network, the most basic form of participation is the provision of a user's Wi-Fi connectivity and Internet uplink to other users. This in turn is the requirement for the basic reward for this user, namely network access at APs provided by other users. Thus, rewarding the user with network access in a coarse-grained reciprocity scheme, requires him to provide network access at his own AP. However, single users do not directly interact with other users but rather perceive the cooperative Wi-Fi-sharing network as one single network [13]. Each user thus interacts with the whole network, in case a user does not share his AP or stops providing network access, access at all other APs, i.e. access to the whole network, will be denied.

A coarse-grained reciprocity scheme thus implements a basic access control mechanism in the Wi-Fi-sharing network. This mechanism requires a means of checking if users cooperate. The result of this check is then used to determine whether or not a user may gain access to the network. We discuss two different ways of implementing such a means: a) login-based access control with a central access control server as implemented by FON and b) decentralized certificate-based access control as proposed by PISA.

The FON network [8] requires the user to log in to a webpage once he wants to access the network. Upon login, the network checks whether the AP that is associated with this user is online and thus providing network access to other users. In case the user does no longer cooperate with the network, i.e. if the user's AP has been offline for a prolonged time, network access is denied. As the status of user cooperation is defined, checked and enforced by a central online authority, the mechanism used by FON is an example for a centralized scheme of ensuring reciprocity.

In PISA [11], no such online central entity exists. Rather than identifying a user by his username and password, PISA employs standardized cryptographic certificates, such as SPKI or X.509 certificates, to express network membership and to identify specific users. Once a mobile user requests access at a foreign AP, the AP checks the provided certificate for validity and only forwards any subsequent traffic if the certificate is valid. Certificates are periodically renewed, with the criteria for renewal being the ongoing cooperation by the user in providing his AP to the network. In PISA, AP availability is implicitly checked as all Internet traffic is routed through the mobile user's own AP. In case this redirection fails, the current AP may stop providing network access to the mobile user. Similar to FON, the certificate is revoked if the user stops cooperating. Although this approach requires short certificate lifetimes to account for user behavior, no central entity is needed for network access.

As shown in Figure 2, a coarse-grained reciprocity scheme can be implemented in a cooperative Wi-Fi-sharing network using these mechanism. While PISA also distributes the authority for user exclusion among all network entities, FON employs a single central entity. However, a user may provide only a fraction of his resources and still get high-quality net-

work access across the network. To account for the actual contribution to the network in measures such as traffic volume or AP uptime, a more fine-grained reciprocity scheme is required.

3.1.2. Fine-Grained Reciprocity

The problem of unfair resource consumption in a cooperative Wi-Fi-sharing network arises in different forms. First, a user might only make a fraction of the total bandwidth available to other users using traffic-shaping techniques at his AP. When using other users' APs, however, he might fully exploit the available bandwidth, thus creating a significant imbalance between the resources he offers and the resources he consumes. Second, a user whose AP is located in an unfrequented area might experience little resource consumption because few users use his AP. As this user might access and use the network in more frequented places, the balance of the system suffers.

A suitable parameter to measure and return a user's contribution in a Wi-Fi-sharing network would be the amount of network traffic that has been provided to others. In [15], Efstathiou et al. describe such a mechanism for community Wi-Fi-sharing networks. In this approach, users that generate traffic at a foreign AP issue a *receipt* for the time they use the AP and the amount of traffic they generate at the foreign host AP. These receipts are cryptographically secured and also denote the pair of users that exchanged resources. When using the APs of other users, the AP owner uses these receipts to receive access to the network as he can prove his cooperation and the amount of resources he offered.

This approach solves both of the above mentioned problems. First, users that offer only little bandwidth will in turn be able to request only little bandwidth in the network as the receipt clearly states the provided amount. Second, users that are not able to generate enough receipts may form groups, for example with friends that own well-frequented APs, and distribute the accumulated receipts of that group among its members. The *Club* network proposed by Efstathiou et al. implements this approach, as shown in Figure 2. However, this approach favors users that provide APs in well-frequented places as forming a group to collectively gather receipts may slow down the acceptance of the system in less well-frequented areas. Furthermore, to avoid disputes over the measured and provided amount of resources, standardized tools need to be used to carry out these measurements. Similar to the case in provider-driven networks, tools that are approved by an external organization or globally standardized support the trust of users in the operation of the measuring system and the overall network.

While this approach mainly considers the consumed traffic volume, a metric that combines multiple factors may give a more detailed measure of user contribution. We discuss such multi-dimensional schemes in the next section.

3.1.3. Fine-grained Quality-based Strategies

Single-dimensional metrics, such as the provided traffic volume, can neither express the combined *Quality of Service* (QoS) provided by a user nor the other users' *Quality of Experience* (QoE) when using the AP of other users. For example, a highly frequented AP may provide a large traffic volume even though the bandwidth at this AP is artificially limited or a transmissions cannot be received by clients due to packet collisions. Combining synthetic QoS parameters with user-provided QoE feedback, however, would allow for an overall evaluation of user cooperation.

To derive the QoS of a service, standardized methods [16] and frameworks [17] are available. The estimation of a specific QoS then includes synthetic parameters that can be measured at both the AP and the client, such as delay, bit rate, packet loss and jitter. Furthermore, the ITU Recommendation G.1000 [17] includes customer requirements, QoS offered by the provider and the achieved QoS as parameters. Following these recommendations, the AP and the client can derive a consensus value of the achieved or measured QoS and issue receipt-like structures as in [15]. Using these receipts to gain access at other APs, a consistent, network-wide QoS-based strategy can be used to recompense user contribution.

Building on QoS measurements as a technical basis, QoE measurements could augment this basis by user-provided feedback. These measurements should reflect perception, context and expectations of the user with regard to the services and system performance of the Wi-Fi-sharing network [18]. In a QoE-based scheme, the user is thus required to rate the current service in periodic time slots to establish a measure of the overall quality and usability of the network access provided by this AP. However, there are no standardized approaches for assessing QoE yet. Hence, providing a consistent metric based on user experience is difficult. QoE standardization activities in ITU-T Study Group 12 (SG12) are ongoing [19], nevertheless the scope of many of the current questions would need to be extended to establish multidimensional QoE assessments in Wi-Fi-sharing networks.

A significant problem assessing user cooperation through QoE ratings is the number of possible reasons for a bad user experience. For example, a slow downlink may be caused by traffic shaping at the AP, indicating defective behavior, or simply by the user being located far away from the access point. Making a distinction between these cases is not possible for the end-user. Hence, purely observing QoE as a metric for cooperation may be far-fetched and can prove error-prone. We are not aware of a cooperative Wi-Fi-sharing network that thoroughly incorporates QoE or QoS metrics in its reciprocity scheme. Standardized tools such as the ITU Recommendation G.1000 framework, could provide a basis for mechanisms that incorporate quality-based metrics for user contribution.

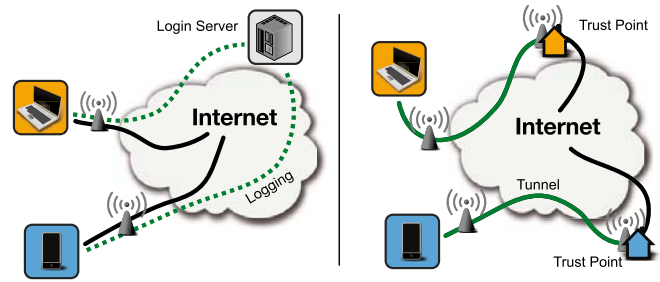


Figure 3. Centralized logging (left) as used in FON and decentralized tunnel approach (right) as introduced in PISA.

Next to user contribution, *user behavior* in the cooperative Wi-Fi-sharing network needs to be monitored and regulated. We discuss frameworks and approaches to mitigating or preventing malicious user behavior in the following section.

3.2. Incentives to Behave

Typically, users behave well in their own home networks since they use the Internet connection they pay and are liable for. In contrast, the use of Internet connectivity at community members' access points may tempt users to misuse these shared resources because of a missing perception of responsibility and liability [20]. Possible misuses range from overuse of the shared resources to committing Internet fraud. Hence, if malicious user behavior is expected, mechanisms must be established within the network to identify and eventually penalize misbehaving users, e.g., by revoking their access rights to the network.

Technical and legal actions against misbehaving users are only possible if their behavior can be observed and documented. Thus, *non-repudiation* is a basic requirement to successfully deal with misbehaving users. In this regard, network traffic generated within the cooperative Wi-Fi-sharing network needs to be clearly traceable and attributable to a specific device or user. In the following sections, we show two approaches of implementing the authority in the network to achieving this attribution, namely *centralized* authentication and logging structures and *decentralized* control mechanisms. We illustrate the use of these approaches in different Wi-Fi-sharing networks, as shown in Figure 2.

3.2.1. Centralized Approach

In a centralized approach, a single entity in the network, the network provider, takes over responsibility to control user behavior and to eventually penalize misbehaving users. We identify two options for the network provider to exercise its control at the technical level: *traffic-based* control and *service-based* control.

For traffic-based control, the network provider requires users to log in before using the network at another member's access point and centrally logs Internet traffic and transaction data generated by the user. The network provider maintains

a history of past user logins in combination with the stored traffic information. This logged information enables the network provider to identify the original user in case of misuse of network resources. A real-world example of a centralized log-based approach using a single login server is FON, as illustrated in Figure 3. While no actual traffic traverses the login server, it is contacted in each connection attempt by a mobile user. The general logging mechanism resembles the technique ISPs are forced to implement under the data retention act [21] in Germany and the EU. In addition, Wi-Fi-sharing networks that route all traffic from the cooperative network through a central backbone use logging techniques that equal the logging of a wired ISP.

A basic requirement of this log-based incident handling is that the network provider is a trustworthy entity because otherwise it cannot prove the responsibility of a user in a law suit. Hence, this approach is only possible if the logging is performed by an ISP-like trusted entity. Provider-less Wi-Fi-sharing networks without dedicated central entities (e.g., Freifunk and Funkfeuer) cannot use logging as the basic mechanism because it leaves users with their word against the word of other users in a law suit. Here, the question which user is more trustworthy (the user that logged the actions or the user that presumably acted inappropriately) is difficult to assess.

With service-based control, on the other hand, a network provider limits network access to a restricted set of services. While this does not enable the provider to identify the traffic originating user, it allows control over which content is accessible in the network. If the network provider ensures that no illegal information is accessible within the set of provided services, this renders traffic-based control mechanisms unnecessary. Restricted services are typically offered to unregistered clients at commercial hotspots as the main incentive to gain new users.

One important aspect common to both options is the requirement for the network provider to exercise control over the hardware and software used in the deployed user APs. Otherwise, malicious users can easily bypass the mechanisms used on the APs enabling enforcement of user behavior on both the traffic and service level.

3.2.2. Decentralized Approach

A fully decentralized cooperative Wi-Fi-sharing network does not provide a central point in the network where user control and penalties can be implemented. Likewise, mechanisms such as distributed logging of traffic information at individual APs do not suffice as a malicious user may provide false control information, i.e. traffic logs, about other users. Instead, decentralized networks need to mitigate the incentive for a user to misbehave at an architectural level.

In PISA [11], as an example for a decentralized cooperative Wi-Fi-sharing network, we remove the motivation for a user to misbehave at another member's AP. We achieve this by redirecting all Internet traffic over a secure tunnel to the home network of the mobile user. The home network then forwards

the traffic to the original destination in the Internet. Figure 3 illustrates the flow of traffic from the mobile user to the Internet. This effectively makes the user's home network appear as the egress point of all traffic generated by the user – a situation comparable to when the user accesses the Internet from within his home network. Hence, this approach encourages benign user behavior by ensuring direct legal liability for the traffic the user generates at other members' APs. Furthermore, only legitimate devices can connect to the user's home network by pairing user devices with the user's home network in an initial step and by performing subsequent mutual authentication between the end-points during the establishment of the secure tunnels. This prevents outsiders and malicious insiders to exploit the secure tunnel mechanism.

While the actual use of the network occurs on the technical level, a network-specific framework for this use is necessary. This framework is based on standards such as IEEE 802.11 for communication and SSL for security and regulates the relations and rights of participants in the network. Furthermore, the general operation of this network with respect to the economic interests of insiders and outsiders needs to be defined. We discuss the contractual means of regulating both in-network operation as well as economic regulation of Wi-Fi-sharing networks in the next section.

4. COOPERATION BY CONTRACT

Technical agreements and implementations directly establish and regulate the use and user behavior inside of the network. However, this set of fine-grained rules needs to be embedded in a contractual framework that defines the more general network aspects. Examples for such higher-level rules are the definition of security standards that are required for communicating in the network or agreements about the rights and duties of users. Rules that are defined in contracts may thereby i) define a framework for benevolent actions in a cooperation, ii) define a framework for network-wide consequences of malicious actions and iii) define the obligations of a network user and his role in the network. As a best practice, these rules should rely on technical standards and software wherever possible to ensure an easy acceptance and a lasting contribution. Above all else, contracts serve as a fixed point of reference which allows for an assessment of actions in the network and defines eventual consequences.

There are numerous examples for contracts in cooperative networks. *Pico peering agreements* [7] as used in Wi-Fi-sharing communities such as FON [8], e.g., regulate which services cooperating users have to provide as participants of the network. Such a contractual definition is necessary to check if access point owners are behaving correctly. The sustained operation of Wi-Fi-sharing networks directly depends on such checks to prevent misuse of the network and defection of benign users that put trust into other participants when they joined. Also, these checks form the basis for exclusion of defecting AP owners.

Next to network regulation, contracts also form a basis of cooperation between the network provider and ISPs to satisfy the economic interests of both parties. First, a contract may define the monetary arrangements between the network provider and the ISP and defines a set of rules that allows the ISP to sue in case of illegal actions of the network provider or users. Second, payments and subscription fees for users of the network are determined by contracts prior to network operation. Certain conditions may thereby require technical mechanisms to be realized. For example, network access for a user who bought a day pass for a commercial Wi-Fi-sharing network needs to be technically revoked after the day pass has expired.

Contracts thus serve as a single-network framework regulating user behavior and economic interaction. On a broader scope, a similar framework is needed to govern the fundamental rules and regulations concerning the establishment, operation and limits of Wi-Fi-sharing networks. As this exceeds the scope of a specific network but rather applies to every such network, applicable laws that provide a legal framework for Wi-Fi-sharing networks are required.

5. THE IMPACT OF LEGISLATION ON COOPERATIVE NETWORKS

The legislative power defines the legal framework that provides the operational context for Wi-Fi-sharing networks in general at the national level. To do so both on a legal and contractual level but also with regard to approved technical means, laws in this context depend on standards as a basis and as orientation points. As these laws provide regulation in a general fashion independent of any specific network, they regulate both the initial establishment and design of networks as well as the legal rules of behavior for insiders and outsiders. In case actions by insiders or outsiders violate these laws, enforcement of the appropriate consequences or exclusion from the network can be achieved through judicial power. Legislation may thus provide the basic rules of conduct for cooperation in two different ways: first by definition of the legal environment and second by protection of interests of insiders and outsiders.

One example for a definition of the environment in which cooperation takes place are laws that define the liabilities of the involved parties. In this sense, the regional court of Mannheim, as a representative of the judicial branch in Germany, decided in 2006 [22] that owners of Wi-Fi access points are partially liable for any Internet traffic generated from within their local network. The decision was based on existing legislation, where the defendant has to prove his innocence (§138 Abs. 2 ZPO), and was recently confirmed by a verdict at the German Federal Court in 2010 [23]. Hence, in Wi-Fi-sharing, the AP owner is responsible for any traffic generated by other users at the shared AP. Such regulations have a significant impact on the design and operation of a Wi-Fi-sharing network and thus on the incentives for a user to join and participate. While such decisions and laws in general do not directly set or mandate standards with regard

to Wi-Fi-sharing, they refer to current standards and mandate the use, e.g., of well-established security standards to protect the local wireless network [23].

The legislation may also protect the business interests of insiders and outsiders, an example for which is the decision made at the regional court of Cologne in 2009 [24]. Here the verdict explicitly forbids providers of a Wi-Fi-sharing community to rent out Wi-Fi-based network access at private homes to (non-)community members, as these use the Internet uplinks of third-party ISPs. The decision is based on recent competition regulations in Germany. Regulations such as this one impose restrictions on how cooperative network operation may affect parties outside of the network.

However, legislative mechanisms typically have a very broad scope and do not provide rules for specific cooperative network scenarios such as for distinct Wi-Fi-sharing networks. If rules are required on a more fine-grained network- or location-specific level, contracts and technical measures are used based on individual terms and conditions for a network.

6. CONCLUSION

In this paper, we analyzed cooperative Wi-Fi-sharing networks as a prime example of cooperation on different levels. Cooperation happens on the technical level inside of a given network, the contractual level that defines the general network operation and the legislative level on which the general rules for Wi-Fi-sharing networks are established. While cooperation benefits all participating parties, selfish and defective behavior needs to be accounted for. Based on this notion, we discussed incentives and frameworks to motivate and regulate user behavior on each level with the goal of providing the maximum benefit for each user while achieving a sustainable network operation.

At the technical level, we discussed possibilities of supporting cooperative behaviors of users and frameworks that allow for checking user behavior and eventually penalizing malicious user actions. The contractual level establishes the general rule of operation in which implementations on the technical level need to be realized. Furthermore, the economic interests of parties acting inside of the network and parties outside of it, such as ISPs, are regulated on this level. On the legislative level, the economic interests of outsiders in competition are regulated and fundamental laws regarding the legal liability of insiders and outsiders are given. Traversing these three levels from the bottom up, the scope of rules widens from a per-network scope to a country- or continent-wide scope, such as in Germany and the European Union, respectively.

We expect cooperative Wi-Fi-sharing networks to gather continuous interest as a cost-efficient way to provide Internet access to mobile users. However, we assume their scope to be local, i.e. city-wide, with high bandwidth connections and specialized services and a cooperation between geographically close users. This is because techniques such as UMTS provide highly mobile, general purpose Internet

access on a national scope. However, for cooperative Wi-Fi-sharing networks to establish a sustainable local operation, strong incentives for contributing to the network and for using the network in a benign fashion need to be provided. We discussed different reciprocity schemes for creating such incentives through coarse-grained and fine-grained mechanisms that closely couple users' benefits to their contribution in the network. Furthermore, we assume a stable legal foundation for the establishment and operation of cooperative Wi-Fi-sharing networks to be necessary to attract network providers, ISPs and private persons on the basis of clear regulations with regard to legal liabilities and economic interests. While we expect no standards specific to Wi-Fi-sharing due to the different stakeholders and the complexity of single networks, ongoing interest of ISPs and commercial providers could strengthen the case for standardized partial solutions or usage frameworks.

REFERENCES

- [1] M.A. Nowak, "Five rules for the evolution of cooperation," *Science*, 2006.
- [2] Freifunk Community, "Freifunk Website," [Online] Available <http://start.freifunk.net/>, last visited July 12th 2011.
- [3] Funkfeuer Community, "Funkfeuer Free Net Website," [Online] Available at <http://www.funkfeuer.at>, last visited July 12th 2011.
- [4] MIT Roofnet Project, "MIT Roofnet Website," [Online] Available at <http://pdos.csail.mit.edu/roofnet/>, last visited July 12th 2011.
- [5] J. Bicket, D. Aguayo, S. Biswas, and R. Morris, "Architecture and evaluation of an unplanned 802.11b mesh network," in *Proceedings of the 11th annual international conference on Mobile computing and networking (MobiCom)*, 2005.
- [6] Meraki Inc., "Meraki Website," [Online] Available at <http://www.meraki.com>, last visited July 12th 2011.
- [7] Funkfeuer Community, "Pico Peering Agreement," [Online] Available at <http://www.funkfeuer.at/PicoPeeringAgreement.59.0.html>, last visited July 12th 2011.
- [8] FON WIRELESS, Ltd, "FON Website," [Online] Available at <http://www.fon.com/>, last visited July 12th 2011.
- [9] Saunalahti Group Oyj, "Wippies Website," [Online] Available at <http://www.wippies.com>, last visited July 12th 2011.
- [10] Mobile ACcess Project, "Mobile ACcess Project Website," [Online] Available at <http://www.mobile-access.org/>, last visited July 12th 2011.
- [11] T. Heer, S. Götz, E. Weingärtner, and K. Wehrle, "Secure Wi-Fi Sharing on Global Scales," in *Proceedings of 15th International Conference on Telecommunications (ICT)*, 2008.
- [12] T. Heer, T. Jansen, R. Hummen, S. Götz, H. Wirtz, E. Weingärtner, and K. Wehrle, "PiSA-SA: Municipal Wi-Fi Based on Wi-Fi Sharing," in *Proceedings of 19th International Conference on Computer Communications and Networks (ICCCN)*, 2010.
- [13] T. Heer, R. Hummen, N. Viol, H. Wirtz, S. Götz, and K. Wehrle, "Collaborative Municipal Wi-Fi Networks - Challenges and Opportunities," in *Proceedings of IEEE PerCom Workshops (PWN)*, 2010.
- [14] F.H.P. Fitzek and M.D. Katz, *Cooperation in wireless networks: principles and applications; real egoistic behavior is to cooperate!*, Springer Verlag, 2006.
- [15] E.C. Efstathiou, P.A. Frangoudis, and G.C. Polyzos, "Controlled Wi-Fi Sharing in Cities: A Decentralized Approach Relying on Indirect Reciprocity," *IEEE Transactions on Mobile Computing*, 2010.
- [16] "Framework and methodologies for the determination and application of qos parameters," ITU-T Recommendation E.802, feb 2007.
- [17] "Communications quality of service: A framework and definitions," ITU-T Recommendation G.1000, nov 2001.
- [18] "New definitions for inclusion in recommendation itu-t p.10/g.100," Recommendation ITU-T P.10/G.100 (2006) - Amendment 2, jul 2008.
- [19] Alexander Raake and Sebastian Möller, "Recent multimedia qoe standardization activities in itu-t sg12," IEEE COMSOC MMTC E-letter - Special Issue on Quality of Experience issues in Media Delivery, aug 2011.
- [20] Daithí Mac Síthigh, "Law in the last mile: Sharing internet access through wifi," SCRIPTed 355, march 2009.
- [21] Deutscher Bundestag, "Entwurf eines Gesetzes zur Neuregelung der Telekommunikationsüberwachung und anderer verdeckter Ermittlungsmaßnahmen sowie zur Umsetzung der Richtlinie 2006/24/EG," [Online] Available at <http://dip.bundestag.de/btd/16/058/1605846.pdf>, last visited July 12th 2011.
- [22] Regional Court Mannheim, "7 O 76/06," 2006.
- [23] German Federal Court, "I ZR 121/08," 2010.
- [24] Regional Court Cologne, "6 U 223/08," 2009.

MAKING THINGS SOCIALIZE IN THE INTERNET – DOES IT HELP OUR LIVES?

Luigi Atzori, Antonio Iera, and Giacomo Morabito

University of Cagliari, Italy, l.atzori@diee.unica.it
University of Reggio Calabria, Italy, antonio.iera@unirc.it
University of Catania, Italy, giacomo.morabito@diit.unict.it

ABSTRACT

Current communication and computation technologies make it possible to embed intelligence and communication capabilities in most of the things surrounding us; this leading to the Internet of Things (IoT) concept. To really exploit the potential of the IoT, objects and provided services should be easily discoverable and usable by humans and by other objects. Besides, trustworthiness of the billions of members of the IoT should be a key element in service selection. Existing solutions for service discovery in IoT do not scale with the number of nodes that is expected to be order of magnitude larger than in the current Internet. In this paper we propose to build a social network, that we name the Social Internet of Things (SIoT), that can be used to provide a navigable structure to the IoT. We also provide a framework that can be applied to socially tie things together and a preliminary architecture to be used as a baseline for the implementation of the SIoT. Our work demonstrates that standards should support establishment and management of federations of objects (ruled by social relationships) that represent “communities” of things in the SIoT.

Keywords— Ubiquitous computing, Internet of Things, Social Networks.

1. INTRODUCTION

Current communication and computation technologies make it possible to embed intelligence and communication capabilities in most of the things surrounding us. This has led to the *Internet of Things* (IoT) concept. The IoT is a world-wide network of interconnected objects uniquely addressable, based on standard communication protocols [1]. As such, the IoT can be seen as a *smart environment* paradigm and has the potential to radically change the way we interact with the environment and other people.

The IoT vision can be fully achieved only if objects are able to cooperate in an open way. Unfortunately, current implementations enable the cooperation among objects only if belonging to the same closed group. Sort of gateways are needed to allow specific groups of objects (aware of each others) to communicate and cooperate.

Necessary condition to reach the desired open cooperation

is the ability of nodes to discover the available services they need. Solutions for *service discovery* can be classified as [2]

- **Centralized:** A server is deployed in the network where a *publish/subscribe* scheme is utilized. Obviously, such solution does not scale as the number of nodes in the IoT will be huge.
- **Request broadcast-based:** Service requests are broadcast throughout the network. Nodes containing the requested services will reply.

Obviously such an approach is reliable, however, it is not scalable as it involves global broadcast of the requests which is highly inefficient when the network size increases. Furthermore, it involves processing even in nodes that do not have the required service, which causes a reduction in the overall system efficiency.

- **Advertisement-based:** Nodes that want to share a service broadcast information about it. Nodes that in the future might be interested in such a service store such *publish* messages in a cache. This approach has two problems. First: the size of the cache memory is expected to increase proportionally to the number of nodes in the IoT. Second: implementation of functionality that locate objects providing the required service is needed. Reliability of such a task becomes difficult to achieve if we introduce mobility into the picture.

In this paper we propose a paradigm for service discovery which makes use of social relationships between objects. We call this the *Social Internet of Things* (SIoT). More specifically, in the SIoT objects establish social relationships with each others. Objectives of such relationships are twofold:

- Give the IoT a structure that can be shaped as required to guarantee network navigability (see [3], for example) so that service discovery can be performed effectively while guaranteeing scalability.
- Create a level of trustworthiness which could be used to leverage the level of interaction between things that are friends (or friends of the friends, etc.).

The proposed approach is distributed and therefore is expected to guarantee higher scalability and better reaction to the frequent state changes.

Proliferation of embedded computing and communication devices in most objects surrounding humans is a necessary step towards the vision of a *fully networked human*. However, the risk exists that the number of such devices will soon become too large and scalability problems will emerge consequently. In this context, efficient cooperation between smart objects creating trusted, dynamic social-like communities contribute to solve the above scalability issues.

Furthermore, autonomous (without actions performed by human user) organization of objects in communities, cooperating towards the implementation of added value services, enables a paradigm shift to the vision of the fully networked human vision; it, in fact, supports the connection of human users with resources and services rather than nodes.

Observe that exploitation of social networks in the context of the IoT has been investigated in [4]. There, it was proposed to exploit (human) social network relationships to share the resources offered by a smart thing. More specifically, smart things support web services that can be used by friends of their owner. Online social networks can be used to authenticate and authorize *friends* of the thing's owner. In this context, also the idea of the possibility of associating social potentialities to smart objects, towards shaping a Social Internet of Things emerges from [5]. Notwithstanding, the authors' attention is more focused on envisaging a generic IoT architecture by integrating both RFID and smart object-based infrastructures than on defining social relationships among objects on which to base objects' interactions of a real social networks of smart objects.

Note that the approach we have in mind is different in two major ways:

- We are interested in establishing and then exploiting social relationships between things, not between their owners.
- We use social relationships so that things can crawl the IoT and discover services and resources.

The reasons and relative benefits of the foreseen SIoT are listed in Table 1 and compared with those that are recognized as the most important advantages of the networks between humans.

With the institution of the *Internet of Things - Global Standards Initiative*¹ (IoT-GSI), ITU plays a crucial role in the standardization process of solutions for the IoT. Our contribution demonstrates that standards for the IoT should support the establishment and management of federations of objects (related by social relationships) that can interact in a more strict way and rely on each others for the execution of complex tasks. This, however, should not prevent the establishment of *loose* interactions between objects not related by social relationships.

¹See <http://www.itu.int/en/ITU-T/techwatch/Pages/internetofthings.aspx>.

Finally, we believe that standard guidelines for the interactions between different networks of social smart objects should be provided at the earliest steps of the deployment phase. In fact, in the context of online social networks the standardization process has begun late (compared to the deployment phase). Accordingly, today large part of the most popular social networks websites do not comply with standard-like guidelines such as those defined by the *OpenSocial*² initiative.

2. MAKING THE IOT SOCIAL

Basic idea of this work is the definition of a “social network of intelligent objects” – which we name the *Social Internet of Things* (SIoT) – in analogy with social networks of human beings, where the value of social relationships has been conceptualized in the so called *social capital* [6]. Bringing such a concept in the IoT would allow to successfully extend the use of models designed to study social networks [3] also to deal with IoT related issues (related to extensive networks of interconnected intelligent objects). The first issue to address is the definition of a kind of social behavior among objects. This may derive from observing typical information exchanges and possible interactions among smart objects, which are called to implement applications and services for the IoT.

Within the overall architecture we envision, the following tasks shall be fulfilled: (i) define a sort of notion of social relationship among objects, (ii) define the “degrees of social relationship” that can be established, (iii) study its evolution over time in a perspective of constant evolution and updating of the “Social Internet of Things”, and (iv) investigate how this social relationship can be codified and supported by current technologies.

We address social relationships among objects in the IoT and the *degree* and dynamics of such relationships in Sections 2.1, 2.2, and 2.3, respectively. A preliminary study of the implementation of the above concepts through current technologies is provided in Section 3.

2.1. Social relationships between things

In defining the types of social relationships between objects we must consider that sociological studies have demonstrated that most value from social relationships can be gained when the structure of the social network is characterized by highly connected clusters which partially overlap with each others [7, 8]. Accordingly, as for human being, we first consider a “parental” form of socialization. In SIoT, what we define “Parental object relationship” is correlated to the membership of a set of objects to the same production batch and is established only among objects usually with the same nature and originated in the same period by the same manufacturer. Moreover, like humans do, objects can establish social

²<http://www.opensocial.org>

Table 1. Reasons for which humans use and things may want to use social networks.

Reason for Humans	Reason for Things
Become visible/increase popularity	Publish information/services
Find resources/find old friends	Find information/services
Obtain context information and get filtered information	Get environment characteristics
Discover new resources and find new friends	Find new services/updated information

relationships whenever they come into contact to share personal (e.g. cohabitation) or public (e.g. work) experiences, named “co-location object relationship” and “co-work object relationship”, respectively. These relations are determined whenever objects (homogeneous or heterogeneous) are either used always in the same place or collaborate to provide a common IoT application. An example of co-location object relationship is the case of different objects (sensors, actuators, etc.) used in the same environment to implement either home or industrial automation applications. Examples of co-work object relationship, by contrast, involve objects that do not have constant co-location relationship but are used together and cooperate for applications such as emergency response (sensors of body area networks, environmental sensors, etc.) and telemedicine.

A further type of relationship among objects is a consequence of their belonging to the same user (e.g., mobile phones, music players, game consoles) and the resulting high probability of interaction and data exchange with each other. We name this “ownership object relationship”.

The last type of relationship is established when objects come into contact, sporadically or continuously, for reasons purely related to relations of *friendship* among their owners, which are in touch during their lives (e.g., devices and sensors belonging to friends who attend each other, classmates, travel companions, colleagues). We name this “social object relationship”.

2.2. Degrees of social relationship

A classification of the “degree” and the “structure” of social relationships among objects is a prerequisite for the definition of adequate models of interaction, based on the nature of their relationships. This is also the basis for defining the type of information exchange among objects belonging to each structure. Again, we can draw inspiration from typical studies in the fields of Sociology, Anthropology, Social Psychology, or Cognition. Several activities in these fields start from a widely accepted classification of social relations proposed by Alan Fiske in his “relational models” theory [9] and studied, among others, by Nick Haslam in [10]. From these theories, four basic relational frames or structures derive from the four elementary models of the Fiske’s theory. In *Communal sharing* relationships, equivalence and collectivity membership emerge against any form of individual distinctiveness. *Equality matching* is based on egalitarian relationships characterized by in-kind reciprocity and balanced

exchange. *Authority ranking* relationships are asymmetrical, based on precedence, hierarchy, status, command, and deference. *Market pricing* relationships, finally, are based on proportionality, with interactions organized with reference to a common scale of ratio values.

Next step is to relate these patterns of interaction among human beings to possible relational modes of smart IoT objects. Communal sharing can be definitely associated with behaviors of objects, not relevant individually but with only a collective relevance. For example, this type of relation is associated to objects forming a “swarm”, according to which it is not important the service offered by the single object but the service that the entire swarm can provide to users. Equality matching may represent all forms of information exchange among objects that operate as equals in the perspective of providing IoT services to users while maintaining their individuality. While with objects in communal sharing relationship the service is associated to the whole group, in the second case every object has associated a service that it advertises. Authority ranking is a type of relationship established among objects of different complexity and hierarchical levels (such as, RFID reader and Tags, master and slave terminals in Bluetooth) exchanging information in a highly asymmetric fashion. In this case, the service advertised is usually associated to the whole group of objects (e.g., the whole coalition composed of master and slaves) or to the object of highest rank, which then coordinates those of lesser rank to provide the service. The last type of relationship, Market pricing, can be associated to objects have to work together in the view of achieving mutual benefit. In many IoT applications, this implies that the participation in this relationship is considered only when it is worth the while to do so. Table 2 relates different types of object relationships, relational models and object interactions between each other. Also exemplary families of applications belonging to each category are shown.

2.3. How Things socialize?

At this point, it is necessary to understand if occasions actually exist to establish the defined relationships among objects and when the established relationship likely changes.

A “parental object relationship” is easy to implement, because such a tight link can be created, for example, among objects belonging to the same production batch, directly during the item production. Surely this inter-object link will not change over time and is only updated by events related to the disruption/obsolescence of a given device. Suitable proce-

Table 2. Object relationships, relational models, and interactions

Category of “object relationship”	Applicable relational model	Type of object interaction	Application examples
Parental object relationship	Communal sharing Equality matching Market pricing	Swarm Balanced Cooperative	Best practice sharing
Co-location object relationship	Communal sharing Equality matching Authority ranking	Swarm Balanced Unbalanced	Environmental monitoring Building automation Industrial automation Data fusion Automatic identification of goods in storing area
Co-work object relationship	Communal sharing Equality matching Authority ranking	Swarm Balanced Unbalanced	Emergency and first responder deployments Data distribution Telemedicine Military applications Logistics
Ownership object relationship	Equality matching Authority ranking	Balanced Unbalanced	Remote control of devices Personal data storing and distribution Multimedia content fruition Infomobility and positioning
Social object relationship	Equality matching Authority ranking Market pricing	Balanced Unbalanced Cooperative	Personal data exchange Cooperative sharing and downloading Distribute gaming Cooperative and hybrid positioning

dures of relationship refresh, based on the periodical check of the existence and the functioning of a given friend object are required.

Also “co-location” and “co-work object” relationships are easily implementable, as the establishment of the social relationship among objects become part of the initialization/implementation of either a “location based application” profile or a “situation-based application” profile. Changes in this kind of relationships are more frequent. Surely, it may be dynamically updated based on the evaluation of parameters such as: time duration of either the co-location or the co-working, frequency of the interaction events, object reputation gained during last interactions, etc. All these parameters shall be monitored through suitable policies whose decision are based on flexible and rich object profile descriptions, and on digital reputation management policies. Maybe, the way to establish an “ownership object relationship” is the most natural to envisage. Associating one another all devices owned by the same user is in fact a common procedure performed by anybody to allow them to exchange data. A ownership object relationship is the logical generalization of this concept through a more complex device profile. Variations happen following either the natural obsolescence of owned objects or any change in the ownership.

The implementation of a “social object relationship”, may for example naturally follow the social interaction of human beings. Similarly to people exchanging their contacts (phone numbers, e-mail addresses, etc.), it is easy to implement ad-

hoc procedures for the exchange of profiles of the devices they own. Actually, also the case in which objects episodically come into proximity may fall into this category of object relationship. In this latter case objects are transported by their owner in a given area and discover nearby devices with a profile of any interest: functional similarity, complementarities, same trademark, etc. The device, if properly authorized, may decide to establish a relationship with other objects, even transparently to the user, by exchanging the social profile. The driving idea is that a device with similar functional behavior may become a best practice to follow to solve a problem that could raise. The duration of these social relationships is ruled by policies exploiting ad-hoc defined metrics to measure the opportunity of maintaining a given link.

3. AN ARCHITECTURE FOR THE SOCIAL INTERNET OF THINGS

In Section 3.1 we first identify the components that can be distinguished in current *Social Network Services* (SNSs) used by humans. Then, we analyze how such components should change to implement the SIoT in Section 3.2. Finally, in Section 3.3 we provide an overview of a preliminary architecture implementing such components.

3.1. Components of SNSs

The definition of an architecture for SIoT should start from the analysis of the solutions currently adopted by the SNSs used by humans.

Unfortunately, we found that there is not a common reference architecture. Indeed, the number of the components and relevant functions may differ significantly from one implementation to another. A partial analysis in this direction is provided in [11] and [12]. On the left hand side of Figure 1, we provide a sketch of the resulting logical architecture.

Central part of the architecture is the *Profiling* of the member, who is asked at the beginning to describe himself/herself and will be adding further information and updates about his/her personalities during all the virtual/digital life.

Then, the uploaded profiles must be visible within the system (and externally) by a *Social graph* module that publicly or semi-publicly displays the connections between the member and their “friends”.

Strictly linked to this module is the *Social presence* one that provides the users with the functionality to traverse the connections (e.g., to view profiles associated with the list of “friends”).

Participation tools are then used to allow the users to keep in touch with the other members, such as: e-mail, instant messaging, chat rooms, blogs, message boards, telephony, videoconferencing, and others.

Each member is usually provided with tools for controlling his/her own visibility (search, profile viewing) and how he/she prefers to interact or be contacted by other entities. This is the *Relation control* tool.

Another important module is used in the current SNSs and is gaining more and more importance. It is the *Service API* component, which represents the interface that allow either third-party services to be included in the SNS (so that the user can benefit of additional services) or external sites to incorporate content into their services provided by the SNSs (e.g., Facebook and OpenSocial).

All the activities carried out through these modules are stored in a layer of metadata where the use of ontology and semantic web has become mandatory [13], with major initiatives already carried out, such as: the Friend-of-a-Friend (FOAF; www.foaf-project.org) project and the Simple Knowledge Organization Systems (SKOS) model.

On top of the components for SNSs shown in Figure 1, applications are developed, spanning from the creation of events to the online gaming, from the management of virtual farms to the collaborative creation of multimedia content (music, movies, logos).

3.2. Components of SIoT

We envision a system for making things socialize on the Internet which takes some major components of existing SNSs for humans as described in the previous subsection. However, some major differences must be introduced which are

mainly due to the limited computing capabilities of smart things and the different objectives of SIoT. On the right hand side of Figure 1 we show the SIoT main components. Underlined bold fonts have been used to highlight the differences with respect to SNSs.

At the center of the system three basic components can be envisioned:

- ***ID management***: assigning an ID that universally identifies all the possible categories of objects in the real world is not an easy task. To maintain the current object identification schemes, we foresee the deployment of a system where existing mechanisms can inter-operate. This can be made possible by adopting a simple XML-based protocol that allows for specifying the ID mechanism adopted other than the ID itself. This system should include at least: IPv6 addresses, Universal Product Code (UPC), Electronic Product Code (EPC), Ubiquitous code (Ucode), OpenID, URI.
- ***Object profiling***: it includes static as well as dynamic information about the object. Objects should be organized in classes, where each class is defined on the basis of the main object features. In this context, appropriate classification strategies should be identified. A viable option could be to distinguish objects based on their computing and communication capabilities. Alternatives could be based on either the type of services they offer or the type of interfaces they implement.
- ***Owner control***: the owner has to be able to define activities that can be performed by the object, the information that can be shared (and which other object is involved in the sharing), as well as the type of relationships to set up. Accordingly, specific policies need to be defined for every possible operation that can be performed by/on each member. To this purpose, different security and access control policy definition languages are already available and can be used [14]. Owner control includes the functionality of the Relation control component in the SNSs.

As to the other satellite components, in Figure 1 we report the most important³:

- ***Service discovery***: this component replaces the Social presence. Service discovery is a fundamental component which is aimed at finding which objects can provide the required service in the same way humans seek for friendships and information in the SNSs.
- ***Relationship management***: this is fundamental in the network of objects since these have not the intelligence of humans in selecting the friendships so that this intelligence needs to be incorporated in the SIoT. Objective

³Note that *Social graph* is not a major component of the SIoT functionality. Indeed, a social graph tool may still be implemented to allow humans to visualize their (and not only) objects relationships; this is, however, a minor functionality.

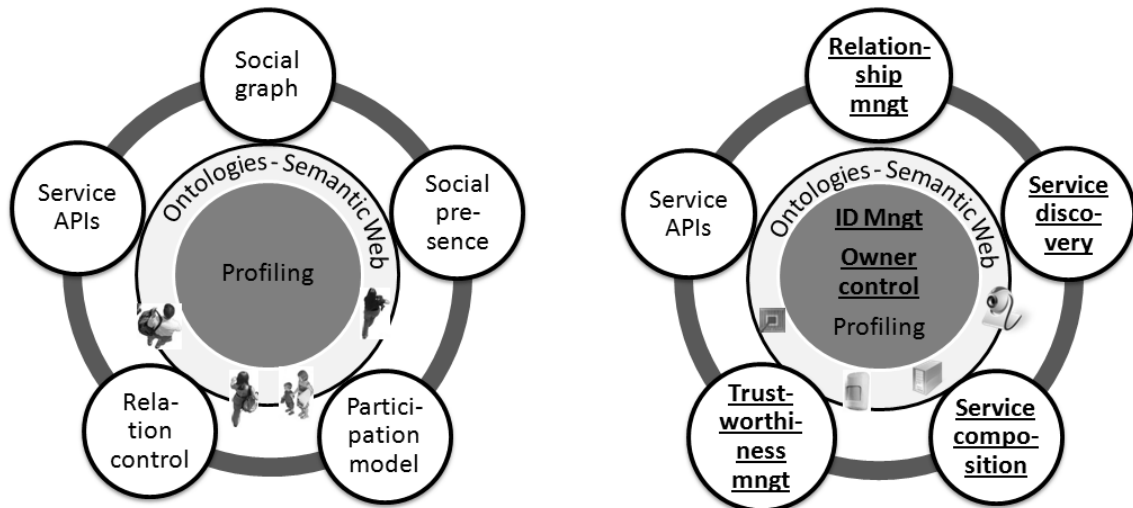


Figure 1. Basic components of social networks platforms for humans (on the left) and for objects (on the right)

of this component is to allow objects to start, update, and terminate relationships with other objects. The selection of which friendship to accept is based on the human control settings previously described. It is then driven by a set of rules that are defined on the basis of the way the objects get into contact with new objects in the physical and virtual worlds.

- **Service composition:** this component enables interaction between objects and replaces the Participation model. The interaction most of the times is related to an object that wishes either to retrieve information about the real world or to find a specific service provided by another object. Indeed, the main potentialities we see in deploying SIoT is its capability to foster the retrieval of information about the real world and services provided by other objects. Leveraging on the object relationships, the Service discovery provides with the way to find the desired service, which is then activated by means of this component. In general, when the requested service corresponds to the provisioning of information about the physical world, the service composition process can be performed according to either a *reactive* or a *proactive* approach. In the *reactive* approach, one of the applications developed on top of the SIoT triggers the request for an object providing a specific data. The service discovery components drive the discovery of the potential sources and then a reactive composition is performed to get the requested information. This is done by making objects interact by means of the available technologies for service composition (mainly according to either a RESTful or a SOA approach).

In the *proactive* approach the source of information about the real world can directly expose the generated data (or metadata) on its own social network showcase so that every other member (or a subset according to the authorization policies) can directly acquire the in-

formation when needed.

This component will also include the functionality of crowd information processing. This is aimed at processing the information obtained from different objects so as to obtain the most reliable answer to the information query on the basis of the different visions, similarly to what has been proposed in [15].

- **Trustworthiness management:** this is aimed at understanding how the information provided by the other members has to be processed. Reliability is built on the basis of the behavior of the object and is then strictly related to the relationship management module. Trustworthiness can be estimated by using notions well known in the literature such as *centrality* and *prestige*, which are crucial in the study of social networks [16].
- **Service APIs:** This component is analogous to the one required in SNSs.

To support the deployment of the model, a specific ontology is needed to record and represent the objects profiles, their friendships, as well as the relevant relationships. This has to be designed taking into account the objective of managing the relationships but also considering that the same ontology is used in the other components, especially for service discovery and trustworthiness management.

3.3. SIoT architecture

In this section we briefly introduce a possible preliminary architecture which is aimed to implement the SIoT components described in Section 3.2. We propose a system made of three main layers at the server side, as shown on the right hand side of Figure 2.

The *Base layer* encompasses the database for the storage and management of the data and relevant descriptors, the ontologies database, the semantic engines and the communications.

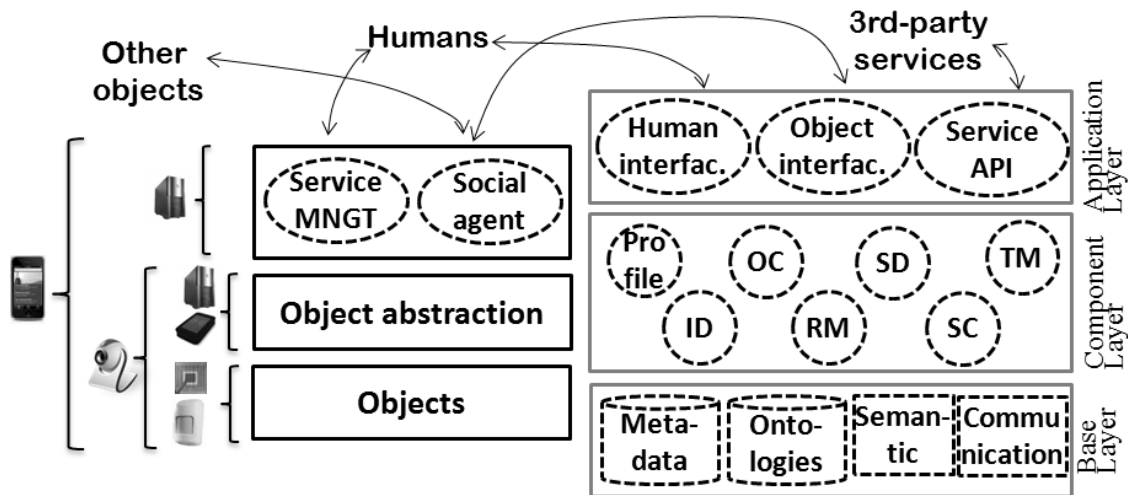


Figure 2. Architecture for the SIoT: client side (left) and server side (right). Acronyms stand for: OW - owner control, RM - relationship management, SD - service discovery, SC - service composition, TM - trustworthiness management

The *Component layer* includes the tools for the implementation of the functionalities described in the previous subsection: profiling, ID management, owner control, relationship management, service discovery, service composition, and trustworthiness management.

The *Application layer* is where the interfaces to the objects, the humans and the services developed by third-party are located.

This high-level architecture sketch may be mapped in a single site, deployed in a federated way by different sites or deployed in a cloud.

At the object side (see left side of Figure 2), the first layer of the architecture – named the *Object layer* – is where the physical objects are located and are reached through their specific communication interfaces.

Due to the fact that a large and heterogeneous set of objects can take part of the network, each one providing specific functions accessible through its own dialect, an *Object abstraction layer* is needed to harmonize the communication of the different devices through common language and procedure. Accordingly, there is the need to introduce a wrapping layer, consisting of two main sub-layers: the interface (upper) and the communication (lower) sub-layers. The first one provides an interface exposing the methods available through an appropriate interface. It is responsible for the management of all the incoming/outcoming (from/to the upper layer) messaging operations involved in the communication with the object. The second sub-layer implements the logic behind each service methods and translates these methods into a set of device-specific commands to communicate with the real-world objects. Some objects may be very elementary, such as an RFID-tagged object, while others may be equipped with an embedded TCP/IP stack, like TinyTCP, mIP or IwIP, which provide a socket like interface for embedded applications. In the first case a gateway is required to implement such abstraction layer, while in the second case this layer can be implemented in the object itself.

In the third layer, the *Social agent* is devoted to the communication with the SIoT servers to update its profile and friendships, as well as to discover and request services from the social network. It also implements the methods to communicate directly with other objects when they are close geographically or when the service composition needs direct communications between objects. Finally, the *Service management* represents the interfaces with the humans that can control the behavior of the object when communicating with the social network. The social management module and the social agent are usually implemented in an external server but sometimes can be located in the device itself when equipped with enough processing capabilities, as in the case of the smartphone.

4. NEEDS FOR STANDARDS

In our vision there will be more than one SIoT platform working in parallel, which should be interoperable not to limit the potentialities of a network of billions of things and not to require each thing to duplicate accounts, as it is currently happening to human users of social network services. At the first access of each new member, the owner interacts with the platform servers to create the account, insert the object profile data, set the control parameters. The thing then will start crawling the network to look for friends among the platform members (parental and ownership relationships) and managing the relationships when encountering other members (social and co-working/location relationships). The object will also make available its own services (e.g., information from the physical world) to the rest of the network. During these processes all the information related to the object profile, activity and relationships are then stored in the SIoT platform. Whenever an object wishes to move to another platform this information should be transferable in the new systems and this can be done only if a standard representation has been adopted, otherwise the object history

is lost (*blocked* in the original platform). Additionally, an object can encounter another potential friend that is a member of a different platform. For the friendship to be created it is required that each platform exposes to the external world well-known functionalities to retrieve object identity, profile, and services.

These issues are already encountered in the social networks between humans, where the Facebook user profile and history cannot be exported entirely and the interaction with other platforms is limited. Lately with respect to the impressive market success of well-know platforms, some initiatives are going on towards the definition of common languages. The open social is a set of common application programming interfaces (APIs) for web-based social network applications, developed by Google along with MySpace and a number of other social networks. These APIs are intended to allow for accessing data and core functions on participating social networks. Other efforts are devoted to the definition of the syntax to describe people, the links between them and the things they create and do (FOAF; www.foaf-project.org). Similar efforts are needed in the SIoT context, and we hope that this time the standards will be available when the market will require SIoT services!

5. CONCLUSIONS

In this paper we have introduced the concept of the Social Internet of Things (SIoT) that can be exploited to implement scalable service discovery in the IoT. We have investigated how social relationships between objects can be established and managed. Finally, we have proposed a preliminary architecture which is able to implement the major components of the SIoT that have been identified by starting from the components that can be distinguished in current SNS used by humans.

The development of the SIoT concept requires a large research effort devoted to the study of the structure of networks of socializing things as well as to the detailed definition and assessment of the operations required to implement the components identified in Section 3.2.

REFERENCES

- [1] L. Atzori, A. Iera, and G. Morabito, "The internet of things: A survey," *Computer Networks*, vol. 54, no. 15, pp. 2787–2805, October 2010.
- [2] D. Chakraborty, A. Joshi, Y. Yesha, and T. Finn, "Towards distributed service discovery in pervasive computing environments," *IEEE Transactions on Mobile Computing*, vol. 5, no. 2, pp. 97–112, 2006.
- [3] J. Kleinberg, "The small-world phenomenon: an algorithmic perspective," in *Proc. of ACM Symposium on Theory and Computing*, 2000.
- [4] D. Guinard, M. Fischer, and V. Trifa, "Sharing using social networks in a composable web of things," in *Proc. of IEEE PERCOM 2010*, March–April 2010.
- [5] Anthony C. Boucouvalas Evangelos A. Kosmatos, Nikolaos D. Tselikas, "Integrating rfids and smart objects into a unified internet of things architecture," *Advances in Internet of Things*, vol. 1, pp. 5–12, 2011.
- [6] J. S. Coleman, "Social capital in the creation of human capital," *American Journal of Sociology*, vol. 94, no. Supplement, 1988.
- [7] R. S. Burt, *The Social Structure of Competition*, Cambridge, Massachusetts: First Harvard University Press, 1992.
- [8] R. S. Burt, "Structural holes and good ideas," *American Journal of Sociology*, vol. 110, no. 2, September 2004.
- [9] A. P. Fiske, "The four elementary forms of sociality: framework for a unified theory of social relations," *Psychological review*, vol. 99, pp. 689–723, 1992.
- [10] Nick Haslam, "The four elementary forms of sociality: framework for a unified theory of social relations," *Cognition*, vol. 53, pp. 59–90, 1994.
- [11] D. M. Boyd and N. B. Ellison, "Social network sites: Definition, history, and scholarship," *Journal of Computer-Mediated Communication*, vol. 1, no. 13, 2007.
- [12] Mike Gotta, "Reference architecture for social network sites," July 2008.
- [13] J. Breslin and S. Decker, "The future of social networks on the internet," *IEEE Internet Computing*, vol. 11, no. 6, pp. 86–90, 2007.
- [14] D. Diaz-Sanchez, A. Marin, F. Almenarez, and A. Cortes, "Social applications in the home network," *IEEE Transactions on Consumer Electronics*, vol. 56, no. 1, pp. 220–225, February 2010.
- [15] A. Kansal, S. Nath, Jie Liu, and Feng Zhao, "Senseweb: An infrastructure for shared sensing," *IEEE Multimedia*, vol. 14, no. 4, pp. 8–13, 2007.
- [16] S. K. Bansal, A. Bansal, and M.B. Blake, "Trust-based dynamic web service composition using social network analysis," in *Proc. of IEEE Workshop on Business Applications of Social Network Analysis*, 2010.

NET-CENTRIC WORLD: LIFESTYLE OF THE 21ST CENTURY

Daniel Kharitonov
Juniper Networks Inc., Sunnyvale CA USA
dkh @ juniper dot net

ABSTRACT

In this paper, we research the potential of information communication technologies (ICTs) for changing our society from a commute-centric to a network-centric environment. We propose to formalize the key attributes of ICT-based telecommuting experiences from both economic and human interactivity perspective. We introduce the notion of network-eligible transactions and disclose the link between degree of network centrality and worker settlement radius, postulating that media-rich network services have a strong potential to increase the physical distance between work and home locations. We also highlight notable technology challenges and opportunities of migration from location-based to mobile living, signifying the needs for new services and standards development.

Keywords— ICT, net-centric, telecommuting, nomadic

1. INTRODUCTION

Since prehistoric times, the concentration of human activity has been synonymous with density of population. Having started as early as 3,000-4,000 BC, the continuous process of urbanization still goes on today, with the UN estimating that about half of the world is now living in metropolitan areas [1]. With large cities becoming focal points for opportunities, services, and wealth, they tend to attract the massive daily migration of humans also known as *the commute*. Commuting allows workers to reside beyond walking distance from their jobs in exchange for certain inconveniences such as unproductive time loss (averaging over 100 hours per year in the U.S. [2]), pollution of the environment, and transport expenses. While most large cities incessantly invest in mass transit infrastructures, the ongoing shift of economies in developed countries from goods to services [3] suggests the possibility that a growing percentage of commuters could, in fact, use ICT facilities in lieu of their physical presence at manufacturing worksites. And this percentage could be quite significant. A 2010 survey of U.S. government employees [4] revealed that 55% were eligible for teleworking, but only 8.67% of respondents used this opportunity. An even larger gap was found in the ICT sector, where the 2008 survey of 1,500 U.S. professionals [5] found that 37% were genuinely interested in telecommuting to such a degree that they would accept a pay cut, but only 7% could effectively work remotely. Such a discrepancy signifies that the economic and social impact of ICT has not reached its full potential in the workspace, and many aspects of networked humanity

remain unidentified. In this publication, we intend to explore the mechanism of telecommuting relative to the modern state of communications technology and uncover the potential social consequences of this relation.

2. COMMUTE-CENTRIC WORLD

The traditional view on human behavior at work with respect to commute patterns suggests that employees have two choices—show up at the work desk or stay at home and work remotely. This is why many studies and surveys focus strictly on ecological, transport, or economical outcomes of home-work interchange.

However, it is easy to observe that relatively few individuals have an option to reside at arbitrarily selected locations and most spatial-based choices in our lives are dependent upon each other. For example, commute maps tend to strongly correlate with real estate prices (Figure 1), suggesting that the cost of housing plays a significant role in choosing places to live and work.

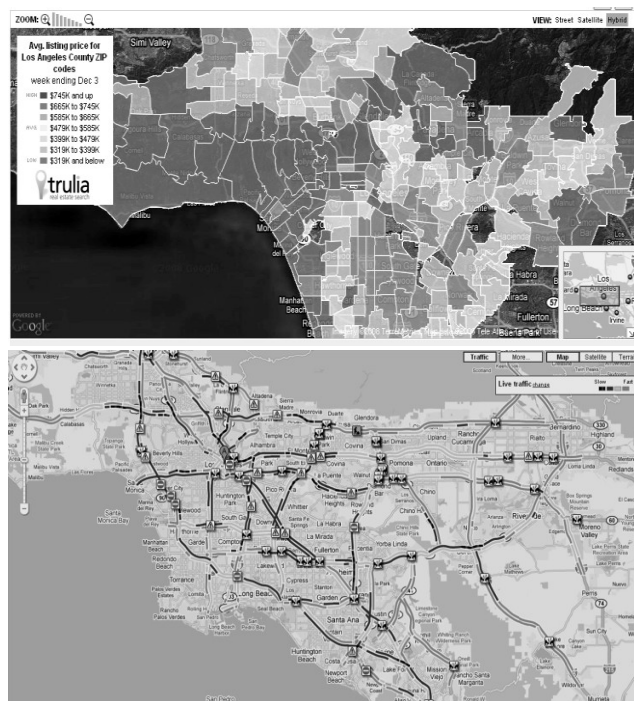


Figure 1. Residence pricing versus rush-hour commute map (source: Google, Trulia)

A graph very similar to Fig.1 is trivial to compile at any location where real-time traffic information is available: the most congested routes tend to be ones connecting the areas with notably different cost of residence.

Whenever a main commute anchor is dropped (e.g., by accepting a job offer or finding a neighborhood with good schools), matching decisions are made to maintain daily activities within a reasonable distance. This means that our lifestyle today is not significantly different from the commute-centric environment of our ancestors, who were bound by pedestrian or equestrian range. A median U.S. worker nowadays lives within 16 miles of work [6], median UK commuters settle within 5 to 10 miles [7], and Canadian workers reside on average within only 7.6 kilometers from their offices [8].

But why exactly do we need to stay physically close to colleagues, businesses, and services while living in what seems to be an increasingly information-driven society?

The answer could be in the fact that we seem to assign tangible value to nonverbal interactions. In his foundational study on how humans store and retrieve information, Edward T. Hall postulated, “language, the system most frequently used to describe culture, is by nature poorly adapted to this difficult task” [9].

The multisensory aspect of human existence was subject to numerous psychology and performance studies. In one example, research of scientist collocation found that the mean citation index for (first-last) author relationship in publications decreased as the physical distance between them increased in the three categorized ranges (same building, same city, or different city) [10]. In another example, applied psychology work discovered that high intensity telecommuting exacerbates the negative impacts on quality of interpersonal relations [11], suggesting certain workspace conflicts (at present state of technology) cannot be effectively resolved without direct human contact.

Such and similar research may shed some light on why large-scale telecommuting (as function of technology) is still not reality, even among the well qualified and eligible social groups.

This brings the logical question—are we bound to live in a commute-centric world and if not, what can be done to change that?

3. TAXONOMY OF HUMAN INTERACTIONS

Significant amount of academic work on cognitive engineering and individual performance suggests that humans are highly capable of multidimensional judgements, with processing in correlated dimensions (such as being able to hear and see a person) improving security of channels and reducing information loss [12].

In fact, there appears to be a broad range of highly interactive, multidimensional experiences, which underpins important functions, cues, and customs of a human society.

For example, numerous studies postulated that spontaneous and informal transactions like hallway conversations, dinners, and face-to-face brainstorming could be important to creating and maintaining productivity in the workplace [13][14]. Similarly, certain high value services (such as

medical consultations and wealth management) are perceived to be more productive when exercised at considerable length in person or across a multitude of channels versus “purely electronic”, mono-channel engagements [15][16].

If we adopt the view that the main value of interpersonal communications lies in richness of informational channels, the necessary conclusion should be that the online workflow (including telecommuting experiences) can be vastly improved with a transition to multichannel information exchanges (such as telepresence and virtual reality systems). This position is supported by evidence that human information-processing capabilities improve with redundant message coding across multiple modalities [17].

However, the replacement of eligible physical modalities with their virtual equivalents is not straightforward.

First of all, digital interactions are grafted over a complex web of technical appliances such as local area networks, wide area exchanges, user terminals, applications, and so forth. As a result, the quality and precision of electronic experiences can go down when details are lost “in translation” due to noise, latency, analog-to-digital conversion, compression, and other technology artifacts.

Second, certain sensory dimensions – such as olfactory, gustatory and kinesthetic experiences are hard to reproduce remotely, at least with existing telepresence equipment.

Finally, whenever we introduce the network into human-to-human transactions, the cost of delivery changes significantly.

Assuming it takes the same time to conduct an interactive session face-to-face or electronically, the main direct cost associated with “mortar-and-brick” participation is related to transport, i.e., the ability to meet a colleague, business partner, or physician in person may cost from pennies to thousands of dollars (a long-haul flight), with the U.S. average being 50 cents per mile (Table 1).

Indirect costs can run much higher. A worker may need to pay a premium for a house that is within an acceptable distance to work, an employer may have to sustain the soaring cost of office space in the middle of a good business district, and so on.

Table 1. Human-human information exchanges

Exchange	Media	Delivery cost
Message	Written memo or note	\$0.44/letter
Verbal	Conversation	\$0.5/mile
Visual	Face-to-face meeting	\$0.5/mile
Multimodal	Lunch, hallway talk, physical treatment, brainstorm session	\$0.5/mile

On the opposite, network-based interactive transactions are priced according to “buckets” of connectivity (Table II). For instance, we may largely assume that multimodal telepresence and virtual reality (VR) systems are free of telecom charges within the corporate local area network (LAN), but they may not work over consumer-grade

Internet connections such as residential broadband lines. On the other hand, running a private leased circuit from home to office for full-scale telepresence can be feasible in large metropolitan areas but priced out of reach for all but the wealthiest telecommuters.

and quality of ICT services (network). Since the price and availability of network services are largely decoupled from their physical location, we can reasonably claim that this new behavior model starts to drift away from legacy, location-driven society. To reflect this difference, we will refer to the new, nomadic lifestyle as “net-centric.”

Table 2. Human-network-Human Information Exchange

	Virtual Experiences				
	Message	Verbal	Visual		Rich multimodal
Media	<i>IM/SMS</i>	<i>Phone</i>	<i>Video Stream</i>	<i>Telepresence</i>	<i>Virtual Reality, 3D Video, etc.</i>
Quality/reliability	Medium	High	Low	High	High
Minimum bandwidth	160 Bytes	9 Kbps	0.2-2 Mbps	2-4 Mbps	5-20 Mbps
QoS requirements	Low	High	Medium	High	Very high
Proximity/Cost					
Local/LAN	\$0	\$0	\$0	\$0	\$0
Metro area/DSL	~\$0	~\$0 (VoIP)	~\$0 (IP Video)	\$0.5/min ^c	\$10/min ^c
National mobile	\$0.1 ^a	\$0.25/min ^a	\$0.05/min ^b	Not supported	Not supported
International mobile	\$0.5 ^a	\$4 ^a	\$10/min ^b	Not supported	Not supported

a) U.S. average GSM voice and short text tariffs b) U.S. 3G data tariffs c) U.S. leased line tariffs based on one hour/day usage

Further, as employees leave their residential areas and work on the go, their mobile carrier will also charge them for airtime. A cellular service provider will typically bill for all messages, calls, and data transactions conducted by smartphones and wireless capable tablets, while frequently imposing restrictions on media experiences (audio, video) that may become crippled in quality or accessibility.

Finally, telecom expenses can mount fast outside the user’s home country, as crossing the border often invites roaming fees. It is still not uncommon for smartphone users to face abnormally high telecom charges accumulated abroad over rather trivial usage profiles [18].

4. NET-CENTRICITY

Assuming employers are generally willing to support remote collaboration, we can now formulate a hypothesis that economy of telecommuting is primarily driven by the confluence of available interaction levels and related costs. An act as simple as reconciliation of a business discussion may be impossible without a trip to meet one’s peers and shake hands; at the same time, a transaction as complex as surgery can be successfully done remotely despite extremely high telecom and robotic equipment costs [19].

Thus, we can presuppose that telecommuting is only practical when it offers a suitable compromise between the cost of information exchanges and the ability of ICT infrastructure to sustain an effective workflow. If achieved, such compromise should mark an important change in human behavior—the increasingly connected work ecosystem becomes less dependent on physical distances (commute) but more dependent on availability, economy,

4.1. Net-Centric Factor

The first question that comes to mind when defining characteristics of a net-centric world is what percentage of duties can be fully performed over existing telecom infrastructure. In the pre-Internet era, very few occupations were eligible for full-time telecommute, with the rest of the economically active population glued to physical workplaces. Today, a significant percentage of the population in developed countries may, in fact, perform duties remotely, at least partially [20]. Thus, every job can be described with a metric that reflects the percentage of work that can be robustly and economically done over the network. Let’s call such a metric a net-centric factor (NCF):

$$NCF = \text{online tasks} / (\text{offline tasks} + \text{online tasks}) \quad (1)$$

For example, a professional technical writer, who does not depend on personal collaboration with co-authors or publishers, may achieve an NCF close to one. On the other hand, a hair stylist will likely have an NCF = 0 simply because specialized machinery for remote coiffeur services is economically prohibitive to build, given the prevailing haircut rates. Considering that every active worker may have a unique combination of possible online and offline actions and duties, NCF is highly personalized. For example, a person who may effectively come to the office three days a week has a de facto NCF factor of 0.4 (forty

Table 3. Sample Economic impact of NCF

Distance to Work/ Time to Work	House/ Month	Commute		Monthly Cost of Housing Plus Transport						
		Mode	Cost*	NCF = 0	0.2	0.4	0.6	0.8	0.9	0.95
0 miles / 5 minutes	\$5,000	Walk	\$0	\$5,000	\$5,000	\$5,000	\$5,000	\$5,000	\$5,000	\$5,000
5 miles / 20 minutes	\$3,500	Tram/Bus	\$25	\$4,000	\$3,900	\$3,800	\$3,700	\$3,600	\$3,550	\$3,525
10 miles / 30 minutes	\$2,500	Car	\$40	\$3,300	\$3,140	\$2,980	\$2,820	\$2,660	\$2,580	\$2,540
25 miles / 45 minutes	\$1,500	Car	\$71	\$2,980	\$2,636	\$2,352	\$2,068	\$1,784	\$1,642	\$1,571
40 miles / 60 minutes	\$1,200	Car	\$100	\$3,200	\$2,800	\$2,400	\$2,000	\$1,600	\$1,400	\$1,300
100 miles / 120 min.	\$1,100	Car	\$200	\$5,500	\$4,620	\$3,740	\$2,860	\$1,980	\$1,540	\$1,320
1000 miles / 180 min.	\$900	Train/Air	\$480	N/A	\$6,980	\$5,460	\$3,940	\$2,420	\$1,660	\$1,280
2500 miles / 300 min.	\$1,100	Air	\$900	N/A	N/A	\$12K	\$8,300	\$4,700	\$2,900	\$2,000
2500 miles / 300 min.	\$1,200	Air + hotel [†]	\$900	N/A	N/A	\$5,400	\$4,100	\$2,900	\$2,500	\$1,850
6000 miles / 840 min.	\$1,300	Air + hotel [†]	\$2,340	N/A	N/A	N/A	\$7,580	\$4,440	\$2,870	\$2,085

* Roundtrip cost, including time loss at \$0.5/minute and transport at \$0.5/car mile or \$200/\$600/\$1,500 for short/mid/long-haul airtickets

[†] Housing cost includes \$200/night hotel surcharge on commute days.

percent tasks can be done offline), while a peer in the same job may have an NCF factor of 0.1 or even less¹.

It is also important to note that NCF merely describes the potential for effectively doing the job remotely and has to be augmented by availability and cost of technology. From a practical standpoint, NCF denotes the minimal frequency of commute required to maintain normal productivity level at work. An NCF of 0.6, for example, allows for commute twice a week, an NCF of 0.8 once a week, and so on. Also of note is that high NCF values do not necessarily mean a proportional reduction in transport distances or expenses, as a teleworker may come to the office more frequently or choose to reside further away from it.

4.2. Net-Centric Economy

When discussing the taxonomy of human interactions, we have mentioned that the cost of in-person transactions is linear and consists of commute expenses plus an indirect premium for residing within an acceptable commute radius. On the other hand, the cost of pure networked transactions is discrete and is entirely driven by connectivity. Thus, the combined economy of living in the net-centric world can be formalized with this equation:

$$Sv > \sum_{i=1..N} C_i t + C_c * (1 - NCF) \quad (2)$$

where

Sv denotes the value of remote work due to better location, cheaper housing, better living conditions, etc.

¹ Although there is some evidence that employee output may change by the mere act of telecommuting or fluctuate with tenure, skill or task interdependency [21][22], in this paper we assume that at any career point, their NCF can empirically established.

$Sum_i (C_i t)$ denotes the sum of telecom transactions across all N media types required to support online tasks. This includes amortization cost of all necessary software programs and hardware appliances.

C_c denotes the cost of daily commute, including transportation, security, time, and other expenses needed to support offline tasks.

One of the possible ways to quantify Sv is to note that the cost of real estate is typically inversely proportional to a settlement radius. For instance, if an employee works in San Francisco's financial district, the cumulative cost of housing and transport (based on a five day work week) may look similar to that shown in Table III.

When we calculate housing costs based on a typical estate pattern similar to that shown in Figure 1, the immediate vicinity of premium office space (0–5 miles) commands the highest prices, which gradually decrease as the residence moves away from the business center and into suburbs. At the same time, commute costs build up both with distance and the number of commute days per week. This explains the empirical “sweet spot” found by surveys—without telecommuting (NCF = 0), it is most economical to settle within a (certain) city, country, and region-dependent optimal distance from work (U.S. average being 16 miles).

However, as NCF increases, so does effective settlement radius. Working remotely two days a week (NCF = 0.4) makes it feasible to reside a bit closer to work, but also strongly motivates workers to move further away from the office. Shown in bold in Table III, the acceptable telecommuting solutions (cost of housing plus transport less or equal to that of the best location with NCF=0) clearly demonstrate that the settlement radius increase is proportional to NCF.

Higher NCF values may also result in new lifestyle options.

A high-intensity telecommuter coming to the office once a week (NCF = 0.8) can economically reside within regional jet or a high-speed train reach and be qualified for such work-home combinations as “San Francisco–San Diego” or “Zurich–Berlin.” Additionally, the equation (2) suggests even cross-continental work habits may make economical sense. With an NCF of 0.8 or more, an employee may live away from Northern California to as far as Hawaii (2,400 miles) or Montreal (2,600 miles) - considering the difference in median house prices, this may actually make a

case for relocation. Even more surprisingly, the carbon footprint of cross-continental commuters can still beat their office-dwelling colleagues. With the U.S. national average of 15,000 miles per driver, any employee using a car with ordinary fuel efficiency could just as well spend about 50 hours per year on commuter jets.

Finally, nomadic and ultra long-haul commutes make an extreme, but still a sound business case. With the ability to reside at the worksite temporarily (e.g., using hotels for accommodation), telecommuting may span hemispheres.

In this latter case, the cost of housing should not be the only (or most important) reason for living at remote locations, so Table III assumes that ultra long-haul commuters pay more than the lowest neighborhood prices for their choice of residence and ability to stay close to transport hubs.

If we plot Table III as a function of cost of living relative to commute distance, we will observe that all existing workers residing at a nontrivial distance from the office can financially benefit from an increased amount of telecommuting (Figure 2).

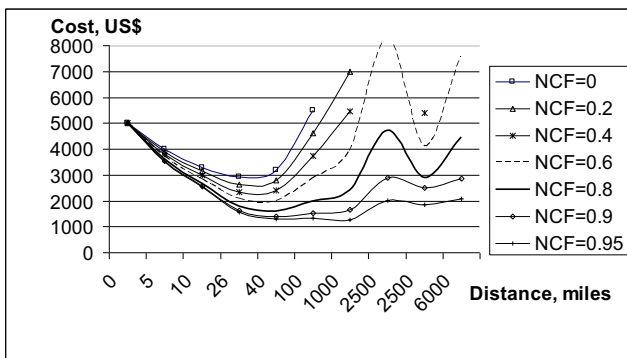


Figure 2. Cost of housing and transport relative to commute distance for various NCF values

This forms the basis for the ICT business case. If telecom services can act as enablers for higher intensity telecommute, consumers and service providers can benefit together from a transition to the new work model.

5. ICT COSTS AND OPPORTUNITIES

So far we have intentionally stayed away from quantifying the expression $Sum_i(C_i)$, i.e., we did not put any bounds or restrictions on cost of telecom services required to support high-intensity telecommuting.

Such boundaries can be trivially established by resolving expression (2) for known values of S_v , NCF, and transport. If we plot the cumulative financial gain from telecommuting due to lowering transport expenses using sample data from Table III (Figure 3), we notice that workers residing within immediate vicinity to work (0-5 miles) do not have financial drivers to practice low intensity telecommuting (NCF 0.2 to 0.4). This category of workers may still realize some savings (less than \$500 per month) from higher intensity telecommuting patterns, but they are not likely to be motivated to increase spending on telecom products beyond their normal utility packages.

Quite predictably, workers with residences beyond the average distance may realize sizeable profit from even low-intensity telework patterns - such as coming to the office three to four times a week. These people should be financially interested in sustaining their net-centric lifestyle, as they can definitely increase their telecom spending beyond the minimum package and occasionally may afford services with recurring monthly costs up to \$500 (or even higher). A typical user from this group would be an executive or highly paid professional whose telecom expenses can be partially covered by the company or may remain insignificant relative to salary.

However, the most interesting case is seen at the median commuter radius (10-40 miles), where cost-conscious workers may gain \$200 to \$500 per month with minimal telecommuting efforts and up to \$1,000 or more for higher intensity telecommuting.

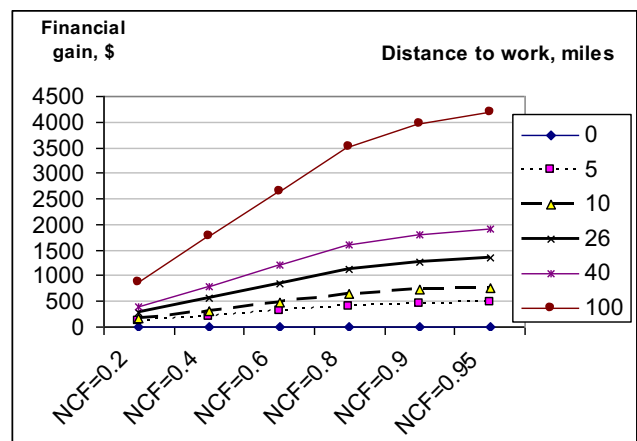


Figure 3. Financial gain due to reduction in cost of housing and transport relative to commute distance

This latter group of customers makes the “bread-and-butter” business case—if telecom providers can provide quality work-home communications at residential locations, they should be able to increase monthly account charges by \$50 to \$200 – in other words, nearly double or triple the current average revenue per user (ARPU).

However, matching the content of Table II against this target revenue also reveals a paradox—the current telecom industry does not provide products that can enable multimodal experiences in a suitable price range.

This gap is surprising, because the ability of humans to absorb information across sensory channels is biologically constrained [23][24] while the progress of codecs, presentation, and broadband access speeds continues at a steady rate. In fact, we can postulate that the ability of ICT systems to transport the amount of information matching the capacity of all human sensory channels is unquestionable, with the only problem being how to cross this barrier economically.

Moreover, the majority of urban population in developed countries already has access to broadband Internet in the speed ranges quite suitable for high quality video streaming [25], while a growing percentage of fiber-connected residents may afford to run virtually any available streaming application. According to Organization for Economic Cooperation and Development (OECD), the median advertised broadband download speed in 2010 was 15 Mbps with prices ranging from \$0.13 (Japan) to \$11 (Mexico) per megabit per second [26]. Therefore, in theory, rich multimedia experiences can be economically delivered in most significant urban hubs worldwide.

In reality, however, Internet service providers (ISPs) remain mostly oriented towards best-effort services, with no guaranteed connections (virtual leased lines) offered to consumers even within their “home” network much less across different service providers.

For instance, a popular voice over IP (VoIP) and video conferencing applications Skype uses an array of audio codecs, including G.729 with lowest bitrate of 8Kbps [27]. At the same time, business-grade VoIP platform Skype Connect™ manual recommends the minimum of symmetrical 33Kbps connection speed (up to six sessions over 256Kbps/512Kbps ADSL service), suggesting 4x the bandwidth over-provisioning to cover for lack of explicit QoS on the Internet connections users [28].

In another example, a leading US streaming provider Netflix reveals that their subscribers on top US networks are able to watch TV and movies at speeds ranging from 1400 Kbits per second to 2700 Kbits per second, with “no client being able to sustain 4800 <Kbits per second> stream from start to finish” [29]. Considering the fastest service provider from Netflix list (Cablevision) in 2011 offered the minimum download access speed of 15Mbps, it took over 5x of over-provisioning to maintain one video streaming application. It is even more interesting to note, that Netflix application typically runs between directly connected networks - last-mile Internet service provider and content-delivery operator like L3, Limelight or Akamai.

So practically speaking, while ordering targeted QoS parameters from any broadband provider today is not possible, consumers and businesses have to pay for access speeds several times higher than the bitrate required for applications they are interested in.

By extension of this example, if we consider running rich media session with over Internet in business environment with quality parameters similar to that of needed by Skype (0.2% or less packet loss, 10ms or less jitter and 200ms or less of delay), a broadband connection required to support telepresence or virtual office sessions may need sustainable

access speed ranging from 40 to 200 Mbps—something not feasible in the nearest future, especially over copper or airwaves.

6. CALL FOR STANDARDIZATION

In the previous section we hinted at the possibility of new, high margin telecommunication services to support interactive applications. For example, a high-speed, guaranteed QoS “virtual leased line” between home and office might, in fact, become a popular service if priced to satisfy the restrictions of our equation (2). We can also foresee a market segment for novel types of consumer collaboration and media applications such as virtual offices, virtual multimodal meeting rooms, and so on.

However, the task of developing signaling, forwarding, and billing solutions for inter-provider QoS-aware tunnels in a generalized, N-service provider format (Figure 4) presents a notable challenge for vendors, network architects, and standards organizations alike.

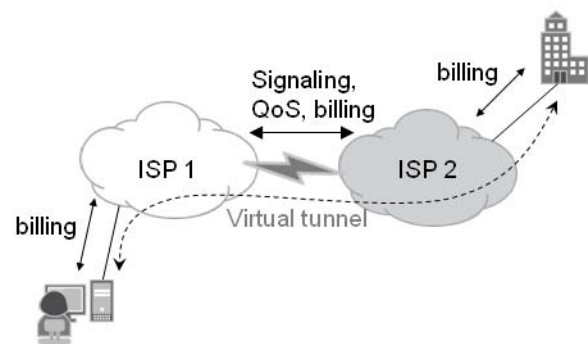


Figure 4. Virtual tunnel between home and office

On one hand, service providers are hesitant or unwilling to invest into developing proprietary application architectures that may not find a matching interface at their Internet peers. Therefore, early involvement of international standards organizations (such as ITU-T) is a must.

On another hand, standardization of all-inclusive network orchestration architectures is a slow, daunting task that requires designing consistent and unified policy management in a system with complex moving parts. The partial list of problems to solve includes mechanisms for connection admission and control, resource reservation, runtime verification of service-level agreements (SLAs), multiparty billing and packet handling programming for transit network devices (routers and switches). Last (but not least) are security concerns that include control for misconfigurations, runaway client devices and network resource abuse by humans and robots as well as integrity and confidentiality guarantees for client data.

This is why, despite the strong body of work on external network-to-network interface (E-NNI) definitions by various organizations including Optical Internetworking Forum [30], Metro Ethernet Forum [31] and pioneering efforts by IPSphere framework group within Traffic Management Forum [32], practical development of session-

based inter-provider QoS services and interfaces remains in the early stage.

The relatively poor condition of de facto and de jure standardization in this area can also be (partially) explained by scarcity of session-based IP services suitable for immediate monetization. However, this deficit works both ways—the lack of session-based services is also an inhibiting factor for development of novel interfaces, interactions, and learning technologies. Thus, fostering and encouraging standardization efforts in this area should resume from making a clear, executable mission statement on the technical subject and related business case.

7. STANDARDIZATION PROPOSAL

The concept of net-centricity assumes close, robust relation between remote workers and media-rich corporate services over network infrastructure. We consider this concept to be pivotal to a future stack of standards defining provider-agnostic application-aware networking (AAN) [33].

If we take a closer look at Figure 4, we may notice it can be simplified into two possible architectures – (1) Content platform based interactive services and (2) “walled-garden” based interactive services.

In a first use-case, a corporation willing to offer rich media experience to remote workers moves its content (such as virtual office environment or 3D telepresence sessions) to a commercial content delivery network (CDN). Considering that CDNs maintain direct peering with all major service providers, this move guarantees that QoS planning, delivery and external network-to-network interfacing remains constrained within the latter, thus greatly simplifying the original IPSphere service planning model [32]. Once last-mile service provider authorizes and accepts service request from the user, it is routed to the nearest CDN operator, which in turn bills content owner based on usage. In that case, access operator acts as both Element Owner (participating in cross-domain design and delivery) and Administrative Owner (offering its own transport services for retail). This allows for session admission and control to run only once (at customer interface) and billing to be complete in two cycles (CDN to corporation and ISP to CDN), while reliably serving the needs of large national and international businesses and their remote employees.

In a second use case, a corporation willing to offer rich media experiences to remote workers moves its content platform directly into “walled service garden” of the last-mile service provider. This model reduces the number of parties to two, but has disadvantage of lower scaling parameters (one content platform is needed per every supported ISP) and better fits regional businesses.

Both architectures are significantly simplified relative to all-encompassing QoS architecture that is required to support an arbitrary number of applications over chain of service providers with complex mix of capabilities. On the other hand, our proposal can be seen as stepping stone for evolved application-aware services – such as subscription-based gaming, remote medical diagnostics, 3D webcasts and others.

The implications of proposed standardization efforts can be significant.

With 90% of the world’s metro areas already residing within only 250 ms of network delay [34], the net-centric lifestyle based on robust, media-rich electronic workflow has strong potential of crossing borders and enabling innovations and virtual communities in ways that are not feasible or even foreseen today. Of particular interest we can also note the confluence of application-aware network services and mobile / rural broadband coverage, which may contribute towards acceleration of human development both in urban territories and communities that insofar have fallen behind the digital age.

8. CONCLUSIONS AND FUTURE WORK

In this paper, we investigated the socioeconomic aspect of telecommuting. We have introduced the notion of Net-Centric Factor (NCF) and studied the links between intensity of telecommuting and feasible commute distances. Our formulated value of telecommute allowed us to show that network centrality allows remote workers to increase their effective settlement radius above and beyond the limits characteristic to legacy, commute-centric lifestyle.

Further, we have looked at economic incentives for telecommuters to increase their NCF and have found that such requests cannot be served with the currently available “best-effort” broadband infrastructure, thus pointing towards new network-based service opportunities.

Our work indicates that internet developers and international standards organizations have strong potential to develop new, high-margin and QoS-guaranteed consumer and business services. In a proposed extension of this work, we consider focusing on practical use-cases and simplification of existing architectures and orchestration abstractions down to practical, executable essentials.

REFERENCES

- [1] “World Urbanization Prospects: The 2009 revision highlights”, UN Department of Economic and Social Affairs: March 2010
- [2] “Americans Spend More Than 100 Hours Commuting to Work Each Year”; US Census Bureau; March 2005; Community Survey [CB05-AC.02](#)
- [3] “The Service Economy” OECD Business and Industry Policy Forum Series; 2000
- [4] “Status of Telework in the Federal Government 2010” U.S. Office of Personnel Management; Feb 2011
- [5] “The Dice Report” Dice Holdings; [Jun 2008](#).
- [6] “Poll: Traffic in the United States” ABC News; Feb [2005](#).
- [7] “Report on Motoring” Royal Automobile Club UK [2006](#).
- [8] “Commuting Patterns and Places of Work of Canadians” Statistics Canada: [2006 Census](#)

- [9] Hall, Edward T. Beyond Culture. New York: Random House, 1977
- [10] Lee, K., Brownstein JS, Mills RG, Kohane IS, “Does Collocation Inform the Impact of Collaboration?” PLoS; 2010 Dec 15; vol 5 (12)
- [11] Ravi, S. Gajendran and David A. Harrison, “The Good, the Bad, and the Unknown About Telecommuting: Meta-Analysis of Psychological Mediators and Individual Consequences”, Journal of Applied Psychology; 2007; Vol. 92, No. 6
- [12] Wickens, C. and Hollands, J. Engineering Psychology and Human Performance. Upper Saddle River, NJ: Prentice Hall, 2000
- [13] Sutton, RJ. and Hargadon A, "Brainstorming groups in context: Effectiveness in a product design firm" Administrative Science Quarterly; Dec 1996; v.41, no.4, p.685-718.
- [14] Kraut, Robert E, Fish Robert S., Root Robert W., and Chalfonte Barbara L., “Informal Communication in Organizations: Form, Function, and Technology” Morristown: Bellcore, 2002
- [15] Sunikka, Anne “Predominantly Electronic or Personal Service Delivery? A Case in the Wealth Management Context,” ECIS 2009 Proceedings; 2009; Paper 166
- [16] Lin, Chen-Tan et al, “Is Patients’ Perception of Time Spent With the Physician a Determinant of Ambulatory Patient Satisfaction?,” Archives of Internal Medicine; Vol 161, June 11, 2001.
- [17] Miller, J. “Channel Interaction and the redundant-targets effect in bimodal divided attention”. Journal of Experimental Psychology; 1991; 17, pp 160-169
- [18] Falch, Morten, Henten, Anders and Tadayoni Reza, “International roaming: is there a need for EU-regulation beyond 2010?” Emerald Publishing; Volume 11 (4): pp 19-33; June 2009
- [19] “Telesurgery to Impact Medical Care” MedMarkets; Volume 3, issue 9; Sep 2004
- [20] “Telework and sustainable development: A case study with the Global -Sustainability Initiative (GeSI)”, Digital Europe; IST-2000-28606; Apr 2003
- [21] Butler, S., Aasheim C. and Williams S., “Does telecommuting improve productivity?” Communications of the ACM, vol. 50, issue 4; April 2007
- [22] Turetken, O., Jain, A., Quesenberry, B. and Ngwenyama, O. “An Empirical Investigation of the Impact of Individual and Work Characteristics on Telecommuting Success” IEEE Transactions on Professional Communication, vol.54 no.1; March 2011
- [23] Deering, Michael F., “The Limits of Human Vision,” Sun Microsystems, 1998.
- [24] Milind N. Kunchu, “Temporal resolution of hearing probed by bandwidth restriction,” Acta Acustica, Vol. 94, Pgs. 594–603, 2008.
- [25] “Fixed and wireless broadband subscriptions per 100 inhabitants” OECD Broadband Portal; Jun 2010
- [26] “Average advertised download speeds by country” OECD Broadband Portal; Sep 2010
- [27] “Skype Protocol” Wikipedia, the free encyclopedia
- [28] “Skype Connect”, User Guide, ver. 4.0; Skype Ltd; 2011
- [29] Florence, K. “Netflix Performance on Top ISP Networks”, Netflix Tech Blog; January 2011
- [30] “Intra-Carrier E-NNI Signaling Specification 1.0” Optical Internet Forum; 2004
- [31] “MEF 26. External Network Network Interface (ENNI) Phase 1”; Metro Ethernet Forum; January 2010
- [32] “QoS aspects of IPSphere” TM Forum, TR157; Release 1.0; May 2011
- [33] “Application-aware Networks: Evolving carrier business models with application-aware technology” Juniper Networks, Inc; 2009
- [34] “Summary statistics for all Archipelago Monitors” CAIDA; Archipelago Project Statistics, RTT quartiles for all monitors

REFLEXIVE STANDARDIZATION OF NETWORK TECHNOLOGY

Ian Graham

University of Edinburgh Business School, Edinburgh, United Kingdom.

ABSTRACT

This paper investigates a JTC1 working group to identify how formal standards processes are evolving in response to globalization and the emergence of consortium standardization. It is found that being part of the formal standards development systems provides a source of legitimacy, but also limits the freedom for the process to replicate the structures of consortia. Their standardization process is deeply reflexive, with focuses on maintaining legitimacy and negotiating the boundaries where their activities impinge on other processes. It is argued that the structure of committees of multiple national standards bodies feeding national requirements into the global processes by responding to ballots resolutions and nominating representatives is increasingly anachronistic in a world of global communications, more open standards development, global technology companies and the weakening of the ability of states to identify a national interest in technology policy.

Keywords— Formal standardization, JTC1, sociology of standards, globalization.

1. INTRODUCTION

Standardization is crucial to the development of network technologies, where the value to the user is enhanced through interoperability. For this reason innovative network technologies are increasingly being developed within global inter-organisational standards development processes. This standardisation may be within the formal procedures of ISO or ITU or in less formal consortia of firms. These processes are important because they shape the functionality of the technology and influence its cost. Network technologies are a driving force in globalizing societies, and one manifestation of globalization is the emergence of supra-national technology consortia. Consortia offer their members the possibility of developing standards within processes they can design to meet their needs, unrestricted by the established rules and structures of formal standardization. However, the growth of consortium standardization increases the risk of fights for legitimacy between competing processes [1]. The aim of this paper is to consider critically a formal standards process to argue that the complexity of standardising network technologies has fundamentally changed away from concentrating on the use of the technology towards a focus on ensuring the

process is recognised as **the** legitimate standards body within a clearly defined area of standardization.

2. A SHORT HISTORY OF FORMAL STANDARDIZATION: THE CONSORTIUM CHALLENGE

Formal international standardization can be traced back to the founding of the International Telegraph Union in Paris in 1865 [2], creating a forum for the negotiation of standards to ensure network interoperability. In 1901 the Engineering Standards Committee in the UK was established by the Institution of Civil Engineers to develop industry-wide standards. Its first standard, in 1903, rationalized the specifications for rolled steel sections [3]. The Engineering Standards Committee evolved in 1931 into the British Standards Institution (BSI). The basis of the BSI was as an open, industry-led voluntary organization to develop standards for industrial sectors, supported by government but controlled by industrial members. Similar voluntary bodies have emerged in Germany (DIN), France (AFNOR) and the United States (ANSI) to develop national standards. In 1905 the International Electrotechnical Commission (IEC) was established as an international body to agree global electrical technology standards. Unlike its national representative members, the IEC was not open to all, but was based on achieving consensus among its national members. Participants in the IEC processes were therefore nominally representing their country's interest. This model based on achieving consensus between national member bodies was adopted by ISO (International Organisation for Standardization) when it was established in 1946 to develop international standards. ISO, IEC and ITU became the dominant bodies in the development and ratification of international standards across an enormous range of domains. This global formal standards process has been described as encompassing four principles of organization: *expertise, representation, user orientation and participation* [4].

Schmidt and Werle [5] identified that the hierarchical structure of national representation was perceived as a barrier to the development of standards and was leading organizations to co-operate and form "para-standardization" bodies, citing as an early example the European Computer Manufacturers Association (ECMA) that was founded in 1961. Closed consortia, where a group of firms collaborate in a technological field, are processes to develop *de facto*

standards [6] where market selection determines whether the output of the process is accepted a standard. However, by opening a consortium to any interested firms and breaking, or at least weakening, links with the patents of the founding members, consortia begin to resemble less the collaborative development of a proprietary technology and more the privatising of a formal standards development process. This process of legitimation by consortia is repeating the process by which the national standards bodies formed as collaborations between industrial firms and whose nascent organizations were then recognized by national governments [7,3].

It has been argued [8] [9] [10] that the rise of consortia has been a response to the formal standards processes becoming too slow, bureaucratic and unrepresentative to develop standards meeting user needs. Exogenous factors that have also been proposed as triggering this change include the 1993 National Co-operative Research and Production Act in the United States reducing the anti-trust risks of informal inter-organizational alliances [11], the emergence of the Single European Market [8] and the Agreement on Technical Barriers to Trade [12].

3. PROCESSES OF STANDARDIZATION

The two central questions in the economic analysis of standardization have been explaining why standards emerge and quantifying the benefits of standards for the economy [13,14,15,16,17]. By comparison, only a limited number of economic studies have examined the institutional standard development process, most notably Farrell and Saloner [18] who claimed that standardization through committees is more efficient than market standardization, but will take longer to produce the standard. David and Shurmer [8] argued that firms choose informal standard consortia over formal processes due to their characteristics, including flexibility, speed and the ability to tailor their membership and internal organization and procedures to the specific task at hand.

To address how the procedures of standards bodies affect their ability to develop standards, Schmidt and Werle [5] analyzed the organizations co-ordinating standards development as emerging institutions. This follows DiMaggio and Powell's [19] definition of an institution as "a system of rules that structure the courses of action a set of actors might choose". As they are emerging in response to perceived weaknesses in the existing formal standards bodies, we might expect standards consortia to operate using radically different procedures, but it has been noted that they tend to adopt institutional features from the formal processes [20], including due process, transparency, consensus and voluntarism. This "mimetic isomorphism" [19] provides legitimacy by meeting potential users, participants and regulators expectations of a valid standards process [1].

As the number of standards consortia grows the likelihood of standards processes competing would be expected to increase. While so-called "standards wars", notably between VHS and Betamax in the eighties, have been widely studied [21,14,22], they are the exception. Werle notes that "co-ordination and coexistence is the prevailing structural pattern" [20]. Tamm Hallstrom argues that international standardization can be analyzed based on four principles of organization: expertise, representation, user orientation and participation [4]. She identifies potential conflicts faced because participants in the process are simultaneously representing their area of technical expertise, the national interest of the country that has accredited them and the interests of potential users of the technology. In a survey of participants in formal standards working groups Jakobs et al. found closer identification with country and employer in ITU than in ISO groups and that experience in the processes was a significant factor in influencing the process [23]. There is therefore a wealth of theoretical argument and some empirical evidence that the processes of standardization are socially complex and influence both the ability of the bodies to develop standards meeting user needs and to be recognized as the legitimate body to be developing these standards.

4. A FORMAL STANDARDS WORKING GROUP IN ACTION

To investigate the negotiation of formal standards a working group within ISO was selected. It was chosen because it was a new process, so the people involved in forming the group were still active in it, and its remit, to develop mobile item identification, crossed over with other standardization processes within the formal processes of ISO/IEC, within national standardization initiatives and with industrial consortia, so it was expected that the actions carried out to gain legitimacy in this contested field could be uncovered. Permission was requested through a national standards body to attend the second meeting of the group in April 2009, which was accepted by the working group. The meeting was attended by the researcher as an observer, the procedures described and the opportunity taken to interview process participants.

In 1987 the relationship between ISO and IEC in information technology standardization was clarified by the establishment of the Joint Technical Committee 1 (JTC 1) as a joint committee of ISO and the IEC. JTC1 has a hierarchical structure of subcommittees and working groups. Subcommittee 31 of JTC1, SC31, was established in 1996 to develop standards for automatic identification and data capture, the most prominent being the standards used for radio frequency identification (RFID) tags. The Working Group approached for the study was the latest working group, WG6: Mobile Item Identification and Management.

In June 2007 SC31 endorsed the setting up of an ad hoc group, Mobile Item Identification and Management, to investigate establishing a new working group for the

standardization of mobile items interacting with readers. The prime mover of this initiative was ETRI, the Korean research organization, who were interested in adding RFID readers to mobile phones. The ad hoc group was chaired by one of the existing US representatives in SC31. He saw the proposed group as a forum combining the Korean interest in using mobile phones to read RFID tags, Japanese interest in incorporating 2D bar code readers mobile phones and developments in the US within IEEE and NIST to link networks of sensor devices. The scope of the proposed working group therefore widened as the ad-hoc group enrolled complementary areas of interest not covered by existing working groups. All three technologies did not have a clear home in the existing structure of JTC1 committees. A proposal was presented to SC31 to set up a working group to standardize mobile reading of tags. In June 2008 SC31 established their sixth working group, WG6, to develop these standards. A JTC1 working group requires Work Items defining specific tasks, and WG 6 was set up with 8 work items. The first meeting of WG 6 was held in Vienna in April 2008 with representatives from nine countries and liaison representation from GS1, the RFID standardization consortium, and IEEE.

The scope was defined as: “Standardization of automatic identification and data collection techniques that are anticipated to be connected to wired or wireless networks, including sensor specifications, combining RFID with mobile telephony, and combining optically readable media with mobile telephony.” It was argued by one of the drafters that “the scope of a committee is usually written with specificity to prevent infringing on another committee and broad enough to allow new opportunities to be incorporated”. The scope was drawn up to encompass the convergence of mobile telephony, mobile commerce and RFID. The convener saw that the WG should be a location that future technologies in these areas could “find a friend”. A participant described this broadening of the remit as “scope creep”. It was argued by participants that the ISO/IEC policy of allowing standards to incorporate existing intellectual property so long as it was licensed globally on the “reasonable and non-discriminatory” basis makes JTC 1 an attractive route for technology developers to spread the uptake of their intellectual property, and with the limited resources within the process it is quicker to adopt or adapt existing specifications than develop completely new ones.

As with the existing SC31 working groups, the scope was defined in terms of the technical function of the standards and not in terms of applications which would use the standards. The scope could include using mobile telephones to read radio tags, mobile telephones to read bar-codes and the remote reading of sensors. WG6 therefore offered a route for national standards initiatives in each of these areas to feed into the JTC1 processes to influence international standards. While the work group could act as a route into ISO standardization for existing proprietary or national standardization, its scope impinged on other areas of standardization, both within and outside

JTC1. Members of WG1 were then appointed to liaise with other internal processes and with outside standards processes to avoid unnecessary duplication. External liaisons included: ETSI, the European body developing mobile phone standards; ITU, the formal international developer of telecommunications standards; the Near Field Communication Forum, a consortium developing short range high frequency radio communication; and GS1, the developer of item identification standards in logistics.

WG6 planned to hold annual physical meetings supplemented by teleconferences midway between meetings. The second meeting was held over two days in a hotel in a mid-Western US city and was hosted by the group’s convener. The meeting was attended by 16 official delegates, out of an eligible membership of 38, from only eight countries.

The first agenda item was a “Ballot Resolution Meeting” to resolve issues raised during the balloting of national bodies on a previously circulated draft of a standard. Each national standards body within JTC1 had an opportunity to accept the standard, abstain or to “disapprove”, raising questions the resolution meeting then had to address. The reasons for disapproval and the comments from national bodies were projected on to a screen and the meeting worked through them. Many of the comments were proposed drafting changes which, after discussion, were either accepted or rejected. The greatest part of the discussion concerned proposed changes to the scope of the specific standard and proposed changes to the definitions of terms. Comments on definitions largely centered on ensuring clarity and on making sure the definitions were consistent with the definitions used elsewhere in ISO standards, so the discussion were dominated by participants with wide experience of other standards, both in their content and their drafting.

The discussion of proposed changes to the scope of the standard was more complex. The standard stated in its scope that it would be restricted to consumer applications: “An application is considered a consumer application if at least one of two interacting entities is a private individual (consumer) and the interaction is taking place in the public domain. Consequently, a Mobile RFID consumer application is defined as Mobile RFID equipment (e.g. mobile phones equipped with an RFID interrogator) being used in a consumer application.”

One national body proposed that this be changed to include the standard being used in enterprise applications in private spaces. In the discussion of the proposed change it was clear that the restriction to consumer uses had been because developers were focused on putting tag readers on consumers’ phones, but the discussion also clarified the boundary between WG6’s scope and activities within other WGs, where the development of tag technologies for enterprise applications were located. It was argued that nothing would prevent closed enterprise applications using

the standard once published. A consensus was reached that the scope was too narrow, but to change the scope of the standard required referral to JTC1 and a ballot of national bodies as it affected the area being standardized. Also it was agreed that the effect of the changes to the body of the standard represented a substantive change, which would also require re-balloting.

In this protracted discussion participants nominally representing their national bodies were not asked to comment on the submissions of those bodies and generally did not interject on behalf of their national bodies. Almost always comments were stated as first person opinions, for example “I am willing to...” or “I do not like...”. The only occasion when this was not true was during the discussion about whether the changes required a re-ballot when the US voting participant said “The US position is....”.

With the formal Ballot Resolution Meeting completed, the meeting then reconvened as WG6 considered the eight ongoing work items in turn, each of which was leading to a standard for future balloting. Each work item discussion was led by a Korean participant and it was clear that the active development within the WG was taking place within the Korean research organization. The discussions predominantly covered the content of the drafts circulated, with repeated discussions of the wording of the standard. The JTC1 directives follow ISO rules for the wording of standards. “Shall” indicates a requirement, “should” indicates a recommendation, “may” indicates permission and “can” indicates possibility. For the recently graduated Korean engineers leading the work items, this required a subtlety of English language usage that they had not been taught when learning English. One of the more experienced participants argued that one benefit of the meetings is developing the standards drafting skills of the participants in a non-confrontational environment. In the discussion of work items it was only briefly, when possible demonstrators of the standards were discussed, that the potential uses for the standards were mentioned.

During the evening between the first and second days, all participants attended a barbecue at the home of the WG convener. From informal discussions it was clear that the European and North American participants knew each other very well through other committees and standards bodies, that there was no separation into cliques and they were very open to include the Asian participants in their discussions. Several participants independently said that if it was not for the good social atmosphere in the group they would be far less willing to travel to the meetings. On the second day the liaisons with other SC31 working groups, external standards bodies and standards consortia were presented by the members responsible for the liaisons. These presentations focused on stating that the other body was aware of WG6’s activities and summarized what these bodies were doing that might impinge on WG6.

5. DISCUSSION: GOVERNANCE IN FORMAL STANDIZATION

In WG6 we see a formal structure that draws legitimacy from the institutions of formal standardization: the language of formal standardization, the procedures of JTC1 and the balloting of national standards bodies. But cutting through this formal structure is an informal, less institutionalized, structure, enrolling resources, including labor, existing intellectual property and expertise, and maintaining constructive links with other bodies. The German sociologist Ulrich Beck claims that a distinctive feature of current society is increasing reflexivity, with social processes focusing inwards on themselves and changes in the processes becoming an increasingly significant outcome of the processes themselves [24]. Early standards processes were focused on using participants’ expertise to develop a standard, but in comparison WG6 is a deeply reflexive organization, with a large part of its activity concerned with shaping the scope of its activities and negotiating its relationships with other bodies.

A further aspect of this reflexivity is the ambiguity about how the standards developed will be used. Each participant had their own ideas about how they might be used, but it was not necessary that these expectations are aligned and there is the risk that they will not be achieved. This ambiguity about use makes it difficult to imagine, following Tamm Hallstrom’s principles of organization [6], how WG6 can draw on user orientation. It is also difficult to see that national representation is a governing principle of the organization. The members of WG6 are nominated by national standards bodies and the ballot resolution dealt with issues raised by national bodies, but in both cases it was unclear whether, except for the involvement of the Korean research organization, any of the activity represented the articulation of a national interest. With the information technology sector highly globalized, many of the participants could have been nominated by more than one national standards body. It was argued by participants that anyone with relevant expertise interested in participating would be able to find a way to participate. Participants in the meeting were not blind to cultural heterogeneity, so the process cannot be seen as a flattening out of cultural differences or cultural hegemony. In the social setting of the barbecue national differences formed the main part of most discussions, but participants displayed for the two days what Beck has termed the “cosmopolitan outlook”: a “global sense, a sense of boundarylessness. An everyday, historically alert, reflexive awareness of ambivalences in a milieu of blurring differentiations and cultural contradictions” [25].

With this cosmopolitan outlook and the attenuation of the need to represent a national interest, the formal standards development process looks less to participants like a hierarchical process and becomes more like a global network of engaged experts. This can be seen as a convergence between the formal standards processes and the informal standards consortia. Castells claims that the

organization of the constitutive processes of society is changing from hierarchies to networks as a result of three trends: the growth of information and communication technology, a crisis of industrialism and the cultural challenge mounted by freedom movements [26]. The emergence of global standards development processes based on consensus, whether inside formal standards bodies or in consortia, fits with Castells' thesis. The explanation for the emergence of this reflexive mode of standardization therefore moves from claiming that the new processes are more efficient or more effective than conventional processes, which has been a methodologically difficult claim to justify, to considering how these bodies are shaped by achieving legitimacy and are seen as the right way to develop standards.

6. CONCLUSION

This paper sought to use a case study of a formal standardization process to assess how the processes of formal standardization are responding to the rise of consortia and the globalization of technology development. The clearest trend observed is the opening up of the process to a wider range of actors. However, this openness is seen to weaken participants' identification with the national interest of the country they nominally represent. The need to ballot national standards bodies creates delays in reaching agreements. The objections of national bodies are interpreted by participants in the working group along a continuum from helpful identification of overlooked issues, through misunderstanding about what is proposed, to anonymous strategic obstruction. Participants in the global process believed that individuals raising issues in national bodies should be active directly in the working group.

The case has shown that reflexively maintaining the legitimacy of a standards process has become a more significant activity within it in a globalized world of potentially overlapping standards processes. This is achieved by being open to all potential participants and by negotiating relations with other bodies. Being part of the formal standards development systems provides a further source of legitimacy, but also limits the freedom for the process to replicate the structures of consortia. The structure of committees of multiple national standards bodies feeding national requirements into the global processes by responding to ballots resolutions and nominating representatives is increasingly anachronistic in a world of global communications, more open standards development, global technology companies and the weakening of the ability of states to identify a national interest in technology policy.

REFERENCES

- [1] R. Werle, and E. Iversen, "Promoting Legitimacy in Technical Standardization," *Science, Technology and Innovation Studies*, vol.2, pp 19-39, 2006.
- [2] W. Bellchambers, J. Francis, E. Hummel and R. Nickelson, R. "The international telecommunication union and development of worldwide telecommunications," *Communications Magazine, IEEE*. 22(5): 72 - 82. 1984.
- [3] C. D. Woodward, *The Story of Standards*, BSI, London. 1972.
- [4] K. Tamm Hallstrom, Organizing the Process of Standardization. In N. Brunsson & B. Jacobsson (Eds.), *A World of Standards*: 85-99. Oxford: Oxford University Press. 2000.
- [5] S. K. Schmidt, and R. Werle., *Co-ordinating technology: Studies in the international standardization of telecommunication*. Cambridge: MIT Press. 1998.
- [6] P. A. David and S. Greenstein, S. "The economics of compatibility standards: an introduction to recent research." *Economics of Innovation and New Technology*, 1(1-2): 3-41. 1990.
- [7] W. Kuert, The Founding of ISO, in *Friendship Among Equals*, ISO. Geneva, p 13-21. 1997.
- [8] P. A. David, and M. Shurmer, "Formal standards-setting for global telecommunications and information services. Towards an institutional regime transformation?" *Telecommunications Policy*, 20(10): 789-815. 1996.
- [9] R. Hawkins, "The rise of consortia in the information and communication technology industries: emerging implications for policy." *Telecommunications Policy*, 23(2): 159-173.1999.
- [10] M. Weiss, and C. Cargill, Consortia in the Standards Development Process, *Journal of the American Society for Information Science*, Vol. 43, Iss. 8, pp 559-569. 1999.
- [11] J. Tate, National Varieties of Standardization. in *Varieties of Capitalism*, Peter A. Hall and David Soskice(eds). Oxford, Oxford University Press. 2001.
- [12] W. Mattli, "The Politics and Economics of International Institutional Standards Setting: an introduction." *Journal of European Public Policy* 8:3, pp. 328-344. 2001.
- [13] M. J. Bonino, and M. B. Spring, "Standards as change agents in the information technology market." *Computer Standards & Interfaces*, 12(2): 97-107. 1991.
- [14] C. Shapiro, and H. R. Varian, *Information Rules*. Boston: Harvard Business Press. 1999.
- [15] P. A. David, "Clio and the Economics of QWERTY." *The American Economic Review*, 75(2): 332-337. 1985.
- [16] K. Blind, *The Economics of Standards: Theory, Policy, Evidence*, Edward Elgar, Cheltenham. 2004.
- [17] P. Swann, P. 'User Needs for Standards: How Can We Ensure that User Votes are Counted?' in B. Meek et al (eds.) *User Needs in Information Technology Standards*, Butterworth/Heinemann. 1993.
- [18] J. Farrell, and G. Saloner, "Coordination Through Committees and Markets". *The RAND Journal of Economics*, 19(2): 235-252. 1988.

- [19] P. J. DiMaggio, and W. W. Powell, "The Iron Cage Revisited: Institutional Isomorphism and Collective Rationality in Organizational Fields." *American Sociological Review*, 48(2): 147-160. 1983.
- [20] R. Werle, "Institutional aspects of standardization - jurisdictional conflicts and the choice of standardization organizations". *Journal of European Public Policy*, 8(3): 392-410. 2001.
- [21] S. M. Besen, and J. Farrell, J.. "Choosing How to Compete: Strategies and Tactics in Standardization." *The Journal of Economic Perspectives*, 8(2): 117-131.1994.
- [22] P. Grindley, *Standards, Strategy, and Policy*, Oxford: Oxford University Press. 1995.
- [23] K. Jakobs, R. Procter, and R. Williams, The making of standards: looking inside the work groups, *Communications Magazine*, IEEE. 39(4): 102 – 107. 2001.
- [24] U. Beck, "The Reinvention of Politics: Towards a Theory of Reflexive Modernization" in U. Beck, A. Giddens, S. Lash *Reflexive Modernization: Tradition and Aesthetics in the Modern Social Order*, Polity Press. 1994.
- [25] U. Beck, *Cosmopolitan Vision*, Polity Press. 2006.
- [26] M. Castells, *The information age: society and culture: volume1 - the rise of the network society*, Blackwell.1996.

SESSION 4

FREQUENCY AND SPECTRUM MANAGEMENT

- S4.1 Radio Resource Management in OFDMA-CRN Considering Primary User Activity and Detection Scenario
- S4.2 Optimal Pilot Patterns Considering Optimal Power Loading for Cognitive Radios in the Two Dimensional Scenario
- S4.3 Optimal Spectrum Hole Selection & Exploitation in Cognitive Radio Networks

RADIO RESOURCE MANAGEMENT IN OFDMA-CRN CONSIDERING PRIMARY USER ACTIVITY AND DETECTION SCENARIO

Dhananjay Kumar* and Kanagaraj N. N.#

*Department of Information Technology, Anna University, MIT Campus, Chromepet, Chennai, India
Email: dhananjay@annauniv.edu

#Alcatel-Lucent India Limited, Industrial Estate, Guindy, Chennai, India
Email: kanagaraj.nn@alcatel-lucent.com

ABSTRACT

In this paper an adaptive radio resource allocation scheme is developed for OFDMA based cognitive radio network (OFDMA-CRN), which not only considers the dynamic nature of primary users but also includes the detection scenario of unused licensed spectrum. In contrast to the existing research for OFDMA-CR systems which consider either of these issues independently, our approach tackles both of these concerns jointly and finds optimal solution. The proposed sub-carrier and power allocation (SPA) algorithm optimally selects and computes optimal power loading for each sub-carrier thereby increases sum data rate and overall throughput of the system.

Keywords— OFDMA, Cognitive radio, Resource allocation, Rate loss, Detection scenario

1. INTRODUCTION

The high data rate requirement for many internet services in wireless network dictates the development of an intelligent radio resource management technique. The cognitive wireless system is an advanced wireless communications networks that can provide mechanisms for more efficient use of the spectrum through dynamic spectrum access techniques [1].

The ITU-R Working Party 1B defines the cognitive radio system as “a radio system employing technology that allows the system: to obtain knowledge of its operational and geographical environment, established policies and its internal state; to dynamically and autonomously adjust its operational parameters and protocols according to its obtained knowledge in order to achieve predefined objectives; and to learn from the results obtained” [2]. This definition not only outlines the objectives of cognitive radio systems but also justifies its application in radio resource management of future wireless networks.

An important and challenging aspect in radio resource management of CRN is to locate and estimate free bands called “spectrum hole” [3]. To understand and assess spectrum hole, some real-time measurement of idle spectrum in popular bands were conducted by us in business district (Adyar) of Chennai. As can be seen in

Fig.1, in the band of 902-928 MHz many spectrum holes are noticeable. With same experimental set-up measurements were carried out in other bands (e.g., 2.4 GHz), and many vacant spectrum were observed (Fig.2). Although these spectrum usage patterns were dynamic in nature, it provides enough scope for the development of prudent radio resource management system.

The multi-carrier systems can further help in optimizing the usage of spectrum as it allows placing carriers in non-contiguous way. Because of its flexibility and improved performance over other multi-carrier access techniques, OFDMA has become the de facto standard for the next generation wireless networks [4-5]. The OFDMA based cognitive radio (OFDMA-CR) becomes an intelligent choice to support the high data rate requirement of many internet services. Applying CR techniques in OFDMA can alleviate the looming spectrum shortage crisis [6] as any idle spectrum can easily be accommodated in radio resource pool. However, it is a challenging due to the requirement of easy coexistence of both primary (licensed) and secondary (unlicensed) users as well as the wide range of available radio spectrum [7-10].

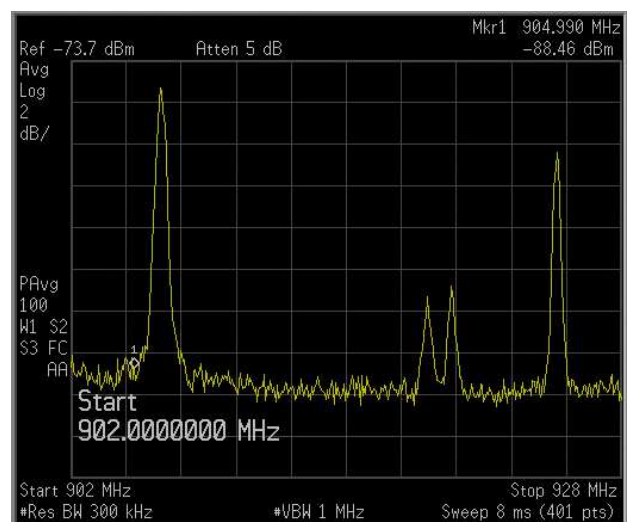


Figure 1. Spectrum uses in 900 MHz band at Adyar, Chennai

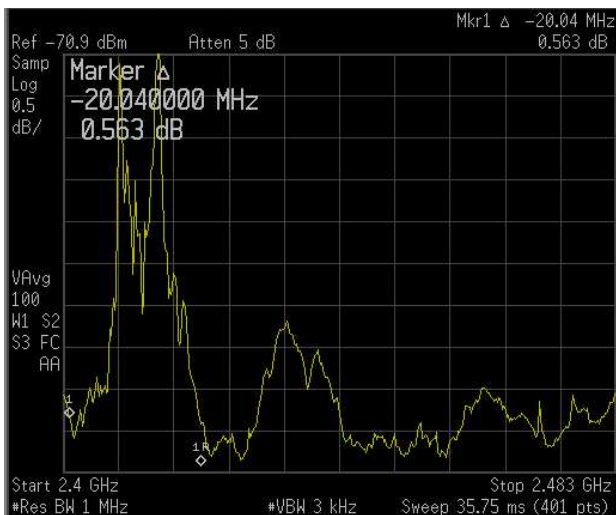


Figure 2. Spectrum uses in 2.4 GHz band at Adyar, Chennai

In our earlier research work [11], to support different types of real-time applications in OFDMA-CRN, an application specific parameter called “adaptability” was defined to signify the dynamic resource requirement of user defined services. The joint sub-carrier and power allocation algorithm maximised the system capacity by dynamically sharing the cognitive spectrum among different secondary users. This concept is prolonged here to include primary user’s activity and detection scenario. Because of change in objective function, instead of using conventional water filling approach, the optimal condition for power allocation is derived here.

The proposed SPA algorithm maximises the system throughput by making use of cognitive (primary user’s) spectrum which is dynamically shared among different secondary group of users. Each user group is characterized by its minimum data rate requirement. SPA contains two major functions: sub-carrier selection and power allocation. The power allocation process is based on optimally derived condition in a typical cognitive environment. The loss in data rate because of return of primary user is modelled as a rate loss function. The detection condition is embedded in objective function to accommodate the loss of channel data rate of secondary user.

2. RELATED WORK

In conventional power allocation methods (e.g., water-filling) high power is assigned onto the sub-carrier with the higher channel gain. In case of cognitive radio network (CRN), the amount of interference produced by allowing communication on a CR user’s sub-carrier is subjected to the location of the sub-carrier with respect to the primary user (PU) working band. Hence, it clear that more power should be allocated to a sub-carrier which is far away from PU’s band [12]. Furthermore, for example, in the uplink channel of OFDMA systems, the additional power constraints for sub-carriers bring new challenges to the traditional resource allocation methods. Many times it

becomes difficult to solve the problem directly but using continuous relaxation, the problem can be converted into a concave maximization problem [13-14].

In both uplink and downlink sub-carrier, the resource allocation problem is a mixed-integer nonlinear programming problem, for which obtaining the optimal solution is known to be *NP*-hard [15]. This dictates the development of computationally efficient sub-optimal algorithms as a common approach. Furthermore, in case of OFDMA based cognitive radio network, as a result of imperfect spectrum sensing, the cognitive radio network might identify certain sub-bands as vacant when, in reality, they are employed by the primary user. This type of false detection will cause lost opportunities for communication and, therefore, will decrease the overall throughput of the network [15].

Modeling a true primary user’s behavior in CRN is difficult as it will require high level of probability theory. A cost function is used to represent the rate loss whenever primary user returns to the sub-channel. The approach in [16], models the system to allocate more power to the sub-carriers that are available fairly more often than the one which gets busy very quickly. Such an activity is noticeable when there is more than one sub-channel that belongs to different primary users. Of course, primary users will vary based on their activity/usage of their licensed band.

The original non-convex optimization problem can be solved in the dual domain with the global optimum obtained when the number of sub-carriers approaches to infinity [17]. Considering the dynamic spectrum usage patterns of primary user, the study in [17] focuses on resource allocation in a secondary OFDMA-based multicast network, which is based on risk-return model in [16]. This work aims the maximization of the expected sum rate of all secondary users while sustaining the bearable interference level at each primary user.

Based on the loss function conception in [17] and [18], our approach includes the problem of detection in sensing of the spectrum hole by secondary user. Although we consider linear model of loss function for simplicity, the goal is to study the combined effect of error in detection and activities of primary user in CRN. Furthermore, as in our earlier study [11], the inclusion of an adaptability factor in resource allocation problem in OFDMA based cognitive radio system to different types of applications remains valid here too.

3. SYSTEM MODEL

The resource requirement of primary users is protected by enforcing the cognitive users to send their data only in unused spectrum called “spectrum hole”. Furthermore, cognitive users need to switch over to their own band while leaving the licensed band whenever primary users return. The system model includes signal representation, channel sensing, and formulation of resource allocation methods. The resource allocation model is defined by deriving necessary conditions in a typical CRN.

3.1. Signal Representation

Assuming uniform sampling, the signal received at the primary and secondary user can be represented using discrete time notation as [19]

$$r_{pu}[n] = h_{pu}[n]s_{pu}[n] + \sum_{k=1}^{K_{su}} h_{p,k}[n]s_k[n] + \sigma_{pu}n_{pu}[n] \quad (1)$$

$$r_{su}[n] = h_{su}[n]s_{su}[n] + \sum_{k=1}^{K_{su}} h_{s,k}[n]s_k[n] + \sigma_{su}n_{su}[n] \quad (2)$$

where $h_{p,k}$ and $h_{s,k}$ are the channel from k^{th} transmitter to the PU and SU receiver respectively. $n_{pu}[n]$ and $n_{su}[n]$ are independent identically distributed (i.i.d.) Gaussian noise normalized to 1 so that σ_{pu}^2 and σ_{su}^2 represents the noise levels at the primary and secondary users. $s_k[n]$ denotes the sample of the signal transmitted by the k^{th} user. h_{pu} / h_{su} and S_{pu} / S_{su} are the channel response and desired signal of the primary/secondary user respectively.

3.2. Channel Sensing

Assuming that, the first N seconds of the frame duration T is used to sense the channel. Since there are NB complex symbol over a period of N seconds, mathematically it can be denoted as [20]

$$H_0 : y(i) = n(i), \quad i = 1, 2, \dots, NB \quad (3)$$

$$H_1 : y(i) = s(i) + n(i), \quad i = 1, 2, \dots, NB \quad (4)$$

where $s(i)$ and $n(i)$ are the signal and noise respectively.

The probability of false alarm and detection can be expressed as follows

$$P_f = 1 - P\left(\frac{NB\lambda}{\sigma_n^2}, NB\right) \quad (5)$$

$$P_d = 1 - P\left(\frac{NB\lambda}{\sigma_n^2 + \sigma_s^2}, NB\right) \quad (6)$$

where λ is the detection threshold and $P(x, a)$ is the regularized lower Gamma function defined as

$$P(x, a) = \frac{\gamma(x, a)}{\tau(a)} \quad (7)$$

where $\gamma(x, a)$ represents the lower incomplete gamma function and $\tau(a)$ is the Gamma function.

3.3. Resource Allocation Model

We consider a CR system with G number of groups employing K number of sub-carriers. Furthermore, it is assumed that there are M_g number of users per group with each user employing a transmit power of $P_{g,k}$ in the channel with signal-to-interfering noise (SINR) of $\gamma_{g,k}$.

If β_k be the linear coefficient associated with the detection error probability, Φ_k related with the probability that the primary user returns and occupies its sub-carriers, and α_g

the adaptability of a g^{th} group, the objective function is formulated as

$$\max_{P_{g,k}} \sum_{g=1}^G \sum_{k=1}^K \frac{\alpha_g |M_g| \log_2(1 + \gamma_{g,k} P_{g,k})}{K} - \phi_k L(P_{g,k}) - \beta_k F(P_{g,k}) \quad (8)$$

Subjected to

$$(i) \sum_{g=1}^G \sum_{k=1}^K P_{g,k} I_k \leq I_{th} \quad (9)$$

$$(ii) P_{g,k} \geq 0, g = 1, 2, \dots, G \quad k = 1, 2, \dots, K \quad (10)$$

$$(iii) P_{g,k} P_{g',k} = 0 \quad \forall g' \neq g \quad (11)$$

$$(iv) \sum_{g=1}^G \alpha_g = 1 \quad (12)$$

In this formulation, weight $\alpha_g \geq 0$ reflects the adaptability of application in a particular group. A high value of α indicates that the user application belongs to a type of non-real time category that can easily adapt to a large change in available channel data rate. Similarly a low value of α indicates a real-time service which may require a judicial number of sub-carriers from the primary band as a matter of strategy. In condition (i), I_{th} indicates interference threshold and I_k represents the interference caused by sub-carrier k to the primary user.

The rate loss function $L(P_{g,k})$ is assumed to be linear and detection function $F(P_{g,k})$ is governed by (6) and (7). $L(P_{g,k})$ could be expressed as [16]

$$L(P_{g,k}) = C \cdot P \quad (13)$$

The constant C is normalized average cost per unit power for the secondary user to allocate resources. The loss function in (13) could be non-linear too [17, 18].

The overall objective defined in (8) can be assumed to be optimization of K independent functions:

$$D_k(\lambda) = \sum_{g=1}^G \left\{ \frac{\alpha_g |M_g| \log_2(1 + \gamma_{g,k} P_{g,k})}{K} - \left(\sum_{n=1}^N \lambda_n I_k^{(n)} P_{g,k} + \phi_k L(P_{g,k}) + \beta_k F(P_{g,k}) \right) \right\} \quad (14)$$

Applying Karush–Kuhn–Tucker (KKT) conditions for optimal power allocation,

$$\nabla_{P_{g,k}} D_k(\lambda) = \frac{\alpha_g |M_g| \gamma_{g,k}}{K(1 + \gamma_{g,k} P_{g,k}^*) \log_2} - \sum_{n=1}^N \lambda_n I_k^{(n)} - \phi_k C_1 - \beta_k C_2 = 0 \quad (15)$$

The KKT condition in (15) provides a basis for a closed-form solution, and $P_{g,k}^*$ can be derived as follows

$$P_{g,k}^* = \frac{\alpha_g |M_g|}{K[\sum_{n=1}^N \lambda_n I_k + \phi_k C_1 + \beta_k C_2] \log_2} - \frac{1}{\gamma_{g,k}} \quad (16)$$

Now, the optimal result in (14) can be obtained using (16) as

$$D_k^*(\lambda) = \max_g \left\{ \frac{\alpha_g |M_g| \log_2(1 + \gamma_{g,k} P_{g,k})}{K} - \left(\sum_{n=1}^N \lambda_n I_k P_{g,k}^* + \phi k L P_{g,k} + \beta k F P_{g,k} \right) \right\} \quad (17)$$

The λ is updated in step-size sequence [17] given by

$$\lambda_n^{(t+1)} = \left(\lambda_n^{(t)} - \delta^{(t)} \left(I_{th}^{(n)} - \sum_{g=1}^G \sum_{k=1}^K P_{g,k} I_k^{(n)} \right) \right)^+ \quad (18)$$

δ is chosen sufficiently small to converge to the optimum value of λ .

4. DEVELOPMENT OF SPA

The development of SPA algorithm results in two scenarios: (i) assignment of the same value of α_g to each group and trying to meet their minimum bit rate requirement, and (ii) assignment of different α_g to each group to represent their quality of service (QoS) profile.

Since multimedia and internet users constitute large varieties of applications, based on their QoS profile users can be grouped into different categories. This step not only helps in resource allocation but also in channel management. Furthermore, as stated before, it provides scope for the SPA algorithm to simulate in two broader scenarios.

Computation of λ according to (18) needs seed value, and selection of it has to make sure that optimum power allocation procedure always converges. The SPA algorithm starts with small value of λ , and iteratively proceeds for sub-carrier selection and optimum power allocation (Fig.3).

5. SIMULATION RESULTS

We consider a multi-user OFDM-based CR system with uniform distribution of users. The Rayleigh channel model was used to represent link between a secondary user and the base station. Table-I lists the main system parameters considered in the simulation.

Table 1. System Parameters

Parameters	Description	Value
N	Number of primary users	2
K	Number of OFDM sub-carriers	128
B	Maximum spectrum hole	10MHz
G	Number of groups	10
$ M_g $	Number of secondary users in each group g	4

First, the study of effect of loss and detection parameters is presented here. Although these two parameters are tightly defined and embedded in system model, the simulation results and analysis provide an insight towards improvement of the sum data rate of the CRN. A judicious selection of these parameters will result in enhanced the throughput of the system.

Two scenarios are created to simulate the SPA algorithm. Scenario1 assumes similar user’s applications in each group

and system tries to meet their minimum bit rate requirement whereas Scenario2 deals with heterogeneous requirement of different users in each group represented by their QoS profile. These two scenarios represent quite different perspective of user’s applications. For example, Scenario1 could represent a case where all users are interested in best effort kind of service or a non-real-time service. Similarly Scenario2 can characterize groups of users employing real-time services which are in need of maintaining its severe QoS requirements.

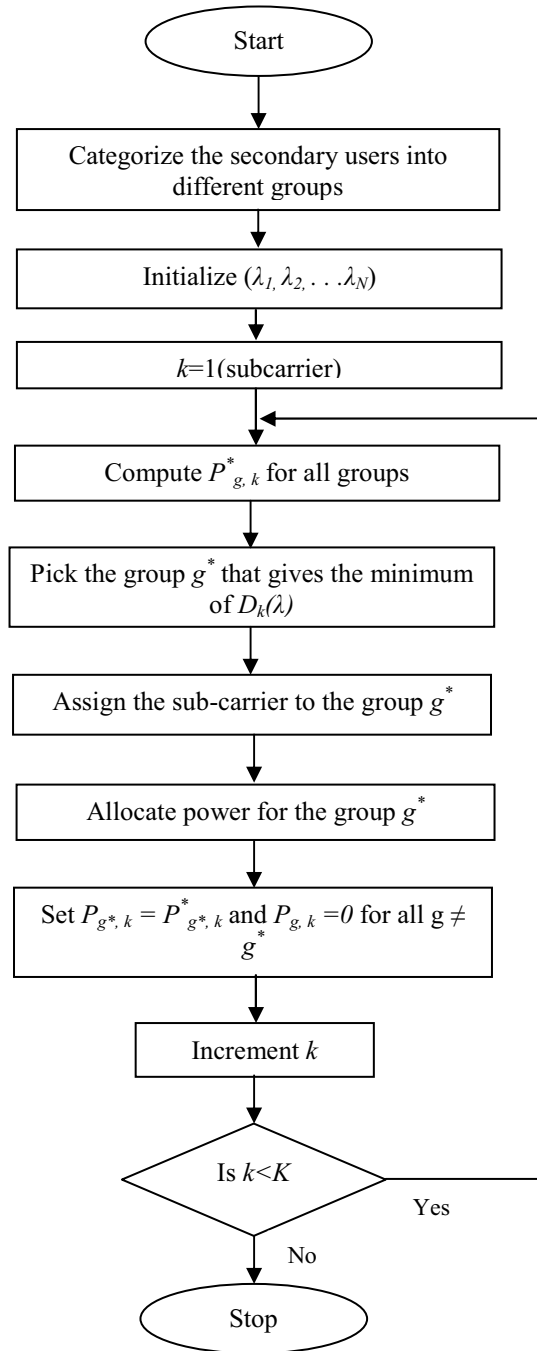


Figure 3. The SPA algorithm

Fig.4 shows the impact of data rate due to the effects of varying Φ_k (probability that the primary user comes back and gets its sub-carriers back) and C_l (rate loss constant). When Φ_k varies from 0.02 to 0.1, C_l is allowed to take any value between 0.0 to 1.0. $C_l = 0$ represents no activity whereas $C_l = 1$ indicates high activity i.e. channel fully occupied by the primary. To observe the near worst case scenario, the coefficient associated with detection error probability was fixed to a low value ($\beta_k = 0.03$). The decrease in data rate (Fig.4) is dictated by increased user activity.

Assuming the probability that the sub-carrier k is taken back by primary user, $\Phi_k = 0.04$, Fig.5 shows the effect of incorrect detection of spectrum by secondary user on its normalized data rate. The inaccurate detection of spectrum not only decreases the throughput of secondary user but also to primary user. The lower value of $\Phi_k = 0.04$ was chosen to study the dominant effect of incorrect detection; otherwise higher value of Φ_k could shadow this investigation.

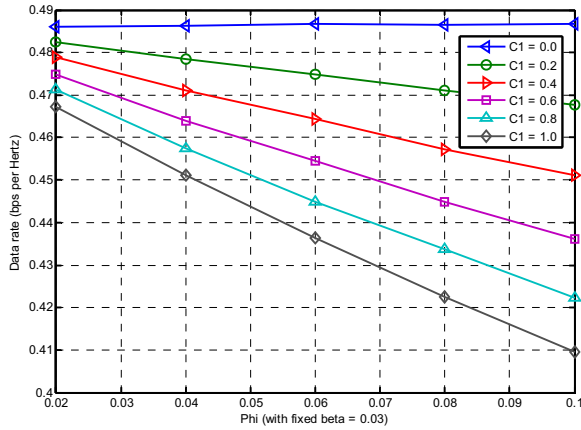


Figure 4. Effects of varying loss parameters

As discussed earlier the SPA algorithm is realized in two scenarios, where each scenario not only differs in assignment of adaptation parameters but also in meeting the minimum rate requirements and QoS. In Fig.6, the simulated results show that the data rate is higher in allocation Scenario1 than Scenario2 with the increased number of group of secondary users. This is attributed to the fact that allocation Scenario1 just tries to meet the minimum data rate requirement for each user. The selection of error detection probability (β_k) and primary user activity (Φ_k) directly affects the sum rate and a higher values of either of these parameters will result in large degradation sum rate and hence the throughput.

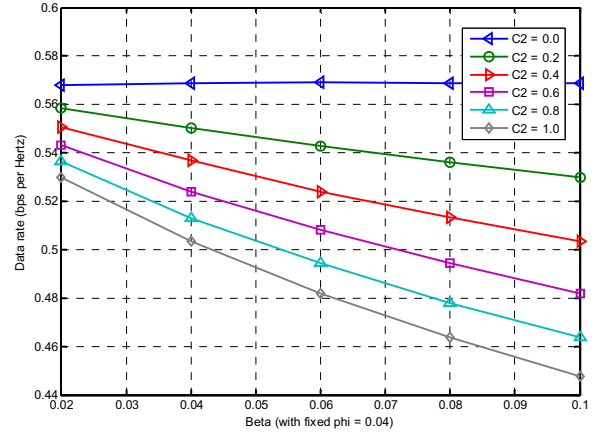


Figure 5. Effects of variation in detection parameters

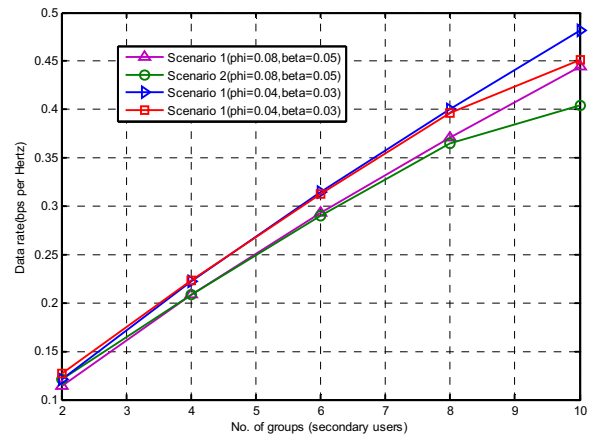


Figure 6. Sum data rates in two proposed scenarios

Fig.7 shows that the throughput is higher for allocation Scenario1 than allocation Scenario2 with the increased number of group of secondary users. The decrease in throughput in Scenario2 at higher number of user group is attributed to the large computation involve at base station in resource allocation as each group has different QoS profile. Although the computational delay depends on many parameters including the processor architecture and memory management, here the main intention is to study the algorithmic processing time that causes packet delay. The simulation Scenario1 assumes same QoS profile for all users, but in Scenario2 because of different user QoS requirement in different group, the computational load on processing node increases. For smaller number of group and users, processing delay for both scenarios is almost the same; only when the number of group containing fixed number of users increases above a threshold value, a different delay profile is noticeable (Fig.8).

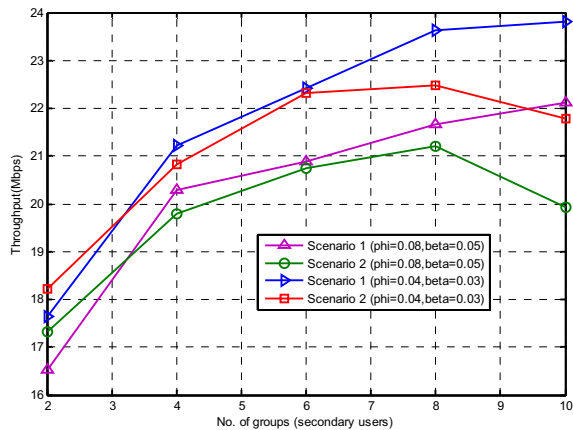


Figure 7. Throughput comparison

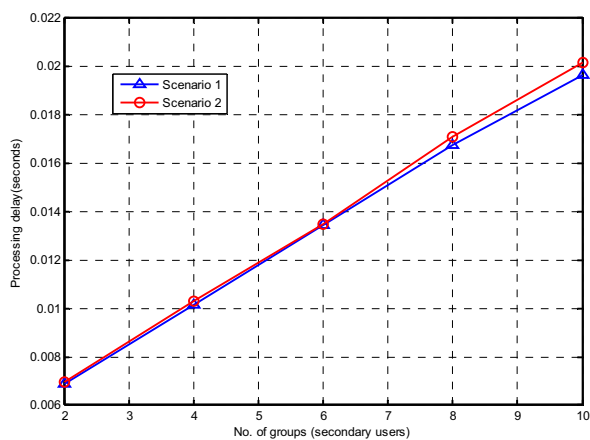


Figure 8. Assessment of processing delay

6. CONCLUSION

We formulated and derived the condition for optimal power and sub-carrier allocation problem in OFDMA-CRN. The rate loss due to the unpredictable activity of primary user and error in sensing spectrum was included in objective function. The effect of presence of both issues i.e. primary user activity and detection was simulated as a function of normalized data rate of secondary user. Although resource allocation in Scenario1 provides better sum rate & throughput, Scenario2 represents a more generic user pattern in a cognitive radio system. The optimized resource allocation strategy formulated here for the SPA algorithm results in enhanced sum data rate and hence better throughput at higher number of secondary users.

The throughput of primary user can be protected by introducing a suitable mechanism in SPA algorithm, which could be the future work of this paper.

REFERENCES

- [1] James A. Hoffmeyer, "IEEE 1900 and ITU-R Standardization Activities in Advanced Radio Systems and Spectrum Management," 4th IEEE Consumer Communications and Networking Conference, 2007 (CCNC 2007), IEEEExplore, pp.1159 – 1163.
- [2] ITU-R SM.2152, "Definitions of Software Defined Radio (SDR) and Cognitive Radio System (CRS)," September 2009.
- [3] Peter Steenkiste, Douglas Sicker, Gary Minden, and Dipankar Raychaudhuri, "Future Directions in Cognitive Radio Network Research," NSF Workshop Report, March 9-10, 2009.
- [4] Megumi Kaneko, Petar Popovski, and Kazunori Hayashi, "Throughput-Guaranteed Resource-Allocation Algorithms for Relay-Aided Cellular OFDMA System," *IEEE Transaction on Vehicular Technology*, Vol.58, No.4, May 2009, pp. 1951-1964.
- [5] Kumar, Dhananjay. Srividhya, S. Mariappan, P. Martheeswaran, M. Chellappan, C., "Dynamic resource management for downlink multimedia traffic in OFDMA cellular networks," ITU-T Kaleidoscope Event: Innovations for Digital Inclusions, (K-IDI 2009) Aug. 31 - Sept. 1, 2009.
- [6] Rui Wang, Vincent K. N. Lau, Linjun Lv, and Bin Chen, "Joint Cross-Layer Scheduling and Spectrum Sensing for OFDMA Cognitive Radio Systems," *IEEE Transaction on Wireless Communications*, Vol. 8, No. 5, May 2009, pp. 2410-2416.
- [7] Dusit Niyato, and Ekram Hossain, "Competitive Spectrum Sharing in Cognitive Radio Networks: A Dynamic Game Approach," *IEEE Transaction on Wireless Communications*, Vol. 7, No. 7, July 2008, pp. 2651-2660.
- [8] Zhiqiang Li, F. Richard Yu, and Minyi Huang, "A Distributed Consensus-Based Cooperative Spectrum-Sensing Scheme in Cognitive Radios," *IEEE Transaction on Vehicular Technology*, Vol. 59, No.1, January 2010, pp. 384-392.
- [9] Nick C. Theis, Ryan W. Thomas, and Luiz A. DaSilva, "Rendezvous for Cognitive Radios," *IEEE transaction on Mobile Computing*, Vol. 10, No. 2, February 2011, pp. 216-227.
- [10] Lifeng Lai, Hesham El Gamal, Hai Jiang, and H. Vincent Poor, "Cognitive Medium Access: Exploration, Exploitation, and Competition," *IEEE transaction on Mobile Computing*, Vol. 10, No. 2, February 2011, pp. 239-253.
- [11] Dhananjay Kumar, S. Mahalaxmi, J. Sharad Kumar, and R. Ramya, "Adaptive Resource Allocation for Real-time services in OFDMA Based Cognitive Radio Systems," Kaleidoscope: Beyond the Internet? - Innovations for Future Networks and Services, 2010 ITU-T, 13-15 December 2010.
- [12] G. Bansal, M. Hossain, and V. Bhargava, "Optimal and suboptimal power allocation schemes for OFDM-based cognitive radio systems," *IEEE Transaction on Wireless Communication*, vol. 7, no. 11, November 2008, pp. 4710–4718.
- [13] C. Y. Wong, R. S. Cheng, K. B. Lataief, and R. D. Murch, "Multiuser OFDM with adaptive subcarrier, bit, and power allocation," *IEEE Journal on Selected Areas in Communication*, vol. 17, 1999 pp. 1747-1758.
- [14] Zhihua Tang and Guo Wei, "An Efficient Subcarrier and Power Allocation Algorithm for Uplink OFDMA-based

- Cognitive Radio Systems,” IEEE Wireless Communications and Networking Conference (WCNC), 2009, pp.1-6.
- [15] Sami M. Almalfouh, and Gordon L. Stüber, “Interference-Aware Radio Resource Allocation in OFDMA-Based Cognitive Radio Networks,” IEEE transaction on Vehicular Technology, Vol. 60, No. 4, May 2011, pp. 1699-1713.
- [16] Z. Hasan, E. Hossain, C. Despins, and V. K. Bhargava, “Power allocation for cognitive radios based on primary user activity in an OFDM system,” IEEE GLOBECOM, Dec. 2008, pp. 1–6.
- [17] Duy T. Ngo, Chintha Tellambura, and Ha H. Nguyen, “Resource Allocation for OFDMA-Based Cognitive Radio Multicast Networks With Primary User Activity Consideration”, *IEEE Transaction on Vehicular Technology*, Vol. 59, No. 4, May 2010, pp.1668-1679.
- [18] V. Vapnik, “An overview of statistical learning theory,” *IEEE Transaction on Neural Network.*, vol. 10, no. 5, September 1999, pp. 988–999.
- [19] Gonzalo Vazquez-Vilar, Carlos Mosquera, and Sudharman K. Jayaweera, “Primary User Enters the Game: Performance of Dynamic Spectrum Leasing in Cognitive Radio Networks,” *IEEE Transaction on Wireless Communications Technology*, Vol. 9, No. 12, December 2010, pp.3625-3629.
- [20] Sami Akin and Mustafa CenkGursoy, “Effective Capacity Analysis of Cognitive Radio Channel for Quality of Service Provisioning,” *IEEE Transaction on Wireless Communications Technology*, Vol. 9, No. 11, November 2010, pp.3354-3363.

OPTIMAL PILOT PATTERNS CONSIDERING OPTIMAL POWER LOADING FOR COGNITIVE RADIOS IN THE TWO DIMENSIONAL SCENARIO

Boyan Soubachov, Neco Ventura

University of Cape Town, Cape Town, South Africa

ABSTRACT

In Orthogonal Frequency Division Multiple Access (OFDMA) based Cognitive Radio (CR) systems, optimal power loading schemes are devised such that the transmission rate is maximized while maintaining interference from Secondary Users (SUs) to Primary Users (PUs) below a specified threshold. The power loading algorithms however do not distinguish between data and pilot symbols. A similar situation exists for optimal pilot patterns where pilots are placed in the optimal positions to achieve the lowest Mean Squared Error (MSE) but no consideration is given to power loading schemes. This paper investigates this scenario and proposes an optimal solution based on the Least Squares (LS) estimator which could be applied to future CR algorithm implementations as well as a possible standardization aspect to ensure an optimal solution to a crucial problem in practical implementations.

Keywords— Power Loading, OFDMA, Cognitive Radio, Pilot Pattern, Least Squares

1. INTRODUCTION

Spectrum shortage has become an omnipresent problem in the telecommunications world due to the immense increases in bandwidth and data rates experienced over the last decades. The need for spectrum in order to provide higher data rates demanded by newer consumer trends has led to fierce competition in terms of spectrum allocation and auctioning while at the same time barring smaller players from entering the market.

A proposed solution to the issue of spectrum crowding is cognitive radio. A cognitive radio system is that of an intelligent, software defined radio (SDR) where the system adjusts its modulation parameters such that it uses licensed frequency bands when they are not used by their licensed users [1]. This allows for a no-interference approach and could be used to achieve optimal spectral usage while being effectively invisible to legacy devices.

While licensed spectrum may only be used by the licensed or primary users, studies conducted have found that the actual, temporal utilisation of licensed spectrum may range anything from 15% to 85% [2]. In certain environments and locations, these figures can be significantly lower. A

prime example of this is in suburban areas where frequency utilization from 100 MHz to 3 GHz can be as low as 7% of the time [3]. This means that the licensed users with exclusive use to their spectrum do not fully utilize their spectrum at all times.

As such, cognitive radio has been proposed as a promising, viable solution to the problem of spectrum crowding and under-utilization. Many proposed implementations involve the use of an OFDM variant known as non-contiguous OFDM (NC-OFDM). This modulation scheme proposes that like OFDMA, different sub-channels are assigned to different users but where they differ is that NC-OFDM requires for sub-channels which interfere to PUs disabled. This then allows the CR systems to conduct transmissions arbitrarily close to the PUs' transmissions while not interfering with the PUs' transmissions themselves. This is critical to the premise of CRs as there must be an interference-free approach to using the PUs' bandwidth.

To successfully implement these principles, two areas of focus have been examined, namely pilot patterns and power loading.

Optimal power loading algorithms for CRs are devised such that the transmission rate for the CR system is maximized (proportional to the power assigned per sub-channel) while interference to PU systems nearby in the frequency spectrum is kept below a specified threshold in order to comply with CR principles. This algorithm, developed in [4], has been shown to be the optimal way to load power per sub-channel in an NC-OFDM transmission such that the optimal data rate is achieved while maintaining interference to PUs below a certain threshold. As such, it is found that the sub-channels closer to the PU, should have less power loaded onto them than the sub-channels farther away from the PU. This can be attributed to the non-ideal characteristics of pulse-shape filtering and windowing methods having non-zero spectral roll-off and as such causing energy leakage outside of their band.

An investigation was also conducted into the aspect of optimal pilot patterns. As the PU transmission bandwidths cannot be known to the SU, a wide- or narrowband PU could appear in the SU's spectrum at any time. This means that some pilots which were used previously in order to estimate the channel impulse response (CIR) and channel frequency response (CFR) would need to be disabled in

order not to interfere with the PU's transmission. As the PU's transmission could take up any amount of sub-channels, this could mean that several pilot sub-channels would need to be disabled, greatly reducing estimation and interpolation accuracy. This would mean an increase in the estimator MSE which would translate to a reduction in the maximum possible data rate.

It has been found that the optimal way to position pilots in CR transmissions is to convert the sub-channels adjacent to PU transmissions to pilot-bearing sub-channels. This allows a decrease in estimator MSE for the sacrifice of two data-bearing sub-channels (one on each side of a PU's transmission) [7].

If one were to consider these two aspects, they are not mutually independent of each other since it is necessary for the pilot-pattern of the system to adapt to changes in the utilised spectrum (such as intermittently appearing and disappearing PUs). When also factoring the criterion for interference to the PU, this would lead the implementation into placing pilots in the sub-channels closest to the PU while reducing the power of those sub-channels significantly so as not to cause any interference to the PU.

It was thus found that a constrained optimisation problem could be used to describe and solve the issue of considering optimal power loading levels while attempting to minimize the MSE by placing pilots close to the PU.

This problem can be further exacerbated by the fact that in many practical applications, the pilot-to-data power ratio (PDPR) is greater than unity such that pilot symbols are assigned more power than data symbols. This allows for a higher SNR for the pilot symbols and in turn provides a more accurate estimate due to there being less noise in the gain measurements at pilot positions.

Optimal pilot placement therefore needs to be found such that estimator MSE is minimized while optimal power loading is applied such that interference to the PU does not exceed a threshold. In this paper, a solution for the Least Squares estimator is proposed and simulated. The results show a significant departure from the trivial case such that pilots are positioned at distances greater than adjacent to the PU. Simulations were performed for 1 side of the PU as the result can be trivially applied to the other side of the PU's transmission.

This paper is organised as follows. Section II describes the system model used and Section III derives and explains the optimal solution to the outlined problem. In Section IV, the simulation parameters are given as well as results of the simulations themselves. The results are discussed in this section and a conclusion is derived from the findings in Section V.

2. SYSTEM MODEL

The system is modeled as an OFDMA transmission where the sub-channels which would interfere with the PU's

transmission are disabled in order to ensure no interference to the PU.

2.1. Channel Model

The multipath channel model which has been used is defined as [5]

$$h(n) = \sum_{l=0}^{L-1} \alpha_l \cdot \delta(n - \tau_l). \quad (1)$$

This is the time domain representation of the received signals affected by multipath where α_l and τ_l are the complex tap gain and delay for the l^{th} path of the channel with a total of L resolvable paths. To obtain the equivalent channel model in the frequency domain, the discrete Fourier transform (DFT) is applied to the time domain representation in (1), resulting in

$$H(i) = \sum_{l=0}^{L-1} \alpha_l \cdot \exp\left(\frac{-2j\pi\tau_l i}{N_{fft}}\right) \quad (2)$$

In the frequency domain representation shown by (2), i and N_{fft} are the sub-channel index and the FFT size respectively.

2.2. Pilot Error

The pilot error for the least squares estimator may be trivially derived from the estimation expression [5]

$$\hat{\mathbf{H}}_p = \mathbf{H}_p + \mathbf{P}^{-1} \mathbf{n}_p, \quad (3)$$

thus the error may be expressed as

$$\mathbf{\varepsilon}_p = \hat{\mathbf{H}}_p - \mathbf{H}_p = \mathbf{P}^{-1} \mathbf{n}_p. \quad (4)$$

where \mathbf{H} represents the channel gain matrix in both the time and frequency dimensions as derived in (2) and \mathbf{H}_p is the subset of \mathbf{H} at the pilot positions such that $p \subseteq i, n$.

2.3. Linear Interpolation Error Bound

The least squares estimate provides us with the estimated channel gain at pilot positions. While this is indeed the information needed for successful channel estimation, the channel gains at data symbol positions still need to be interpolated so as to successfully detect the data symbols. In this paper, a linear interpolator is considered for the sake of simplicity and to demonstrate the proof of concept. The interpolation error bound for a linear interpolator is noted as being dependent on the second derivative of the function being interpolated and the distance between the two interpolation points. It is therefore obvious that the more a function varies on a given interval, the higher the linear interpolation error will be. The upper bound for

linear interpolation error can be expressed as [6]

$$\varepsilon_{\text{int}} \leq \frac{d_{p,p'}^2}{8} \cdot \max \left| \frac{\partial^2 H(p, p')}{\partial i^2} \right|. \quad (5)$$

where p and p' indicate two pilot positions in the two dimensional grid.

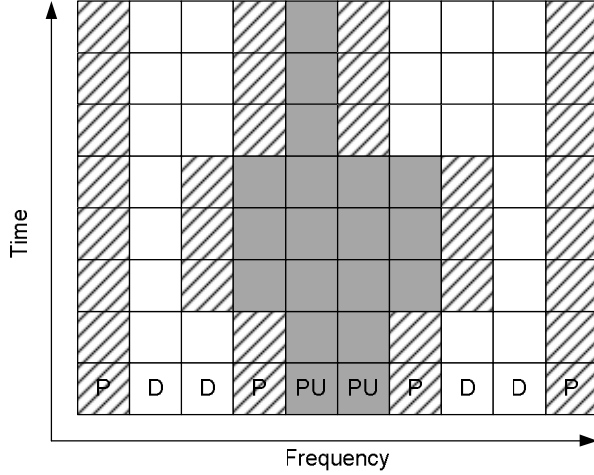


Figure 1. This figure demonstrates the likely situation of varying PU interference bandwidth for a CR system considering just optimal pilot patterns where the PU's transmission symbols (disabled by the CR) are denoted by dark grey, the SU's pilot symbols are denoted by the hatched pattern and the SU's data symbols are denoted by solid white.

2.4. Signal power density spectrum

The transmitted signals in the system simulation are assumed to be of a rectangular pulse shape. The power density spectrum of the rectangular pulse shaping function can therefore be represented as [4]

$$\phi_i(f) = P_i T_s \left(\frac{\sin(f \cdot \pi \cdot T_s)}{f \cdot \pi \cdot T_s} \right)^2. \quad (6)$$

In (6), P_i represents the transmit power of the i_{th} sub-carrier and T_s represents the symbol duration. While the equation only serves for a rectangular pulse shaping function, other equations may be substituted for (6) and the contradiction (and therefore solution) will still hold since all non-ideal filters have some form of spectral roll-off which presents interference to adjacent frequency bands.

2.5. PU-to-SU interference

As the signals between a PU and an SU are assumed to be non-orthogonal, the interference from the PU to the SU undergoes an additional process of smearing. This is due to

the fact that the Fast Fourier Transform (FFT) is performed by the SU [9]. The interference can then be modelled as [9]

$$E\{I_M(\omega)\} = \frac{1}{2\pi M} \int_{-\pi}^{\pi} \phi_{PU}(e^{j\omega}) \left(\frac{\sin[(\omega - \psi)M/2]}{\sin[(\omega - \psi)/2]} \right)^2 d\psi, \quad (7)$$

where ω represents the angular frequency which has been normalised to the sampling frequency and $\phi_{PU}(e^{j\omega})$ represents the power density spectrum of the PU's pulse-shaping filter.

Trivially, the total interference from the PU to the SU can then be described as the integral of the PDS of the smeared pulse shaping function. This integral is then expressed as

$$I_{PU}(d_i, P_i) = \int_{d_i - \Delta f / 2}^{d_i + \Delta f / 2} E\{I_M(\omega)\} d\omega. \quad (8)$$

In (8), d_i represents the spectral distance between the considered sub-carrier and the PU expressed as an integer number of sub-channels whereas Δf represents the width of each sub-channel of the SU (equivalent to the inverse of the OFDM useful symbol duration).

2.6. SU-to-PU interference

As we assume that we have no statistics or information about the modulation scheme of the PU, the interference from the SU to the PU is determined as the integral of the PDS of the spectral roll-off over the PU's bandwidth. This is effectively the integral of (6) over the given PU bandwidth and can be expressed as

$$I_{SU}(d_i, P_i) = \int_{d_i - B/2}^{d_i + B/2} \phi_i(f) df. \quad (9)$$

In this expression, B denotes the bandwidth occupied by the PU's signal such that the integration is performed over the PU's bandwidth with an added frequency 'offset' introduced by the spectral distance between the considered sub-channel and the PU's signal. This offset effectively represents the spectral roll-off measured from $d_i \cdot B$ Hz away as the roll-off, theoretically, extends to infinite frequency in both the positive and negative directions due to the time-limited nature of the pulse shaping waveform.

2.7. Optimal power loading

The optimal power loading algorithm is based on the interference models as specified in (8) and (9). The algorithm also takes into account a specified interference threshold parameter I_{th} such that it serves as a design-specified constraint. The optimal power loading is then derived at the interference threshold due to the fact that the most power is allocated per sub-channel when the

interference threshold is the highest. This achieves maximum channel capacity. The optimal power loading can then be expressed as [4]

$$P_i^* = \frac{1}{\lambda \cdot \frac{\partial I_{SU}}{\partial P_i}} - \frac{\sigma^2 + I_{PU}}{|H(i)|^2} \quad (10)$$

where λ is the Lagrangian multiplier used to find the optimal power level for each sub-channel, $H(i)$ is the channel frequency response gain at sub-channel i and σ^2 is the noise variance or effectively the power of the White Gaussian Noise (WGN) component in the interference noise. As the interference by the optimal power loading is limited by the threshold value I_{th} , the interference from the SU to the PU can be effectively ignored in the optimization solution due to the implicit constraint present in the power loading algorithm.

3. OPTIMAL SOLUTION

To obtain the optimal solution the problem may be formulated for the one dimensional (frequency) case and then repeated for the time dimension such that the estimation error between the concerned sub-channels, namely $i_0 \leq i \leq i_L$, is minimized for each time instance/OFDM symbol. The constrained optimization problem is therefore described as

$$\mathcal{E} = \min_i \left| \mathcal{E}_p \right| + \mathcal{E}_{int} \quad (11)$$

$$= \frac{\sigma^2 + I_{PU}(i)}{P_i^*} + \frac{(i - i_L)^2}{8} \max \left| \sum_{l=0}^{L-1} \frac{-4\pi^2 \tau_l^2}{N_{fft}^2} \alpha_l \exp \left(\frac{-2j\pi \tau_l i}{N_{fft}} \right) \right| \quad (12)$$

subject to,

$$i \leq i_L, \quad (13)$$

$$\sum I_{PU}(i) \leq I_{th} \quad (13)$$

$$\text{and } P_i \geq 0, \quad (14)$$

where $\forall i = 0, 1, \dots, i_L$.

In the context of the optimisation problem, i_L is used to represent the upper limit (i.e. adjacent to the nearest, original pilot sub-channel) of the possible pilot sub-channel placement position and i_0 represents the lower limit (i.e. adjacent to the PU).

As the optimisation problem is that of where to place the new pilot, the error function will always correspond to a decrease in MSE for the linear estimator due to the fact that the extra pilot is added compared to the case where the

system is left as-is and only the interfering sub-channels are disabled [7].

It was found that the optimal power loading equation (10) is of a transcendental form. Where the Karush-Kuhn-Tucker (KKT) conditions would normally be used to solve and provide the optimal solution for constrained optimisation problems, this was not the case due to the function's transcendental nature. It was therefore mandatory to optimise the error function through numerical computation rather than through the use of analytical methods. While this may seem inefficient and computationally expensive, in practice the value of i_L may not be bigger than the pilot spacing. This means that the optimisation problem is constrained to and only considers the sub-channels between the PU and the nearest pilot sub-channel (before insertion of the extra pilot sub-channel) per PU interference block and per OFDM symbol.

The provided solution is for a single side of the PU, this can be identically applied to the other side of the PU's transmission provided the PU's power remains relatively uniform throughout the PU's bandwidth.

4. SIMULATION PARAMETERS AND RESULTS

The simulations were performed through a Monte Carlo analysis such that 10000 runs were executed and a statistical average of the results was taken. The simulation parameters are described in Table 1.

Table 1. System simulation parameters.

Parameter	Value
PU sub-channel bandwidth	3 kHz
Channel path gains (dB)	[0 -15 -20]
Path delay times (μ s)	[0 0.4 0.9]
OFDM symbol duration	341 ms (333 μ s x 1024)
SU sub-channel bandwidth	3 kHz
FFT size	1024
Pilot spacing (frequency, time)	(12,12)
Maximum Doppler shift	24 Hz
PU signal power	20 dBm
Noise floor	-90 dBm

Several results were yielded from the simulations where the primary parameters compared were the error function values as a function of the interference threshold parameter I_{th} and the PU bandwidth size B measured in number of SU sub-channels overlapped.

Figures 2, 3 and 4 all demonstrate the error function value as a function of the spectral distance given different PU bandwidths. Figure 2 displays the results for the interference threshold parameter set to 1 mW, Figure 3 for

an interference threshold of 5 mW and Figure 4 for the interference threshold set at 9 mW.

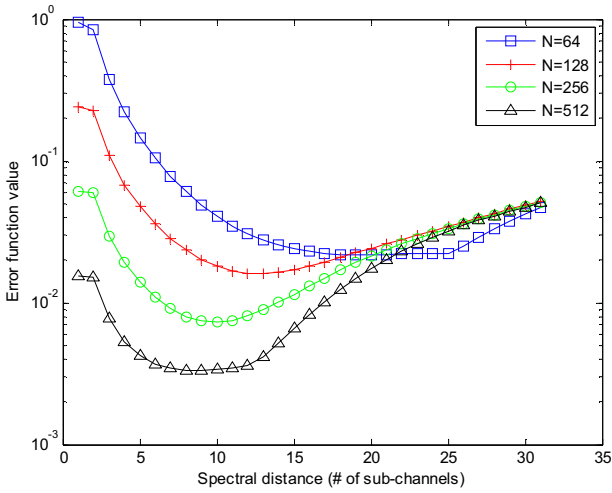


Figure 2. This figure shows the error function value as a function of the spectral distance of the new pilot and the PU for an interference threshold of 1 mW. The legend indicates the PU’s bandwidth relative to the number of SU sub-channels it occupies.

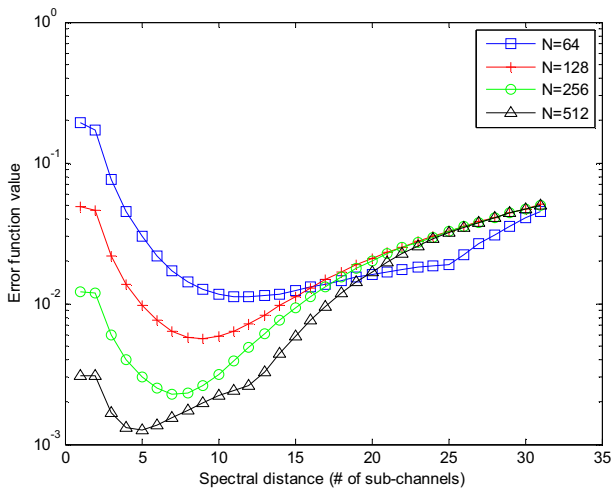


Figure 3. This figure shows the error function value as a function of the spectral distance of the new pilot and the PU for an interference threshold of 5 mW. The legend indicates the PU’s bandwidth relative to the number of SU sub-channels it occupies.

The initial and most obvious fact which may be noted is the fact that the optimal pilot placement (i.e. spectral distance between pilot and PU where the error function is at its lowest) is heavily dependent and influenced by the PU bandwidth. We may note that the optimal spectral distance decreases as the PU bandwidth increases. This may be attributed to the fact that the interference energy from the pilot sub-channel remains constant while the total bandwidth of the PU increases. Thus, the interference power over the total bandwidth of the PU decreases when the PU bandwidth increases. This is also attributable to the

increase in overall MSE as the number of existing pilot sub-channels is decreased, causing the algorithm to compensate.

Figures 5 and 6 both demonstrate the results for when the error function value is compared as a function of the spectral distance with the PU bandwidth being fixed (in this case, $N=64$ in Figure 5 and $N=512$ in Figure 6) while comparing for different interference thresholds.

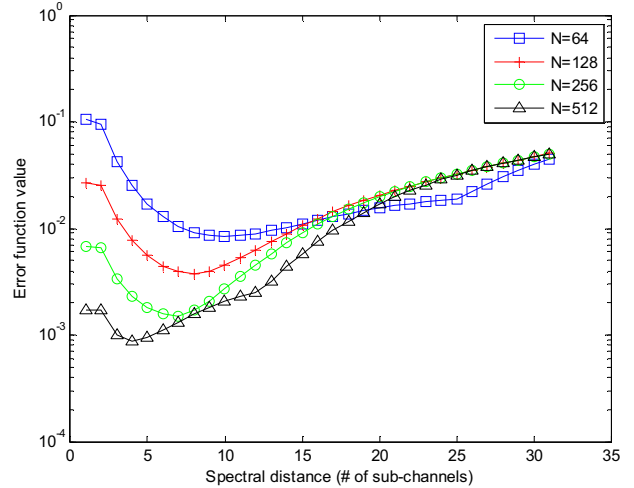


Figure 4. This figure shows the error function value as a function of the spectral distance of the new pilot and the PU for an interference threshold of 9 mW. The legend indicates the PU’s bandwidth relative to the number of SU sub-channels it occupies

The most obvious conclusion which may be drawn from both these figures is that the optimal distance between the PU and the new pilot decreases as the interference threshold increases.

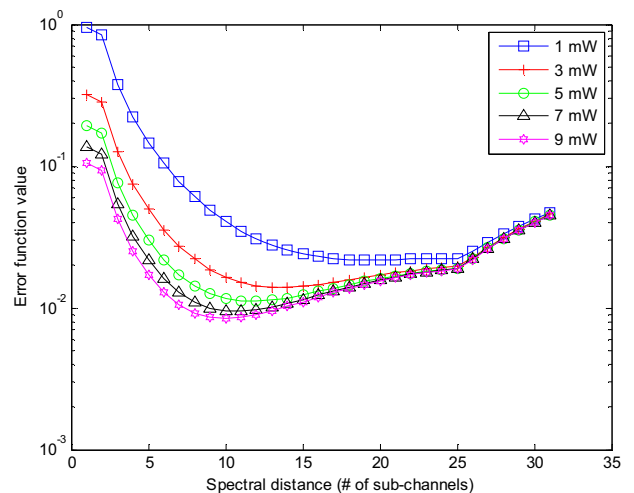


Figure 5. This figure shows the error function value as a function of the spectral distance for a PU bandwidth size of $N=64$ and different interference thresholds.

This is an obvious conclusion due to the fact that when the pilot is placed closer to the PU, the interference increases.

Therefore the higher the interference threshold, the closer the pilot may be. It should also be noted that the actual values of the error function at the optimal points decrease as the interference threshold increases which coincides with the fact that closer pilots have lower MSEs than ones farther away.

Figure 6 demonstrates the same datasets but with the interference bandwidth parameter set to 512 sub-channels. This coincides with the first three figures where we noted that the error function values decrease as the PU bandwidth increases, causing the pilots to be placed closer to the PU.

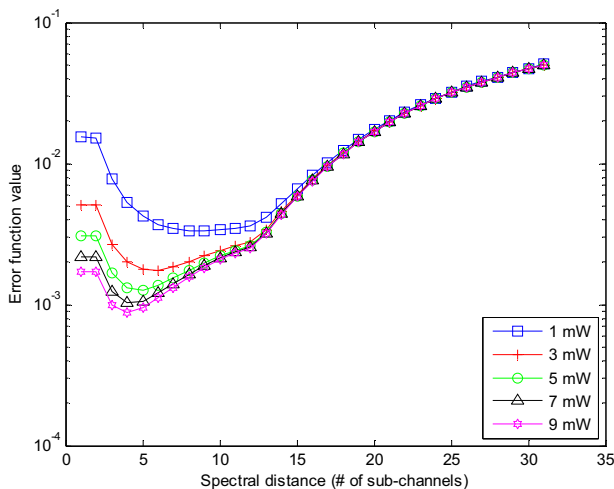


Figure 6. This figure shows the error function value as a function of the spectral distance with the PU bandwidth parameter fixed at $N=512$ and for different interference thresholds.

A major point to note in Figures 2-6 is the prevalence of what seem like discontinuities in some of the curves. This is attributed to the channel frequency response. As the pilots are evaluated at points farther away, the PU-to-SU and SU-to-PU interference terms decrease exponentially while the interpolation error term increases exponentially. Thus, the discontinuities are formed from the second derivative part of the interpolation error term and are due to the second derivative of the CFR increasing very steeply at those points and dominating the error expression defined by (11) and (12).

5. CONCLUSION

A contradiction was discovered where the optimal pilot placement algorithm for NC-OFDM CRs was ignorant to the optimal power loading algorithm and vice versa. A constrained, non-linear optimization problem and

algorithm were proposed such that the lowest LS estimator MSE is achieved while maintain interference to the primary user below a fixed threshold while also considering interference from the PU to the SU. The results showed that pilots should not be placed adjacent to the PU but rather a few sub-channels away.

Another aspect addressed by this research is also the impending issue of the formation of a cognitive radio standard. As pilot patterns and power loading algorithms of proposed cognitive radio standards are defined they will be seen to clash with each other. The research demonstrated has solved this issue and would allow for the optimal solutions to coexist and be implemented as a standard which would overcome the theoretical contradiction of these two areas.

REFERENCES

- [1] J. Mitola III and G. Q. Maguire Jr., "Cognitive radio: making software radios more personal," *Personal Communications, IEEE*, vol. 6, pp. 13-18, 1999.
- [2] 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, pp. 2127-2159, 9/15, 2006.
- [3] V. Valenta, Z. Fedra, R. Marsalek, G. Baudoin and M. Villegas, "Towards cognitive radio networks: Spectrum utilization measurements in suburb environment," in *Radio and Wireless Symposium, 2009. RWS '09. IEEE*, 2009, pp. 352-355.
- [4] G. Bansal, M. J. Hossain and V. K. Bhargava, "Optimal and Suboptimal Power Allocation Schemes for OFDM-based Cognitive Radio Systems," *Wireless Communications, IEEE Transactions on*, vol. 7, pp. 4710-4718, 2008.
- [5] Shichang Zhang, Jun Wang and Shaoqian Li, "A channel estimation method for NC-OFDM systems in cognitive radio context," in *Communication Systems, 2008. ICCS 2008. 11th IEEE Singapore International Conference on*, 2008, pp. 208-212.
- [6] S. D. Conte and C. de Boor, "Interpolation by polynomials," in *Elementary Numerical Analysis, an Algorithmic Approach*, 3rd ed. McGraw-Hill, 1980.
- [7] I. Rashad, I. Budiarto and H. Nikookar, "Efficient pilot pattern for OFDM-based cognitive radio channel estimation - part 1," in *Communications and Vehicular Technology in the Benelux, 2007 14th IEEE Symposium on*, 2007, pp. 1-5.
- [8] Chia-Horng Liu, "Adaptive two-dimensional channel estimation scheme for OFDM systems," in *Cognitive Radio Oriented Wireless Networks and Communications, 2008. CrownCom 2008. 3rd International Conference on*, 2008, pp. 1-5.
- [9] T. Weiss, J. Hillenbrand, A. Krohn and F. K. Jondral, "Mutual interference in OFDM-based spectrum pooling systems," in *Vehicular Technology Conference, 2004. VTC 2004-Spring. 2004 IEEE 59th*, 2004, pp. 1873-1877 Vol.4.

The authors would like to thank Telkom SA, the National Research Foundation (NRF), Technology and Human Resources for Industry Programme (THRIP), the Department of Trade and Industry (DTI), Nokia Siemens Networks, Ericsson and TeleSciences, for supporting this research project.

OPTIMAL SPECTRUM HOLE SELECTION & EXPLOITATION IN COGNITIVE RADIO NETWORKS

Mahdi Pirmoradian, Student Member, IEEE, Christos Politis, Senior Member, IEEE

WMN (Wireless Multimedia & Networking) Research Group, Kingston University London, U K

ABSTRACT

Future networks, especially in the framework of smart cities will be populated with various kinds of equipment accessing wireless communication channels. A much higher device variety will increase spectrum scarcity. Cognitive radio networks will be a key enabling technology in order to cope with the availability of the allocated radio spectrum bands. Cognitive Radio (CR) technology significantly utilizes current static spectrum bands assignment in an opportunistic manner. In this paper, we propose two spectrum opportunity (or spectrum hole) selection schemes; Minimum Collision Technique (MCT) and Maximum Residual Lifetime Technique (MRLT). The proposed techniques are evaluated by average channel utilization, average channel collision and successful secondary transmission bytes over licensed channels in a specific period of time (100s). The numerical results confirm that the MRLT scheme provides higher channel utilization and transmission bytes as well as decreases channel collision compared with the MCT scheme.

Keywords— Channel Utilization; Cognitive Radio; Spectrum Holes;

1. INTRODUCTION

Spectrum has become a scarce radio resource in emerging wireless technologies due to increasing use of wireless communication systems. The WWRF (Wireless World Research Forum) expects 7 trillion wireless devices will be serving 7 billion people by 2017, implying almost a 1000 wireless devices per person [1]. Therefore, the limited usable spectrum resources will be rapidly running out because of this phenomenon. According to an estimation of the ITU-R, the spectrum demand in 2020 will be 1280–1720MHz [2]. Based on the current static spectrum assignment approach, there will be insufficient spectrum for future wireless radio communications. To this respect, optimal spectrum utilisation is a crucial challenge in next generation wireless equipment. The usage of the current licensed radio spectrum without interference and disruption to licensed users' is new concept in recent years. Investigations and reports from Spectrum Policy Task Force show that 85% of spectrum bands are either partially or completely unused in different times and geographical locations [3]. For this reason, spectrum-sharing concept is a

well-known solution to tackle spectrum scarcity, which employs two mechanisms called: Overlay Spectrum Access, also referred to as Opportunistic Spectrum Access (OSA), and underlay spectrum access. OSA, which was first presented by Mitola under the cognitive radio concept, has the potential to dramatically increase spectrum utilisation by allowing license-exempt users to opportunistically re-use licensed spectrum in an interference-limiting manner. The main objectives in OSA context are: identifying, exploiting and managing unused spectrum portions in various locations and times.

Cognitive Radio (CR) is a paradigm for wireless communications, which is seen as the solution to the current low utilisation of the radio spectrum. According to the CR's definitions [4], [5], CR's objective is to observe its radio environment and intelligently adapts its parameters by using internal state and observed radio knowledge. Therefore, the prospect of CR effectively allowing license-exempt users, also called Secondary Users (SUs), to reuse licensed spectrum bands without harming licensed users called Primary Users (PUs). In this case, SUs need to identify and reuse the spectrum bands that are not being used by PUs in specific periods of time. The available spectrum bands for secondary user's usage vary with time and location. A region of location-time-frequency, available for a secondary user is called either Spectrum Hole (SH) or white spaces. Overall, a CR system will be able to coexist with a primary system by exploring and accessing to available SHs without harming spectrum owners [6]. This intelligent process will be done on cognitive engine (see Figure 1), under the Dynamic Spectrum Access (DSA) concept. In contrast, in underlay spectrum sharing, the cognitive radio devices are allowed to use licensed spectrum simultaneously as long as interference at the primary receiver are kept below predefined level. In this work, we focus on OSA mechanism where SU explores and exploits appropriate spectrum holes on vacant licensed channels before changing radio parameters for transmission. We reviewed most relevant topics to the present paper that have been studied by researchers. In [7], authors present an opportunistic spectrum access scheme based on distribution functions, where probability density function of unoccupied channels is estimated and selection is done according to the estimated probabilities. A spectrum hole prediction model based on the IEEE 802.11 standard is proposed in [8]. The distribution of interval time between two consecutive packets in multi user networks is analyzed. Moreover, in [9], researchers propose a spectrum occupancy model in paging band (928-948 MHz).

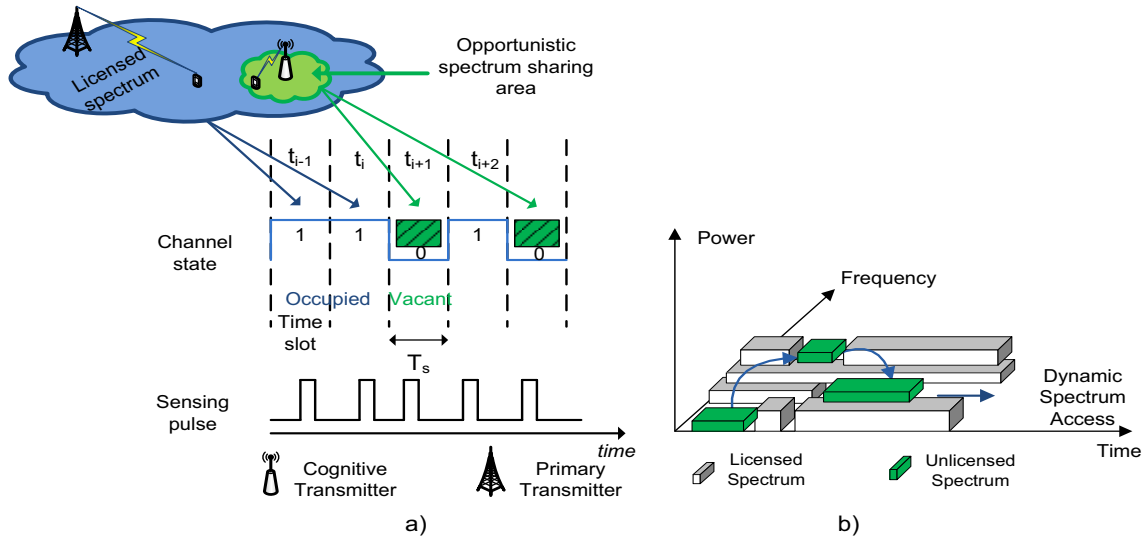


Figure 1. a) The cognitive network reuses unoccupied spectrum bands in various time slots. b) Dynamic spectrum access scheduling.

Where different statistical characteristics of the licensed channels are employed into the proposed model.

According to the literature, predicting PU’s presence lifetime and finding proper spectrum holes among vacant channels significantly impact SU’s activity over licensed networks. In this work we assume that SU monitors the surrounding radio spectrum and intelligently detects proper unoccupied channel in specific frequency bands and time, then communicates over a selected channel with minimal disruption to the licensed spectrum radio nodes. This paper utilizes N licensed channels in specific period of time, and the proposed MCT and MRLT techniques, which are employed at the MAC layer of SU, make licensed spectrum more efficient.

The rest of the paper is structured as follows. In section 2, spectrum hole is defined and described. In section 3, system model, spectrum sensing mechanisms and licensed channels model are explained. Residual lifetime of the idle channels and spectrum hole selection techniques are described and mathematically computed in section 4. Proper channel selection techniques, MCT and MRLT approaches are explained in 5. Numerical results and comparison between the two proposed techniques are presented in section 6, and finally conclusion and future concepts appear in section 7.

2. SPECTRUM HOLES

Spectrum holes can be defined as unoccupied frequency bands (licensed or license-exempt) in various space-time or frequency-time slots in radio environments. In time domain, the definition of a spectrum hole is easy to understand. It is the period of time that the PU is not transmitting. Figure 2 illustrates space-time and frequency-time spectrum holes definition where cognitive radio reuses unused frequency bands either in specific location or licensed band during a time slot. A spectrum hole in frequency domain is a vacant frequency band in which a SU is allowed to transmit during specific time without interfering with primary receivers

across all frequencies. In [5], spectrum holes are categorised into three types: 1) Black spaces; wherein the spectrum is occupied by high power primary user signals. 2) Grey spaces; where the spectrum band is occupied by weak primary signals, and the holes can be used considering interference power constraints. 3) White spaces, in which spectrum bands are clear of licensed signal. In conclusion, the grey and white spaces are the candidates to be exploited by licensed- exempt users.

Figure 3 illustrates the sensing blocks at the SU side and ON/OFF PU’s channel model. It is worth noting that the predefined threshold level at the sensing module, which can be measured and defined locally, affects the accuracy of the binary signal. In this respect, recognising proper spectrum hole on available channels at time (t) is a crucial challenge in CR context. The Quality of Service (QoS) of the primary and secondary users is considered since spectrum handoff and spectrum hole lifetime affects SU data transmission, as well as interference among users in the network. The spectrum-sensing module at the physical layer and resource allocation algorithms at the MAC (Medium Access Control) layer play essential roles in exploring and exploiting optimal spectrum holes in cognitive radio architecture. To this end, robust spectrum sensing modules and efficient resource allocation algorithm at the MAC guarantee appropriate QoS for both PUs and SUs.

In the cognitive physical layer, the accuracy of the primary signal detection at various times and locations needs to be enhanced, while at the cognitive MAC layer, robust and reliable algorithms are needed to allocate radio resources based on extracted statistics obtained by sensing data from PHY layer during observation times. By applying MCT and MRLT algorithms at cognitive MAC layer, unoccupied frequency bands at various times are reused by SU opportunistically. The spectrum hole residual time is a critical parameter, which influences the QoS of the SUs

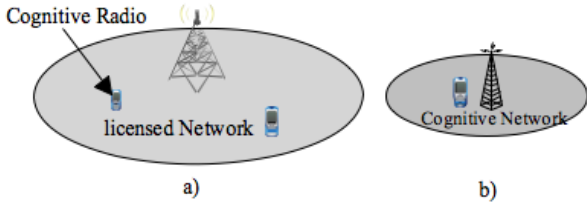


Figure 2. a) Temporary spectrum hole scenario, wherein cognitive radios opportunistically reuse licensed channels. b) Spatial spectrum hole, wherein cognitive network locates outside of the licensed network with caring to avoid interference with licensed users.

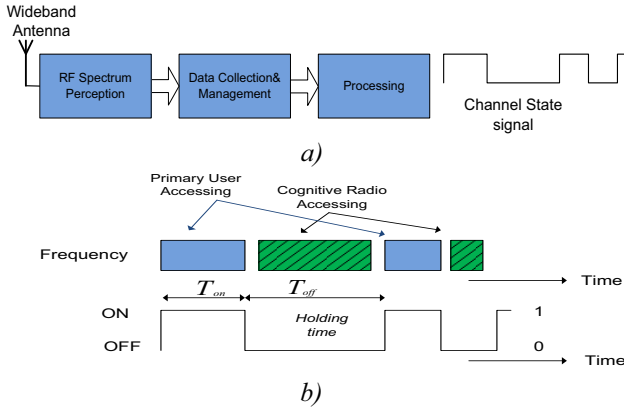


Figure 3. a) Sensing block at the PHY layer of the cognitive radio observes licensed spectrum bands occupation b) Binary ON/OFF signal provided by sensing module at the physical layer.

transmission, energy efficiency and channel utilization of SUs. It means transmission over unoccupied channels with high lifetime causes minor disruption to PUs because of less spectrum handoff at the SU side. On the other hand, more spectrum handoff causes more disruptions, power consumption and QoS might be greatly affected at SU.

3. SYSTEM MODEL

In this section, we explain the assumed network topology. We consider a licensed network with several wireless nodes, which communicate with each other by using N licensed channels, and a cognitive radio, which intends to utilize the licensed channel in an opportunistic manner. The sensing module is at the PHY layer of the SU, and equipped with a wideband antenna. The spectrum sensor at the SU performs perfect and accurate sensing also the probability of primary signal missing is ignored.

3.1. Spectrum Opportunity and Sensing Mechanisms

Identification and utilization of spectrum opportunity is the main function of cognitive radio. Three approaches are proposed to identify spectrum opportunities – *database registry*, *beacon signals*, and *spectrum sensing*. In the database registry approach, the information about spectrum opportunity is exchanged between the licensed and licensed-exempt (LE) users through a central database. In beacon signals mechanism, spectrum status and its

information synchronizes by transmitting beacon signals between licensed and LE users over a common channel. Spectrum sensing approach relies only on LE users and requires them to identify and track spectrum opportunities. The SUs find and exploits unused spectrums without interfering with PUs transmission. In spectrum sensing approach, several techniques such as cooperative centralized and cooperative distributed spectrum sensing have been presented to enhance efficiency of the applied mechanism.

Several spectrum detection methods, such as energy detection, match filter and feature detection, have been proposed as candidates for the sensing module at PHY layer. In spectrum sensing processes, the MAC layer inform PHY layer when and which channels need to be selected and sensed. The current draft of MAC layer in the 802.22 standard employs the quiet period mechanism, which consists of two stages with different time scales: fast sensing and fine sensing. During fast sensing stage, a fast sensing algorithm such as energy detection is done very fast (below 1ms/channel). The outcomes of the measurements are used for the fine sensing stage that follows. During fine sensing stage, more detailed sensing is performed on the selected channel [10].

The above spectrum sensing methods can be performed by two mechanisms namely: *Reactive Sensing*, in which an unlicensed user performs spectrum sensing only when this user needs access to the spectrum (i.e. on-demand basis) and *Proactive Sensing*, in which a licensed-exempt user continuously senses the spectrum, and the sensing results are maintained in the database. When an LE user wants to access the spectrum, it uses the database to locate the spectrum opportunity to be accessed. Lower overhead and longer delays are the main characteristics of reactive sensing. On the other hand, proactive sensing incurs larger overhead and lower access delays. Therefore selection between two sensing methods depends on application requirements [11]. In this paper, the reactive mechanism is considered at the SU side (see Figure 4).

3.2. Channel Usage Model

Surveying, analysing, processing and predicting of unused spectrum in both frequency and time domains require highly computational mathematical scenarios. In this subsection, PU's channel behaviour is explained mathematically. The utilisation of the licensed band by the PU is modelled as a Poisson process with arrival rate parameter μ , and the number of events in time interval $(t, t + \tau]$ can be given by [9];

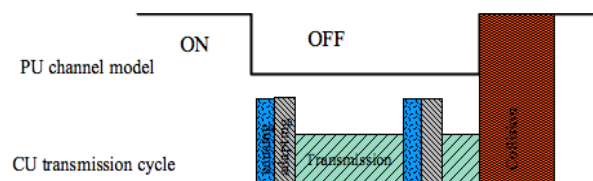


Figure 4. Secondary user transmission cycle over licensed channel, collision occurs when excess lifetime is less than period of transmission time.

$$P[(N(t + \tau) - N(t)) = k] = \frac{(\mu\tau)^k e^{-\mu\tau}}{k!} \quad k = 0, 1, 2, \dots \quad (1)$$

Here $N(t + \tau) - N(t)$ is the number of events in time interval $(t, t + \tau]$.

A single duration of utilization of the licensed band by a PU is denoted by T_{on} and a single duration of the licensed band being idle (unoccupied) is denoted by T_{off} , in the idle state, the secondary user can transmit information without harming primary receiver. The duration between two utilization periods (inter-arrival rate of the PU) are identical independent (i.i.d) random variables, which follows an exponential distribution. Thus, the probability density function of T_{off} (unoccupied time) for the licensed band can be expressed as:

$$f(t, \mu_{off}) = \begin{cases} \mu_{off} e^{-\mu_{off} t}, & t \geq 0 \\ 0, & t < 0 \end{cases} \quad (2)$$

Similarly, the probability density function of T_{on} (occupied band) for the licensed band is denoted as:

$$f(t, \mu_{on}) = \begin{cases} \mu_{on} e^{-\mu_{on} t}, & t \geq 0 \\ 0, & t < 0 \end{cases} \quad (3)$$

In the aforementioned functions, μ denotes OFF and ON time arrival rate. It is intuitive that the period of the OFF time plays an essential role in the quality of the SU transmission over captured licensed channel. Consequently, the SU needs to know the forthcoming state of the licensed channels to adapt its parameters to the new radio environment without exceeding bounding interference level because of its changing frequency. We assume that the sensing results are accurate enough to establish primary channel states.

4. SPECTRUM HOLE LIFE TIME

In this section, spectrum hole lifetime is described and mathematically estimated. This parameter plays an essential role in opportunistic channel access technique at the cognitive radio side. The located sensing task at the PHY layer of the SU senses all available license channels before data transmission begins to destination node. The approached sensing results take into account the channel selection algorithm, which aims to protect PUs from any disruption, interference and collision. As mentioned previously, licensed channels are modeled by ON/OFF binary signals, where ON and OFF duration times are exponential i.i.d random variables with probability density functions, $f_X(x)$ and $f_Y(y)$ respectively. Figure 5 (a) illustrates the sensing process and successful secondary transmission over opportunity spectrum bands via license channels. It is clear that excess lifetime or residual lifetime plays significant role on the performance of the channel selection scenario. In fact SU approaches at high-level QoS in great period of excess lifetime.

Hereafter, we assume, X , Y and γ are identical independent random variables, which denote ON, OFF and remain lifetime of the licensed idle channel respectively (see Figure 5 (b)). The probability density function of time up and time down are $f_X(x)$ and $f_Y(y)$. Consequently, excess lifetime of the vacant channel i at time (t) is given below as:

$$\gamma = t_{n+1} - t \quad (4)$$

Here t and t_{n+1} depict channel sensing instance time and the next PU's appearance time. For simplicity, all mathematical computations are done on one channel and then extended to N channels. The density function of residual idle time computed by SU, can be expressed as [12];

$$f_Y(x) = \mathbb{F}_X(x)/E(X), \quad x \geq 0 \quad (5)$$

Where, $\mathbb{F}_X(x) = 1 - F_X(x)$ represents survival function and $E(X) = E(T_{OFF})$ denotes expected value of period of the OFF time. Therefore distribution function of the residual time can be derived as follows:

$$F_Y(x) = \int_0^x f_Y(x) dx \quad (6)$$

Where $F_Y(x)$ denotes the probability that residual idle channel is less than x . Thus, T_{th}^i is assumed to be minimum required period of the SU transmission time. Also T_{th}^i is constant over all channels. According to the channel model and density function of OFF times (3) and secondary transmission threshold, the probability of collision among SU and PU transmission at time t can be computed as follows.

$$P(Y \leq T_{th}^i) = 1 - \exp(-\mu_{off} T_{th}^i), \quad T_{th}^i \geq 0 \quad (7)$$

Here appropriate idle channel will be selected under following constraint.

$$P(Y \leq T_{th}^i) < \varepsilon \quad (8)$$

Where ε shows the prescribed level of probability of the lifetime on the vacant channel. Moreover, in multi channel scenario, the proper unoccupied channel will be selected under maximum probability of survival channel at time instance t . In this paper, the behaviour of the licensed channels modelled by renewal theory in which for a renewal process alternating between state OFF and ON, the probability $P_{00}^{idle}(t)$ that state OFF is in use at time t , provided the process starts from state OFF is given as [13];

$$P_{00}^{idle}(t) = \frac{\mu_{on}}{(\mu_{off} + \mu_{on})} + \frac{\mu_{off}}{(\mu_{off} + \mu_{on})} e^{-(\mu_{off} + \mu_{on})t} \quad (9)$$

If we switch the role of two states then,

$$P_{11}^{busy}(t) = \frac{\mu_{off}}{(\mu_{off} + \mu_{on})} + \frac{\mu_{on}}{(\mu_{off} + \mu_{on})} e^{-(\mu_{off} + \mu_{on})t} \quad (10)$$

However, (9) represents the probability of idle channel residual lifetime at time t .

transmission, T_{th}^i , is 3.2ms. Also, the mean values of all ON states are 2 second and OFF-periods are assumed to be 1, 5, 3, 6, 2, 1, 7, 1, 4, 3 seconds.

The algorithms (8) and (12) have been employed in order to evaluate the performance of the proposed channel selection techniques (Algorithm). Moreover, ε and δ were assumed to be 0.2 and 0.8 respectively.

Algorithm. Channel selection algorithm using (13) and (14)

1. **Begin**
2. **Inputs** $N, \mu_{off}, \mu_{on}, \varepsilon, \delta, T_{th}^i$
3. **For** $i=1:N$
4. **Sense channels**
5. $\mathbb{N}(t) \leftarrow \text{unoccupied channels}$
6. Evaluate (12)
7. **end**
8. **If** $\mathbb{N}(t)$ is empty
9. **Stop** Transmission
10. **Else**
11. $H_j(t) = \text{argmin}(i | P^i(T^i \leq T_{th}^i) < \varepsilon) i \in \mathbb{N}(t)$
12. **If** ($j \neq 0$)
13. Transmission on channel j
14. **Else**
15. **Stop** Transmission
16. **End** (If)
17. $\bar{H}_{j1}(t) = \text{argmax}(i | T_R^{i \text{ dele } i} \geq T_{th}^i) i \in \mathbb{N}(t)$
18. **If** ($j1 \neq 0$)
19. Transmission on channel $j1$
20. **Else**
21. **Stop** Transmission
22. **End** (If)
23. **End**

Figure 6 reveals secondary utilization of the licensed channels in time duration 100s. In this case, SU observes ten-licensed channels with random interval sensing from 1 to 5 seconds. The initial results show that the licensed channels will be more utilized by MRLT technique. Also four licensed channels will be targeted by SU because of their probability constraint and channel characteristics. Basically channel utilization of the 7th channel is very greater than that of the other channels. We can see in Figure 7 the average channel collision on primary and secondary transmission in the proposed network topology. The results show MRLT scheme will reach fewer collisions compared to MCT scheme during the simulation period. Also Figure 8 reveals the performance of MRLT in sending successful bytes to secondary receiver. As can be seen in Figure 8, MRLT and MCT schemes send 26392500 Bytes and 25659375 Bytes during 100s, respectively. Consequently the outcomes confirm MRLT scheme could be much more applicable and reliable in exploiting spectrum opportunity and dynamic channel access in coexisting cognitive networks.

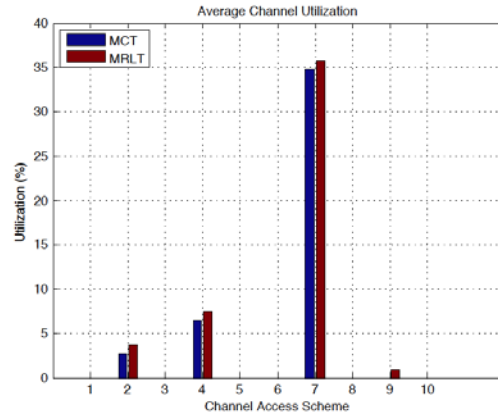


Figure 6. Comparison of the average channel utilization (MCT and MRLT techniques during 100s).

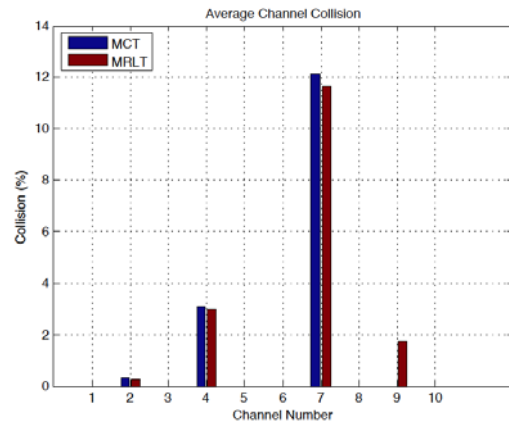


Figure 7. Comparison of the average channel collision (MCT and MRLT schemes during 100s).

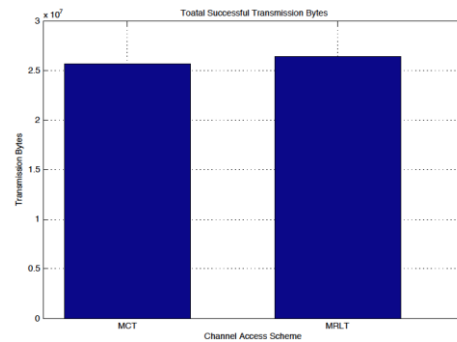


Figure 8. Total successful transmission bytes by secondary user

7. CONCLUSION AND FUTURE WORKS

In this paper, two spectrum opportunity schemes namely MRLT and MCT were proposed and their performances numerically analysed. Also channels' characteristics, such as distribution functions and channel rates were assumed known. The SU determined and selected appropriate spectrum opportunity based on its knowledge of licensed users behaviour and its prediction of the channel state at a future point in time. In this paper, the performance of the schemes such as channel utilization, channel collision and successful transmission bytes were evaluated during period of 100s. The simulation results revealed that MRLT scheme

approached higher capability in terms of utilization, collision and transmission bytes metrics.

However, the proposed spectrum opportunity selection schemes still require further investigation and analysis in cooperative and realistic wireless communication networks. Hence Multi-Users cognitive networks and primary networks might also be considered in the next stage of this work.

REFERENCES

- [1] D.Klaus, D.Sudhir and J.Nigle "2020 Vision," *IEEE Vehicular Technology Magazine*, pp. 22-29, September 2010.
- [2] "Estimated spectrum bandwidth requirements for the future development of IMT-2000 and IMT-Advanced", International telecommunication Union (ITU), Rep. 2010, Available: http://www.itu.int/dms_pub/itu-r/opb/rep/R-REP-M.2078-2006-PDF-E.pdf
- [3] "Spectrum policy task force report (ET Docket-135)", Federal Communications Commission, Tech. Rep., 2002. [Online] Available: <http://hraunfoss.fcc.gov/edocs/public/attachmatch/DOC-228542A1.pdf>
- [4] M. Joseph, "Cognitive radio: An integrated agent architecture for software defined radio", Ph.D Thesis, KTH Royal Institute of Technology, 2000.
- [5] H. Simon, "Cognitive Radio : Brain-empowered wireless communications", *IEEE Journal on Selected Area in communications*, vol.23, no. 2, pp. 201-220, Feb. 2005.
- [6] Z. Qianchuan, S. Geirhofer T. Lang, and B.M. Sadler, "Opportunistic spectrum access via periodic channel sensing", *IEEE Trans. Signal Process.*, vol. 36, no. 2, pp. 785–796, Feb. 2008.
- [7] W. Zhigang, F. Chunxiao, Z. Xiaoying, W. Yuexin, Z. Junwei, L.Jie, "A learning spectrum hole prediction model for cognitive radio systems", in Proc. 10th IEEE International Conference on Computer and Information Technology (CIT 2010), Bradford, UK.
- [8] J. Ohyun and C. Dong-Ho, "Efficient spectrum matching based on spectrum characteristics in cognitive radio systems", *Wireless Telecommunications Symposium*, pp. 230–235, April 2008.
- [9] G. Chittabrata, P. Srikanth, P.A. Dharma, M.W. Alexander, "A framework for statistical wireless spectrum occupancy modelling", *IEEE Trans. on Wireless Communications*, vol. 9, Issue 1, pp.38-44, January 2010.
- [10] C. Carlos, C. Kiran, B. Dagnachew, "IEEE 802.22: An introduction to the first wireless standard based on cognitive radios", *Journal of communication*, vol. 1, no. 1, April 2006.
- [11] H. Ekram, N. Dusit, H. Zhu, *Dynamic Spectrum Access and Management in Cognitive Radio Networks*, New York: Cambridge University, 2009
- [12] M.R. Sheldon, *Stochastic Process*, 2nd ed., New York: John Wiley & Son, 1996.
- [13] C.T. Henk, *Stochastic Models An Algorithmic Approach*, Chichester, New York, Brisbane, Toronto, Singapore:John Wiley & Sons Ltd, 1994, ch.1.
- [14] S.T. Kishora, *Probability and Statistics with Reliability, Queuing, and Computer Science Applications*, 2nd ed., New York: John Wiley & Son, 2002.
- [15] <http://standards.ieee.org/about/get/802/802.11.html>

SESSION 5

OPTIMISATION OF LAYERS 1 – 3

- S5.1 Transmission Analysis of Digital TV Signals over a Radio-on-FSO Channel
- S5.2 A Hybrid MAC with Intelligent Sleep Scheduling for Wireless Sensor Networks
- S5.3 Route Optimization Based On The Detection of Triangle Inequality Violations

TRANSMISSION ANALYSIS OF DIGITAL TV SIGNALS OVER A RADIO-ON-FSO CHANNEL

Chedlia Ben Naila¹, Kazuhiko Wakamori¹, Mitsuji Matsumoto¹ and Katsutoshi Tsukamoto²

¹Graduate School of Global Information and Telecommunication Studies, Waseda University, Japan

²Graduate School of Engineering, Osaka University, Japan

ABSTRACT

Recently, Radio frequency on free-space optical (RoFSO) technology is regarded as a new universal platform for enabling seamless convergence of fiber and FSO communication networks, thus extending broadband connectivity to underserved areas. In this paper, an experimental demonstration of the newly developed advanced RoFSO system capable of transmitting the Japanese integrated services digital broadcasting-terrestrial (ISDB-T) signals over 1km FSO link. Our innovative system combines a new generation full optical FSO system with radio over fiber (RoF) technology. The obtained results can be used for designing, predicting and evaluating the RoFSO system capable of transmitting multiple wireless services over turbulent FSO link.

Keywords— Radio-on-Free space optics (RoFSO), broadband wireless access, Digital TV, atmospheric turbulence.

1. INTRODUCTION

The emerging radio on free-space optics (RoFSO) systems are considered as a promising cost-effective solution able to satisfy the ever-increasing demand for capacity and quality in broadband wireless communication links [1-3]. Such system combines the radio over fiber (RoF) technology comprising heterogeneous wireless services and an FSO link, thus extending broadband connectivity to underserved areas. In RoF implementation, multiple RF signals are multiplexed in the RF domain and then transmitted through the atmospheric link using intensity modulation direct detection (IM/DD) method [4,5]. Nevertheless, the implementation of RoF solutions is dependent on the availability of installed optical fiber cables and the installation costs. In the absence of installed fiber cables, FSO links can conveniently be used to transmit radio-frequency (RF) signals through free space between end-points without the use of the fiber medium [6-8].

Expanding the development of information and communication technology (ICT) infrastructure, connectivity and access has been among the main concerns of the international telecommunication union (ITU) [9]. These ITU initiatives are addressed to bridge the digital divide between the developed countries and the least

developed ones and afford sustainable connectivity and access to remote and marginalized areas at all levels. In this regard, revolutionary broadband wireless technologies, like the proposed full-optical wireless communication RoFSO system described in this paper, are expected to play a vital role in accomplishing these ITU objectives.

Among ICT communications services, significant progress in Digital TV (DTV) technologies has been achieved in order to provide reliable high-quality video, sound and data broadcasting, leading to three main terrestrial DTV broadcasting standards around the world [10]: Digital Video Broadcasting-Terrestrial (DVB-T) in Europe, Advanced Television System Committee (ATSC) in North America and the Integrated Services Digital Broadcasting-Terrestrial (ISDB-T) in Japan. The orthogonal frequency-division multiplexing (OFDM) has been adopted in several high-speed digital communication standards such as DVB-T, ISDB-T, IEEE 802.11 local area network (LAN) and IEEE 802.16 standards. This is due to its increased robustness against frequency selective fading, narrow band interference and high channel efficiency [11]. In our previous work [12], it has been demonstrated that the RoFSO system can be optimally engineered according to a judicious selection of the optical modulation index (OMI) for the transmission of OFDM signals. Besides, it has been shown that the transmission performance of OFDM-based wireless services over RoFSO system is highly sensitive to the atmospheric turbulence, received optical power and to the selection of a proper OMI for optimum performance.

Recently, we have conducted experimental investigations for designing and evaluating a newly developed advanced full optical RoFSO system capable of transmitting ISDB-T signals over 1-km FSO link [13-15]. Our theoretical study proposed in [12] provided guidelines to optimally configure the experimental RoFSO system setup. In this paper, we report on the experimental results of evaluating the transmission performance of the ISDB-T signals over turbulent FSO link in terms of the modulation error rate (MER) and the bit-error rate (BER). This work provides insight on the system design, operation and performance characteristics relevant in implementing economical links for OFDM based wireless services transmission especially in areas lacking fiber infrastructure. The experimental evaluation of this innovative RoFSO system may serve then as guidelines for a standardization work in the ITU.

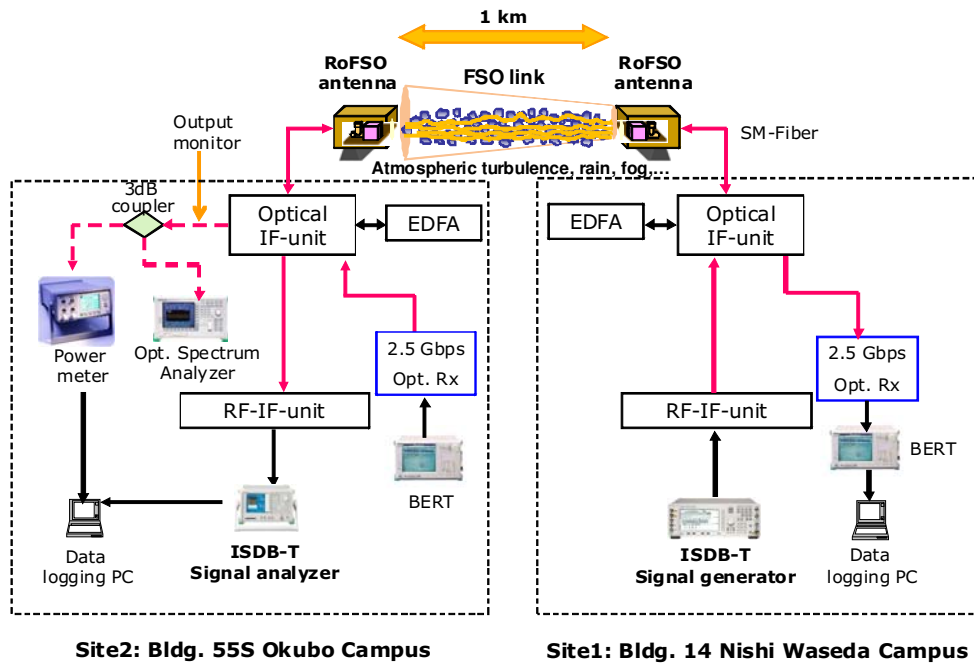


Figure 1. Experimental setup diagram of ISDB-T signal transmission using FSO links.

Table 1. Specifications of the RoFSO antenna.

Parameter	Specification
Operating wavelength band	1550 nm
Transmit power	100 mW (20 dBm)
Antenna aperture	80 mm
Coupling losses	5 dB
Beam divergence	$\pm 47.3 \mu\text{rad}$
Fiber coupling technique	Direct coupling using FPM
Tracking method	Automatic using QPD
	Rough: 850 nm beacon
	Fine: 1550 nm

Table 2. Basic transmission parameters of the ISDB-T system.

Transmission Parameter	Mode 1	Mode 2	Mode 3
No. of OFDM segments	13		
Bandwidth (MHz)	5.575	5.573	5.572
No of carriers	1405	2809	5617
Symbol length (ms)	252	504	1008
No of symbols per frame	204		
Guard interval length	1/4, 1/8, 1/16, 1/36		
Carrier Modulation	QPSK, 16QAM, 64QAM		
Information bit rate	3.65Mbps-23.23 Mbps		
Hierarchical transmission	Maximum 3 layers (A,B,C)		

This paper is organized as follows. In Section 2 we present a detailed description of RoFSO system we have developed. In section 3, the experimental results of OFDM based ISDB-T signals transmission using a prototype RoFSO system, are shown and discussed. Finally, Section 4 concludes the paper.

2. ROFSO SYSTEM SETUP

RoFSO system is designed to provide a reliable transmission of multiple RF signals and wireless services [14,16]. It uses 1550 nm as the transmission wavelength to be compatible with long-haul fiber optic technologies such as erbium doped fiber amplifiers (EDFA). In this system, an optical beam is simply emitted directly to free-space from the fiber termination point using the FSO transceiver and at the receiving end the optical beam is focused directly to the single mode fiber (SMF) core [15]. The proposed full-optical FSO antenna allows the seamless connection of a FSO beam to the SMF, thus eliminating the necessity of converting the transmitted signal from optical-to-electrical

(O/E) or vice versa in conventional FSO. The newly developed RoFSO system is able to provide high data rates in the order of several Giga-bit per second, e.g. 320 Gbps [17] and 1.28 Tbps [18]. It employs an innovative technique for initial antenna tracking and alignment and a technique using fine pointing mirror (FPM) for directly coupling the received optical beam to the SMF core, i.e., fine tracking. For initial alignment, the RoFSO antenna uses auto tracking whereby the beam is automatically realigned toward the opposite receiver [19,20]. The specification of the RoFSO antenna used in the experiments is given in table 1 [1].

In order to evaluate experimentally and characterize the transmission of digital terrestrial television broadcasting ISDB-T signals over a turbulent 1-km RoFSO link, we install two RoFSO antennas on the rooftop of two building with a distance of 1-km in Waseda University campus, Tokyo city. At one site, we set up ISDB-T signal generator (Anritsu MG3700A). At the second site, ISDB-T signal analyzer (Anritsu MS8901A) and other devices for measuring and recording the quality of the received RF and optical signals, weather data (temperature, visibility, rain

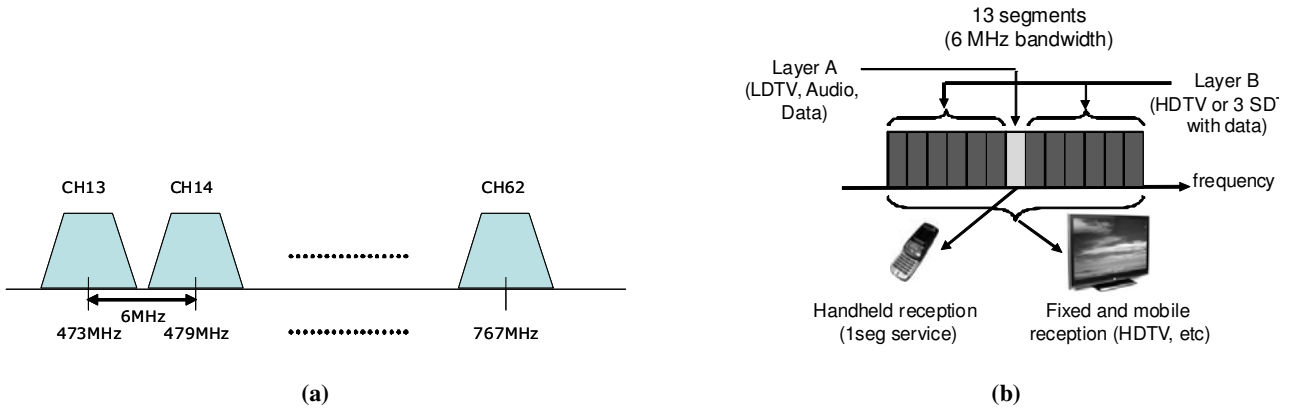


Figure 2. ISDB-T channel: (a) Frequency band and (b) segments and services.

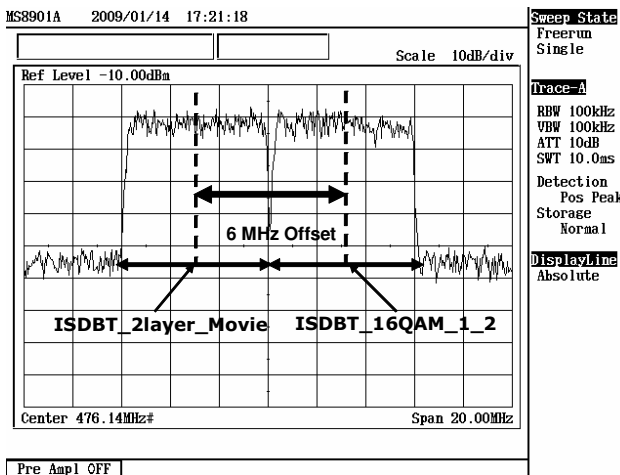


Figure 3. ISDB-T received signal spectrum.

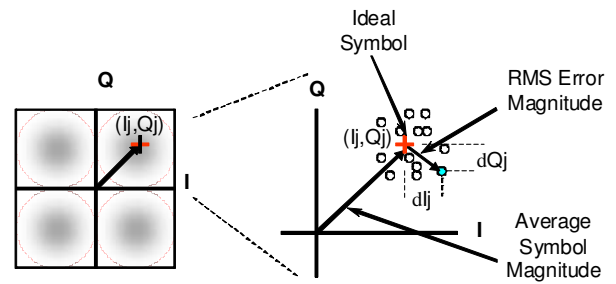


Figure 4. MER definition.

rate etc) as well as atmospheric conditions like scintillation effects are placed. In the RoFSO system configuration two interface units are included; an optical interface unit (Optical IF unit) and a RoF interface unit (RF IF unit). The optical interface unit consists of boost and post amplifiers and an optical circulator to isolate the transmitted and received signals. On the other hand, the RoF interface unit has RoF modules responsible for the electrical to optical signal conversion and vice versa. A schematic diagram representing the experimental setup is depicted in Fig. 1.

3. ISDB-T SIGNALS TRANSMISSION OVER ROFSO SYSTEM

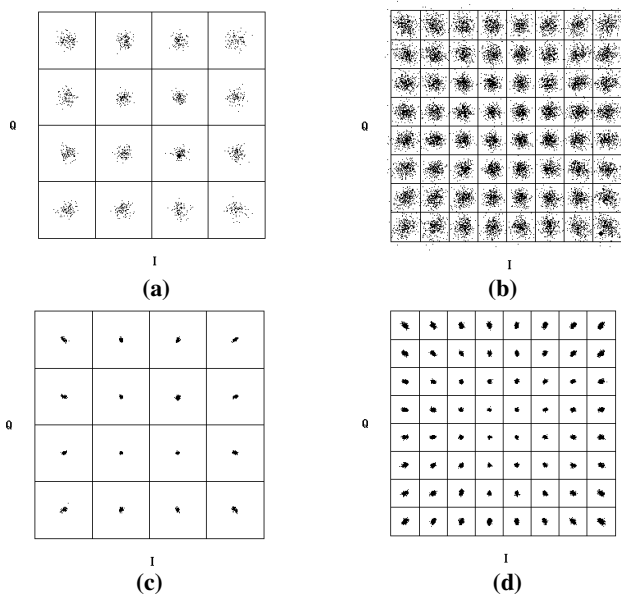
Integrated services digital broadcasting-terrestrial, referred to as ISDB-T, is an international digital television broadcasting standard developed in Japan designed to provide reliable high-quality video, sound and data broadcasting not only for fixed receivers but also for mobile receivers [21]. The ISDB-T system uses the ultra-high-frequency band at frequencies between 470 MHz and 770 MHz, giving a total bandwidth of 300 MHz. The bandwidth is divided into 50 channels numbered from 13 to 62 as

illustrated in Fig. 2 (a). Each channel is further divided into 13 OFDM segments which includes a single segment, (Layer_A or 1seg) for mobile receivers (LDTV, audio and data) and the remainder can be allocated as one 12-segments (Layer_B) for high definition television (HDTV) programs as depicted in Fig. 2 (b). One of the main features of the ISDB-T standard consists in having three transmission modes with different carrier intervals in order to deal with a variety of conditions such as the variable guard interval, information bit rate. The basic parameters of each mode are shown in table 2. In our experimental investigations, the two main metrics used to evaluate the quality of the transmitted ISDB-T signal over RoFSO link are both the bit error rate (BER) and the modulation error ratio (MER).

In our experiment, two signals are set simultaneously with the following waveform patterns (a) ISDBT_16QAM_1_2 and (b) ISDBT_2layer_Movie, both at -20 dBm with a 6 MHz frequency offset. The combined signal at -17 dBm is fed into the RoF module. The OMI for each channel (at -20 dBm input) is 10%. The signal is subsequently transmitted over the RoFSO link. The parameters of the two signals are listed in table 3. It should be noted that, the waveform pattern (a) is mainly used for BER and MER measurement and the waveform pattern (b) is mainly used for evaluation of video and voice data terminals. A received signal spectrum showing the two transmitted ISDB-T signals is depicted in Fig. 3.

Table 3. ISDB-T transmission parameters and required CNIR.

Pattern name	ISDBT_16QAM_1_2		ISDBT_2layer_Movie	
Mode	3		3	
Guard Interval	1/8		1/8	
Layer	A	B	A	B
Number of segments	1	12	1	12
Carrier Modulation	16 QAM	64 QAM	QPSK	64QAM
Inner Coding rate	1/2	7/8	2/3	7/8
Required CNIR [dB]	11.5	22.2	6.6	22.2
Information bit rate [kbps]	624.13	19660	416.08	19660

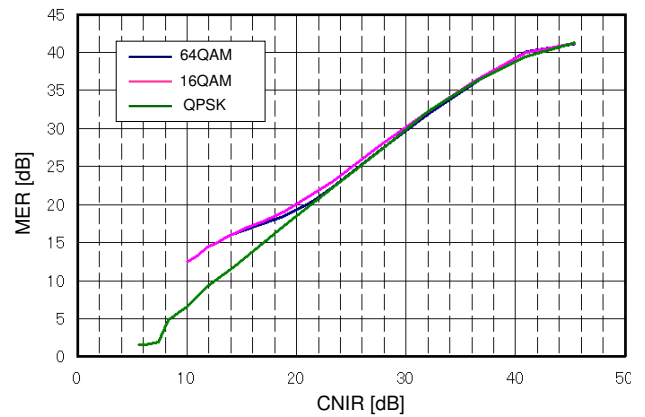
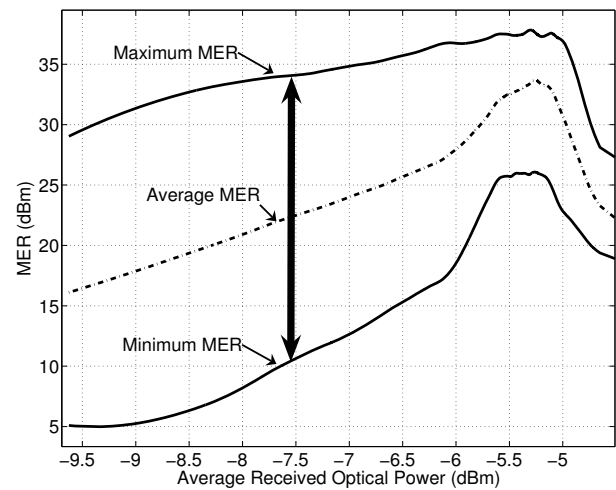

Figure 5. ISDB-T constellation maps (a) 1seg (MER= 23 dB), (b) 12seg (MER= 23 dB), (c) 1seg (MER= 35 dB) and (d) 12seg (MER= 35 dB).

3.1. MER performance

At first, the MER provides a figure of merit analysis to quantify the performance of a digital radio transmitter or receiver in a communications system using digital modulation. It gives an indication of the ability of the receiver to correctly decode the signal, defined as follows [22]

$$MER[dB] = 10 \times \log_{10} \left\{ \frac{\sum_{j=1}^N (I_j^2 + Q_j^2)}{\sum_{j=1}^N (dI_j^2 + dQ_j^2)} \right\} \quad (1)$$

where N is the number of received symbols, the vector (I_j, Q_j) is the j-th carrier's ideal symbol position resulting from a symbol decision using the actually received vector and the vector (dI_j, dQ_j) is defined as the difference between the actually received position vector and the ideal position


Figure 6. Variation of ISDB-T MER with the CNIR.

Figure 7. Variation of the MER versus the received optical power.

vector as shown in Fig. 4. In the RoF link the measured MER will be influenced by both RF noise figure and intermodulation distortion (IMD). An example of modulation analysis constellation for the ISDB-T signal made of Layer_A (16QAM) and Layer_B (64QAM) for different values of MER is shown in Fig. 5 (a), (b), (c) and (d) respectively. The constellation is very useful for analyzing the condition of the received signal by monitoring the modulation symbol movement.

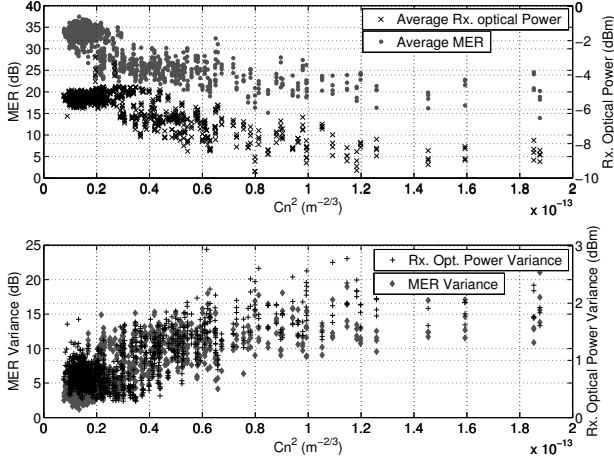


Figure 8. Variation of Average and Variance of ISDB-T MER with the C_n^2 .

The relationship between the measured MER and measured carrier-to-noise-plus-interference ratio (CNIR) for different types of modulation, i.e., QPSK, 16-QAM, 64-QAM is plotted in Fig. 6. It can be observed a linear relationship between the CNIR and the MER with almost equal values for the three modulation schemes.

Fig. 7 depicts the received ISDB-T MER after transmission over 1-km RoFSO link in clear weather condition, accumulated in 24 hours period. The data is collected every 1 second and for each minute the average, maximum and the minimum values are calculated. The MER curves for 16-QAM and 64-QAM are overlapped and thus have the same value, as shown in Fig. 6. From the Fig. 7, the obtained results show that, when the average received optical power is higher than -7.5 dBm, the value of MER is almost higher than 22 dB, satisfying the requirement for ISDB-T signal transmission with the MER to be higher than 11.5 dB for 16-QAM and 22 dB for 64-QAM [24]. However, it can be observed that the performance of the MER decreases when the received optical power is higher than a specific value (>-5 dBm), this is due to the IMD effect inherent to the laser diode nonlinearity. Furthermore, the received MER drops and shows large difference between the minimum and maximum values of the MER, with the decrease of the received optical power. This is due to the strong atmospheric turbulence period, i.e., refractive index fluctuation (C_n^2) is higher than $5 \times 10^{-14} (m^{-2/3})$, which causes a fast attenuation and large fluctuations of the received optical power as it can be seen in Fig. 8. Usually, the atmospheric turbulence remains the major issue to establish a reliable RoFSO link operating over long distances. In clear weather condition, the atmospheric loss is due to the scintillation variance and can be expressed as function of $C_n^2(m^{-2/3})$, $\lambda (m)$ the transmitter wavelength and $l (m)$ the transmission link distance [23]

$$Att_{scint}[dB] = 2 \sqrt{23.17 (2\pi/\lambda)^{7/6} C_n^2 l^{11/6}} \quad (2)$$

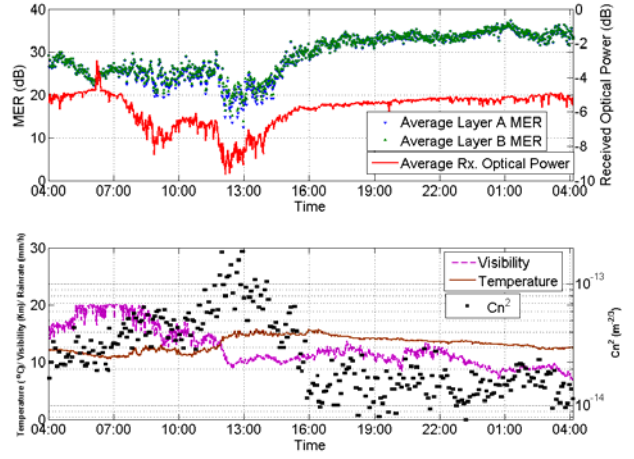


Figure 9. Received ISDB-T MER characteristics.

When the RoFSO system is used to provide connection between original broadcasting stations and retransmitters having no subsidiary relay stations and relatively small service area, the ARIB STD-B3 specifies the CNIR to be at least 35 dB. The measured MER values shown in Fig. 9 are below this specified threshold, due to the effects of atmospheric turbulence. In this experiment, the ISDB-T service signal was transmitted without any power optimization. If a power optimization was performed and the ISDB-T signal carrying wavelength was assigned slightly more power, a margin sufficient to suppress these effects of atmospheric turbulence would have been achieved. The recorded MER values would have improved to meet the specified threshold value of 35 dB which was confirmed by back-to-back measurements.

3.2. BER performance

In this experiment, we also measured the ISDB-T signal BER quality metric parameter. Results of the received ISDB-T Layer_A and Layer_B BER and received optical power characteristics for a clear day are shown in Fig. 10. The data is recorded for a continuous 24 hours of 11th December 2008 and the average received optical power represents the average value of received optical power recorded in 1 minute interval. Similar to the case of the MER, the Layer_A BER, Layer_B BER and the received optical power characteristics are observed to deteriorate causing increased burst errors especially at noon time due to the increase in the atmospheric turbulence induced beam wander and scintillation effects, resulting in random attenuation of the received signal.

The BER characteristics show satisfactory performance with most values being below the error correction limit (2×10^{-4}). Unfortunately, the automatic gain control (AGC) is occasionally inadequate in the case of 12-segments (Layer_B) transmission. However, the system achieves better performance where error free transmission can be observed. To get better insight of this data, the empirical cumulative distribution function (CDF) of

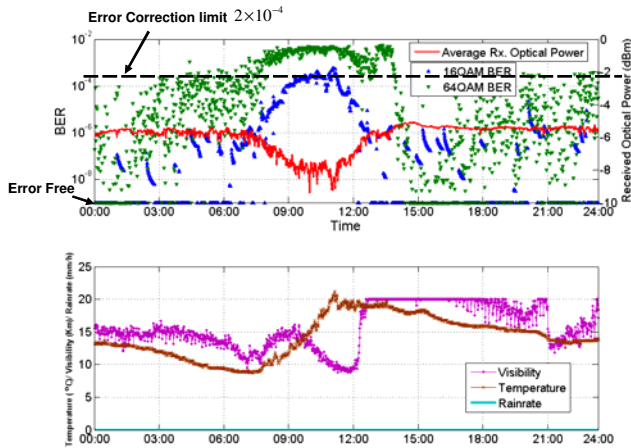


Figure 10. Received ISDB-T BER characteristics.

Layer_A and Layer_B BER is plotted in Fig. 11. It is shown that most likely 99% of Layer_A BER data is better than the required value of 2×10^{-4} with 50% of the data are error free, while for the Layer_B BER 72% of the recorded data is under the threshold with 18% error free. Therefore, under moderate atmospheric turbulence period, the system shows satisfactory BER performance for both Layer_A (16QAM) and Layer_B (64QAM) which are better than the standard requirement, demonstrating the suitability of the RoFSO system for ISDB-T signal transmission.

4. CONCLUSION

An experimental evaluation of a newly developed broadband wireless access technology based on RoFSO system has been presented. Important performance metric parameters for evaluating the quality of OFDM based terrestrial digital TV broadcasting signal transmission using RoFSO links, i.e., CNIR, BER and MER, have been quantified, measured, and characterized. The experimental results show that in the absence of severe atmospheric turbulence, a properly engineered RoFSO link can effectively be used to transmit OFDM based digital TV signals.

The ultimate goal of this work is to develop a robust RoFSO system capable of simultaneously transmitting multiple RF signals, which can operate in environments characterized with strong atmospheric turbulence manifested as scintillation effects, beam wander and angle-of-arrival (AOA) fluctuations. This work represents an attempt, based on a realistic operational scenario, aiming at demonstrating the RoFSO system can be conveniently used as a reliable alternative broadband wireless technology for complementing optical fiber networks in areas where the deployment of optical fiber is not feasible.

REFERENCES

- [1] K. Kazaura, K. Wakamori, M. Matsumoto, T. Higashino, K. Tsukamoto and S. Komaki, "RoFSO: a universal platform for

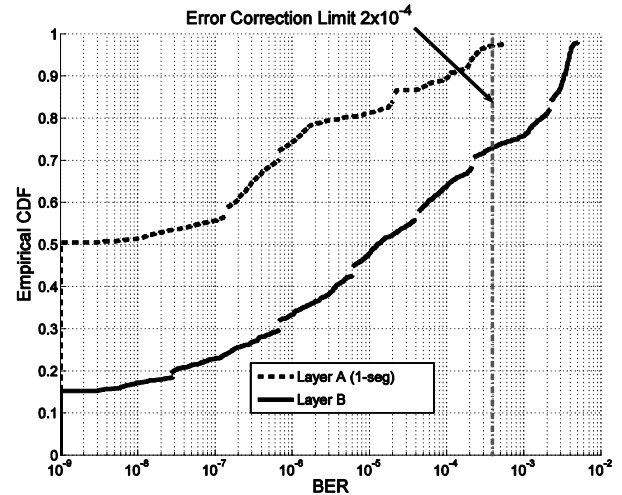


Figure 11. Empirical CDF of ISDB-T BER.

convergence of fiber and free-space optical communication networks," *IEEE Commun. Mag.*, vol. 48, pp. 130-137, Feb. 2010.

- [2] W. O. Popoola and Z. Ghassemlooy, "BPSK subcarrier intensity modulated free-space optical communications in atmospheric turbulence," *J. Light. Technol.*, vol. 27, no. 8, pp. 967-973, 2009.
- [3] C. Ben Naila, A. Bekkali, K. Wakamori and M. Matsumoto, "Transmission analysis of CDMA-based wireless services over turbulent radio-on-FSO links using aperture averaging," *Proc. of IEEE ICC*, Kyoto, Japan, June 2011.
- [4] T. Ohtsuki, "Multiple-subcarrier modulation in optical wireless communications," *IEEE Commun. Magazine*, vol. 41, no. 3, pp. 74-79, Mars 2003.
- [5] R. You and J. M. Kahn, "Average Power Reduction Techniques for Multiple-Subcarrier Intensity-Modulated Optical Signals," *IEEE Trans. Commun.*, vol. 49, pp. 2164--71, Dec. 2001.
- [6] H. Willebrand and B. Ghuman, "Free space optics: enabling optical connectivity in today's networks," Sams Publishing, 2002.
- [7] V. W. S. Chan, "Free-space optical communications," *J. Light. Technol.*, vol. 24, no. 12, pp. 4750-4762, Dec. 2006.
- [8] S. Bloom, E. Korevarr, J. Schuster, and H. Willebrand, "Understanding the performance of free-space optics," *J. Opt. Netw.*, vol. 2, 2003.
- [9] ITU-T Rec. F.2106, "Fixed service applications using free-space optical links," ITU, 2007.
- [10] Y. Wu, E. Pliszka, B. Caron, P. Bouchard, and G. Chouinard, "Comparison of Terrestrial DTV Transmission Systems: The ATSC 8-VSB, the DVB-T COFDM, and the ISDB-T BST-OFDM" *IEEE Transactions on Broadcasting*, vol. 46, no. 2, pp. 101-113, June 2000.
- [11] R. V. Nee and R. Prasad, *OFDM for Wireless Multimedia Communications*, Artech House, 2000.

- [12] A. Bekkali, C. Ben Naila, K. Kazaura, K. Wakamori, and M. Matsumoto, "Transmission analysis of OFDM-based wireless services over turbulent Radio-on-FSO links modeled by Gamma-Gamma distribution," *IEEE Photonics Journal*, vol. 2, no. 6, pp. 510-520, 2010.
- [13] C. Ben Naila, P. T. Dat, K. Wakamori, M. Matsumoto and K. Tsukamoto, "Next generation free space optics system for ubiquitous communications," *IEICE MWP Technical Report*, pp. 75-79, Feb. 2011.
- [14] A. Bekkali, P.T. Dat, K. Kazaura, K. Wakamori, M. Matsumoto, T. Higashino, K. Tsukamoto and S. Komaki, "Performance evaluation of an advanced DWDM RoFSO system for transmitting multiple RF signals," *IEICE Trans. on Fund. of Electron.*, vol. E92-A, no.11, pp. 2697-2705, Nov. 2009.[15] K. Kazaura, K. Omae, T. Suzuki, M. Matsumoto, E. Mutafungwa, T. Murakami, K. Takahashi, H. Matsumoto, K. Wakamori, and Y. Arimoto, "Performance evaluation of next generation free-space optical communication system," *IEICE. Trans. Electron.*, vol. E90-C no. 2, pp. 381-388, 2007.
- [16] K. Tsukamoto, K. Nakaduka, M. Kamei, T. Higashino, S. Komaki, K. Wakamori, Y. Aburakawa, T. Nakamura, K. Takahashi, T. Suzuki, K. Kazaura, K. Omae, M. Matsumoto, S. Kuwano, and H. Watanabe, "Development of DWDM Radio on free space optic link system for ubiquitous wireless," *Asia-Pacific Microwave Photonics Conference (AP-MWP Conference 2007)*, pp. 295-296, Jeju Island, Korea, April 25-27, 2007.
- [17] Y. Arimoto, M. Presi, V. Guarino, A. D'Errico, G. Contestabile, M. Matsumoto and E. Ciaramella, "320 Gbit/s (8X40 Gbit/s) Doublepass Terrestrial Free-space Optical Link Transparently Connected to Optical Fibre Lines," *ECOC 2008*, vol. 2990, pp. 1-2, Sept. 2008.
- [18] E. Ciaramella, Y. Arimoto, G. Contestabile, M. Presi, A. D'Errico, V. Guarino, and M. Matsumoto, "1.28 Terabit/s (32x40 Gbit/s) WDM Transmission over a Double-Pass Free Space Optical Link," *IEEE J. on Selected Areas in Commun.*, vol. 27, issue 9, (Special issue on optical wireless communications), 2009.
- [19] K. Takahashi, T. Higashino, T. Nakamura, Y. Aburakawa, K. Tsukamoto, S. Komaki, K. Wakamori, T. Suzuki, K. Kazaura, A. M. Shah, K. Omae, M. Matsumoto, Y. Miyamoto, "Design and evaluation of optical antenna module suitable for radio-on free-space optics link system for ubiquitous wireless," *SPIE Photonic West LASE 2008 Conference (Proc. SPIE 6877)*, San Jose Convention Center, San Jose, California, January 19-24 2008.
- [20] Y. Arimoto, "Compact free-space optical terminal for multi-gigabit signal transmission with a single mode fiber," *Proceedings of SPIE*, vol. 7199, no. 7, 2009.
- [21] DiBEG, <http://www.dibeg.org/>
- [22] ETSI Technical Report, ETR 290 "Measurement guidelines for Digital Video Broadcasting (DVB) systems".
- [23] L. C. Andrews and R. L. Philips, *Laser Beam Propagation through Random Media*, SPIE, Bellingham, WA, 2005.
- [24] ARIB STD-B21 "Receiver for Digital Broadcasting".

A HYBRID MAC WITH INTELLIGENT SLEEP SCHEDULING FOR WIRELESS SENSOR NETWORKS

Mohammad Arifuzzaman¹, Mohammad Shah Alam², Mitsuji Matsumoto¹

¹Graduate School of Global Information and Telecommunication Studies, Waseda University, Tokyo, Japan; ²Research Institute for Science and Engineering, Waseda University, Tokyo, Japan; ¹arif@fuji.waseda.jp, ²alam@aoni.waseda.jp, ¹mmatsumoto@waseda.jp;

ABSTRACT

In this paper, we present Intelligent Hybrid MAC (IH-MAC), a novel low power with minimal packet delay medium access control protocol for wireless sensor networks (WSNs). IH-MAC achieves high energy efficiency under wide range of traffic load. It ensures high channel utilization during high traffic load without compromising energy efficiency. IH-MAC does it by using the strength of CSMA and TDMA approach with intelligence. The novel idea behind the IH-MAC is that, it uses both the broadcast scheduling and link scheduling. Depending on the network loads the IH-MAC protocol dynamically switches from broadcast scheduling to link scheduling and vice-versa in order to achieve better efficiency. Furthermore, IH-MAC uses Request-To-Send (RTS), Clear-To-send (CTS) handshakes with methods for adapting the transmit power to the minimum level necessary to reach the intended neighbor with a given BER target or packet loss probability. Thus IH-MAC reduces energy consumption by suitably varying the transmit power. The analytical results corroborate the theoretical idea, and show the efficiency of our proposed protocol. Considering the importance of a unique MAC protocol for WSNs, we propose a study for standardization work in the ITU as an initiative which can lead to its rapid adaptation.

Keywords— Medium access control, energy efficiency, wireless sensor network.

1. INTRODUCTION

Wireless sensor networks (WSNs) have become very popular in recent years. Many WSNs are based on proprietary standards for wireless networking, but the recent trend has been increasingly towards the standardization of low power wireless communication. The first step of standardization for such low rate wireless personal area networks was taken in 2003 when IEEE 802.15.4 was approved. IEEE 802.15.4 standard specifies only the lowest part of OSI communication model: PHY layer and MAC sub-layer which are briefly overviewed with the area of our proposed work in figure 1. With 802.15.4, IEEE had a goal in mind for low-cost, low-power and short-range wireless communications. This standardization process continued and went through an

enhancement process. As a result, newer versions like IEEE 802.15.4b, 802.15.4a, 802.15.4c and 802.15.4d were released subsequently. But unlike 802.11 WLAN cards where MAC is usually included as part of the chipset, in WSNs the MAC designer has absolute control on the design of MAC layer. So, on the basis of IEEE 802.15.4 standard though a lot of MAC protocols for sensor networks have been proposed in recent years, still researchers are settled and agreed on one point that a definite and universally accepted standard MAC protocol for wireless sensor network is really needed. We hope our IH-MAC protocol will certainly contribute to the standardization process of MAC layer protocol for wireless sensor network.

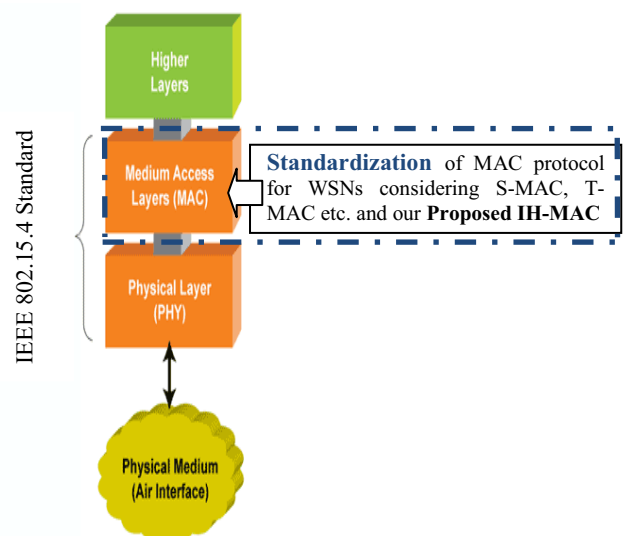


Figure 1. Structure of IEEE 802.15.4 protocol stack and the area of our proposed work

WSN consists of a large number of wireless sensor nodes that are deployed randomly. The sensor nodes are typically small, and equipped with low-powered battery. Unlike other wireless networks, it is generally impractical to charge or replace the exhausted battery. Since prolonging lifetime of the sensor nodes is very important, energy efficiency becomes the most important attribute of design of MAC protocol of sensor networks. Other attributes are fairness, latency, delivery ratio, and bandwidth [1]. Idle

listening is the major source of energy wastage for wireless sensor networks [2]. Therefore, in sensor network, nodes do not wake-up all the time rather prefer energy preservation by going to sleep time to time as explained in figure 2.

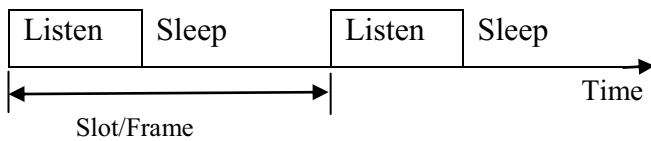


Figure 2. Periodic listen and sleep of a sensor node

After the sleep scheduling, nodes could operate in a low duty cycle which can significantly save energy and extend the network lifetime at the expense of increased communication latency and synchronization overhead. In [3] different sleep scheduling schemes are analyzed and a scheduling methods that can decrease the end to end delay is proposed. But this method does not provide an interference free scheduling. One obvious approach is TDMA MAC which can inherently support low duty cycle operation. Besides TDMA has natural advantage of contention and collisions free transmission [1]. To be interference free, a straightforward approach can be to assign each communication link a slot, and thus the number of slot is equal to the number of communication links of the network. However, this scheme requires much more slots than necessary, which enhance delay and reduces the channel utilization. Moreover, minimizing the number of slot assignment for producing an interference free link scheduling is a NP complete problem [4]. On the other hand performance of broadcast scheduling is worse than link scheduling in WSNs, in terms of energy conservation. Henceforth, we propose a new hybrid MAC protocol for wireless sensor network, called IH-MAC, which combines the strength of CSMA, link scheduling and broadcast scheduling.

The rest of the paper is organized as follows. Section 2 reviews related works. In Section 3 we will elaborate on the design of the IH-MAC protocol. In section 4 we will describe an analytical model, followed by results in section 5. And finally section 6 concludes the paper and mentions some guideline of the scope of future works.

2. RELATED WORK

For sensor network, S-MAC [2] is one of the pioneering works in contention based MAC protocol. In S-MAC nodes operates in low duty cycle and energy efficiency is achieved by periodic sleeping. Nodes form virtual clusters, based on common sleep schedules, to reduce control overhead and enable traffic adaptive wake-up. T-MAC [5] improves the energy efficiency of S-MAC by introducing adaptive duty cycle. T-MAC reduces the idle listening by transmitting all messages in burst of variable length and sleeping between bursts and it maintains an optimal active time under variable load by dynamically determining its

length. In AMAC [6] each node can adjust duration of the active period depending on traffic. In [7] the performance analysis of optimized medium access control for wireless sensor network is done. B-MAC [8] is the default MAC for Mica2. B-MAC allows an application to implement its own MAC through a well-defined interface. Z-MAC [9] dynamically adjusts the behavior of MAC between CSMA and TDMA depending on the level of contention in the network. The protocol uses the knowledge of topology and loosely synchronized clocks as hints to improve MAC performance under high contention. Z-MAC uses DRAND [10], a distributed implementation of RAND [11] to assign slot to every node in the network. TH-MAC [12] is a traffic pattern aware hybrid MAC protocol inspired from Z-MAC. It uses A-DRAND as slot assignment algorithm. A-DRAND is an improved version of DRAND for clustered wireless sensor networks where cluster heads require more slots to relay packets.

Our proposed IH-MAC also combines TDMA and CSMA. But IH-MAC is completely different from Z-MAC and similar hybrid MAC in the sense that in IH-MAC, each node calculates its own slot independently which is very flexible. Moreover, the hybrid concept used in case of IH-MAC does not stand for the same meaning, as the meaning used by other already proposed hybrid MAC. IH-MAC is hybrid in the sense that it combines CSMA, the broadcast scheduling and link scheduling dynamically to improve the energy efficiency. Another important feature of IH-MAC is that it reduces energy consumption by suitably varying the transmit power.

3. INTELLIGENT HYBRID MAC (IH-MAC) PROTOCOL DESIGN

We first define the terminology used in this paper. A *Slot* or *Frame* is defined as the periodic interval, which consists of an active period and a sleep period. A *duty cycle* is the proportion of active period to entire cycle time. A *rendezvous slot* is defined as a slot explicitly dedicated to a pair of nodes to communicate with each other.

3.1. Neighbor Discovery, Clustering and Synchronization

Frame synchronization is done by virtual clustering, as described in the S-MAC protocol [2]. When a node comes to life, it starts by waiting and listening. If it hears nothing for a certain period, it chooses a frame schedule and transmits a SYNC packet. The SYNC packet contains the time until the next frame starts. If the node during start up hears a SYNC packet from another node, it follows the schedule in that SYNC packet and transmits its own SYNC accordingly. Nodes retransmit their SYNC once in a while. When a node has a schedule but it hears SYNC with a different schedule from another node, it adopts both schedules. Adopting both schedules ensures the successful communication between the nodes of different schedule. The described synchronization scheme, which is called

virtual clustering [2], urges nodes to form clusters with the same schedule. So, all the nodes in the networks need not to follow the same schedule.

During this virtual cluster creation, each node creates the one hop neighbor list and with using these a node can easily constitutes the two hop neighbor list. After that each node is given an id such that within a two hop neighbor the id is unique.

3.2. Slot Assignment

Each slot in IH-MAC consists of a fixed length SYNC period, a fixed length data period (For RTS/CTS) and a sleep period that depends on the duty cycle. The duty cycle should be chosen in such a way that the sleep period of a slot is large enough to transmit a data packet along with ACK. All nodes are allowed to transmit in any slot, but the owner of the slot will get the priority. Priority can be ensured by choosing contention window size which is elaborately described in the part III E of this paper. The owner calculation can be performed by each sensor node locally by simple clock arithmetic. For example, if there are 8 neighbor nodes (every node is 1 or 2-hop neighbor to each other), the node 1 will be the owner of the Slot 1, 9, 17.....etc. The procedure is explained in figure 2 where T1, T2..., T10 represent the slot sequences and S1, S2..., S8 represent the sensor nodes. So according to the clock arithmetic (modulo 8) in figure 3, the sensor node S1 is the owner of the slot T1 and T9.

S1	S2	S3	S4	S5	S6	S7	S8	S1	S2
T1	T2	T3	T4	T5	T6	T7	T8	T9	T10

Figure 3. Owner selection of each slot for 8 sensor nodes.

S1	S2	S3	S4	S5	S6	S7	S8	S1	S2
T9	T10	T11	T12	T13	T14	T15	T16	T17	T18

Figure 4. Rendezvous slot selection for 8 sensor nodes by using modulo 16

Now, each node can make some of its owned slot as a rendezvous slot with which it can send message to its neighbor exclusively. The rendezvous slots can be also calculated by clock arithmetic, as modulo m. The value of m is set according to the system requirements, i.e. network load, delay, message buffer size etc. m will be always multiple of node id. For instance, let node 1 wants to create a rendezvous slot. By using modulo 16, the rendezvous slots of node 1 will be a subset of [1, 17...etc.]. The procedure is explained with the figure 4 where T9, T10..., T18 represent the slot sequences and S1, S2..., S8 represents the sensor nodes. If we use modulo 16, node S1 can make slot T17 as its rendezvous slot. Here it is

noticeable that, though node S1 is owner of both slots T9 and T17 but S1 cannot make T9 as its rendezvous slot. It is because 9 is not a subset of [1, 17 ...etc.].

For the sake of scalability the value that we use in modulo operation i.e., m will be always larger than the number of two hop neighbor nodes in a virtual cluster. So when a new node wants to join in the network, at least there will be some slots which are not using as rendezvous and it will be used for the scalability.

3.3. Transmission

Each node will sleep for some time and then periodically wakes up to see whether any other node wants to talk to it. During sleeping, the node turns off its radio, and sets a timer to awake later. If a node wants to send data to another node it will check whether the node itself is the owner of the slot. If it is the owner of the slot it will get priority. If it is not owner of the slot it will contend with other nodes to get the slot. Broadcast packets are sent without Request-To-Send (RTS) and Clear-To-Send (CTS). Unicast packets will follow the sequence of RTS, CTS, Data, and Acknowledgement (ACK). This scheme is well recognized and used, for example in the IEEE 802.11 standard [13].

Now, if messages for a particular node queued in its buffer cross threshold value the node will make some of its owned slots as rendezvous slots. The node will first broadcast the declaration of its making rendezvous slot. The declaration message contains how many slots will be used as rendezvous slot, and between whom the rendezvous will be done. So, remaining neighboring nodes can calculate locally about the slot so that they need not to wake up during those slots. For each rendezvous slot since all the neighbors of both sender and receiver will be in sleep mode, there will be no hidden or exposed terminal problem. So, on those cycles no RTS-CTS are required. Only Data-Ack will work. Part of the energy saving scheme of IH-MAC is pictorially represented in the timing diagram of Figure 5.

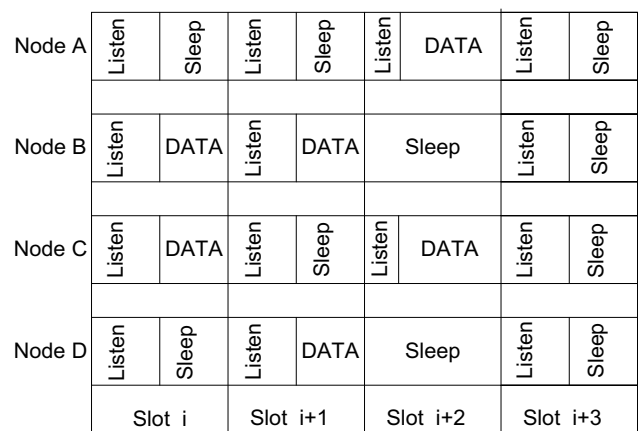


Figure 5. Timing diagram of sensor nodes working in IH-MAC

Consider a simple case scenario of four sensor nodes A, B, C, and D where each node is within transmission range of others and they follow same schedule. In Fig. 5 we consider four consecutive slots namely, i , $i+1$, $i+2$ and $i+3$. Each slot is further divided into two parts, the first one is listen part which is used for SYNC, RTS and CTS, and the next portion is sleep part, which is used for data transmission between two nodes. The proportion of listen and sleep interval depends on the duty cycle of the operation of sensor nodes. We take some arbitrary transmission to clarify the different consequences of working principle of IH-MAC. Now, during slot i , let data transmission occur between node B and C. But node A and D also need to wake up to see whether there is any data for them. But subsequently they go to sleep because either there is no data to receive or send by them or they lose in contention. Similar situation occurs in slot $i+1$ where transmission is occurred between node B and D. In slot $i+2$, node A and C created rendezvous between them. So, on that slot node B and D will not wake up rather they will sleep the whole slot. Therefore, during this period node B and D save energy by operating with zero duty cycle, lingering their sleep time as well as by avoiding transition from sleep to active state. And node A and C will save energy by avoiding RTS and CTS and contention for getting the slot. Another power savings feature of IH-MAC is by adjusting transmission power, which is explained in section 3.4

3.4. Transmission power adjustment

The power adjustment features of IH-MAC allow the sensor nodes to suitably vary the transmission power to reduce energy consumption. This idea is based on power control protocol for wireless ad-hoc network proposed in [14]. IH-MAC transmits the RTS and CTS packets with maximum power P_{max} . When receiver node receives an RTS packet, it responds with a CTS packet at usual maximum power level P_{max} . When the source node receives this CTS packet, it calculates $P_{desired}$ based on the received power level P_r and transmitted power level P_{max} as

$$P_{desired} = \frac{P_{max}}{P_r} \times Rx_{thres} \times c$$

Where Rx_{thres} is the minimum necessary signal strength and c is a constant. The source node uses power level $P_{desired}$ to transmit data packet. Similarly, receiver uses the signal power of received RTS packet to determine the power level to be used $P_{desired}$, for the ACK packet. This method assumes the attenuation between sender and receiver nodes to be the same in both directions. It also assumes the noise level at nodes to be below a certain predefined threshold value.

Since IH-MAC allows data transmission between only one pair of nodes in a slot and all the neighbors of both sender and receiver sleep during transmission, it overcomes the

shortcomings of said technique, like increased collision and degradation of network throughput.

3.5. Contention window size and owner's priority

Owner's priority can be set by using different contention window size for owners and non owners. Owners of a slot picks a random time uniformly over contention interval $[1, CW_{own}]$, while non owners do so within $[1, CW_{nwn}]$. The average window size observed by an owner node would be $(1 + CW_{own})/2$ and for the non owner would be $CW_{own} + (1 + CW_{nwn})/2$. The owner takes hold of the channel every time because of its smaller contention window, provided that owner of the slot has some data to send. Once the slot is chosen, the node transmits at that slot. So, for both owner and non owner nodes, SYNC and RTS transmission in IH-MAC always starts by waiting and listening for a random time within the contention interval. But when a slot is already declared as rendezvous slot for that slot without waiting for contention window the node can initiate transmission.

4. ENERGY CONSUMPTION ANALYTICAL MODEL

An analytical model for the energy consumption of nodes for IH-MAC is explained in this section. For simplicity we consider the case where a sensor node is either in broadcast scheduling mode or in a link scheduling mode. Let d be the duty cycle and t_{SIM} be the simulation time and t_{TX} , t_{RX} , t_{OH} , t_{IDLE} , t_{SLEEP} , t_{TRANS} are denoted as the time spent for transmitting, receiving, overhearing, idle listening, sleep, and radio transitions during sleep to wakeup state of a sensor node, respectively.

So, t_{SIM} can be expressed as

$$t_{SIM} = t_{TX} + t_{RX} + t_{OH} + t_{IDLE} + t_{SLEEP} + t_{TRANS} \quad (IV.1)$$

$$\text{and } t_{SIM} = t_{SLOT} \times N \quad (IV.2)$$

Here, N is total number of slots during time t_{SIM}

$$\text{Again, } t_{SIM} = t_W + t_R \quad (IV.3)$$

Where t_W and t_R represent time period while IH-MAC operates in broadcast scheduling mode and link scheduling mode respectively.

Let n_H , n_{TX} , n_{RX} , n_{OH} , represents the total number of times that a node hears, transmits, receives, and overhears during t_{SIM}

A sensor node consumes energy by transmitting (e_{TX}), receiving (e_{RX}), overhearing (e_{OH}), and idle listening (e_{IDLE}) during the awake state. And during the sleep state very less energy is consumed. During transition (e_{TRANS}) from sleep state to active state energy is also consumed. Since our IH-MAC protocol operate both in broadcast scheduling and link scheduling (Rendezvous) and we have used power adjustment technique, so transmitting energy is further divided into two category, without rendezvous,

$e_{TX(W)}$ and with rendezvous, $e_{TX(R)}$. Similarly, receiving energy can be divided into $e_{RX(W)}$ and $e_{RX(R)}$. Now energy consumption during t_{SIM} can be expressed by

$$e = n_{TX(W)} \times e_{TX(W)} + n_{TX(R)} \times e_{TX(R)} + n_{RX(W)} \times e_{RX(W)} + n_{RX(R)} \times e_{RX(R)} + t_{OH} \times e_{OH} + t_{IDLE} \times e_{IDLE} + t_{SLEEP} \times e_{SLEEP} + t_{TRANS} \times e_{TRANS} \quad (IV.4)$$

Since IH-MAC has the probabtion of adjusting transmission power we use maximum transmission power as $E_{TX(max)}$

and right transmission power as, $E_{TX(right)}$.

When a sensor node transmits a packet, it sends SYNC, RTS, DATA and it receives CTS and ACK.

So, for transmitting a packet energy consumed by a transmitting node is

$$e_{TX(W)} = E_{TX(max)} \times t_{SYNC-RTS} + E_{TX(right)} \times t_{DATA} + E_{RX} \times t_{CTS} + E_{RX} \times t_{ACK} \quad (IV.5)$$

$$e_{TX(R)} = E_{TX(right)} \times t_{SYNC} + E_{TX(right)} \times t_{DATA} + E_{RX} \times t_{ACK} \quad (IV.6)$$

Where $t_{SYNC-RTS}$, t_{DATA} , t_{CTS} and t_{ACK} are required time to send SYNC-RTS, DATA, and to receive CTS and ACK, respectively.

Now, when a sensor node receives a packet, it receives SYNC, RTS, DATA and it sends CTS and ACK.

So, for receiving a packet energy consumed by receiving node is

$$e_{RX(W)} = E_{RX} \times t_{SYNC-RTS} + E_{RX} \times t_{DATA} + E_{TX(max)} \times t_{CTS} + E_{TX(right)} \times t_{ACK} \quad (IV.7)$$

$$e_{RX(R)} = E_{RX} \times t_{SYNC} + E_{RX} \times t_{DATA} + E_{TX(right)} \times t_{ACK} \quad (IV.8)$$

Now, let the sensor nodes Poisson arrival rate of transmitting packet is μ_{TX} and sensor nodes Poisson arrival rate of receiving packet is μ_{RX} during time t_{SIM} . So, the number of times the sensor node transmits and receives packet during t_{SIM} is

$$n_{TX(W)} = \mu_{TX} \times t_{SIM(W)} \quad (IV.09)$$

$$n_{TX(R)} = \mu_{TX} \times t_{SIM(R)} \quad (IV.10)$$

Similarly,

$$n_{RX(W)} = \mu_{RX} \times t_{SIM(W)} \quad (IV.11)$$

$$n_{RX(R)} = \mu_{RX} \times t_{SIM(R)} \quad (IV.12)$$

The overhearing of packets and idle listening occur during listen interval. So,

$$t_{OH} = n_{OH(SYNC-RTS)} \times t_{SYNC-RTS} + n_{OH(CTS)} \times t_{CTS} \quad (IV.11)$$

$$\text{and } n_{OH} = n_H - n_{RX} \quad (IV.12)$$

$$t_{IDLE} = d \times t_W - n_{TX(W)} \times (t_{SYNC-RTS} + t_{CTS}) - n_{RX(W)} \times (t_{SYNC-RTS} + t_{CTS}) - t_{OH} \quad (IV.13)$$

The transition from sleep mode to active mode will occur in every slot. So,

$$t_{TRANS} = N \times t_{SA} \quad (IV.14)$$

Where t_{SA} represents the time required for switching radio from sleep mode to active mode.

So, the energy consumption of a sensor node can be computed analytically using the equation (IV.4)

Now, we also develop energy consumption analytical model of S-MAC, one of the fundamental MAC protocol for sensor network, to compare with IH-MAC. In fact S-MAC protocol is the most popular general purpose MAC protocol specially designed for wireless sensor network. For S-MAC the total simulation time, t_{SIM} can be expressed as

$$t_{SIM} = t_{TX} + t_{RX} + t_{OH} + t_{IDLE} + t_{SLEEP} + t_{TRANS} \quad (IV.15)$$

$$\text{And } t_{SIM} = t_{SLOT} \times N \quad (IV.16)$$

S-MAC protocol operates like broadcast scheduling and no power adjustment technique is used. Therefore energy consumption during t_{SIM} can be expressed as

$$e = n_{TX} \times e_{TX} + n_{RX} \times e_{RX} + t_{OH} \times e_{OH} + t_{IDLE} \times e_{IDLE} + t_{SLEEP} \times e_{SLEEP} + t_{TRANS} \times e_{TRANS} \quad (IV.17)$$

When a node transmits a packet, it sends SYNC, RTS, DATA and it receives CTS and ACK.

So, for transmitting a packet energy consumed by transmitting node is

$$e_{TX} = E_{TX} \times (t_{SYNC-RTS} + t_{DATA}) + E_{RX} \times (t_{CTS} + t_{ACK}) \quad (IV.18)$$

Now, when a node receives a packet, it receives SYNC, RTS, DATA, and sends CTS and ACK.

So, for receiving packet energy consumed by a receiving node is

$$e_{RX} = E_{RX} \times (t_{SYNC-RTS} + t_{DATA}) + E_{TX} \times (t_{CTS} + t_{ACK}) \quad (IV.19)$$

Let, sensor nodes Poisson arrival rate of transmitting packet and receiving packet during the time t_{SIM} are same as before.

So, the number of times the node transmits and receives packet during t_{SIM} is

$$n_{TX} = \mu_{TX} \times t_{SIM} \quad (IV.20)$$

Similarly,

$$n_{RX} = \mu_{RX} \times t_{SIM} \quad (IV.21)$$

The overhearing of packets and idle listening occur during listen interval. So,

$$t_{OH} = n_{OH(SYNC-RTS)} \times t_{SYNC-RTS} + n_{OH(CTS)} \times t_{CTS} \quad (IV.22)$$

$$\text{and } n_{OH} = n_H - n_{RX} \quad (IV.23)$$

$$t_{IDLE} = d \times t_{SIM} - n_{TX} \times (t_{SYNC-RTS} + t_{CTS}) - n_{RX} \times (t_{SYNC-RTS} + t_{CTS}) - t_{OH} \quad (IV.24)$$

The transition from sleep to active mode will occur in every slot. So,

$$t_{TRANS} = N \times t_{SA} \quad (IV.25)$$

So, the energy consumption of a sensor node of S-MAC can be computed analytically using equation (IV.17). For simplicity we avoid considering the collision both for S-MAC and IH-MAC in our analytical model.

Now, if we compare equation (IV.5) & (IV.6) with the equation (IV.18) we see that the consumed power for a packet transmission for the source node is less in IH-MAC than S-MAC. Similarly, if we compare equation (IV.7) & (IV.8) with equation (IV.19) we see that the consumed power for a packet reception for the destination node is less in IH-MAC than S-MAC. Finally if we put these value in equation (IV.4) and (IV.17) we can conclude that the IH-MAC is more energy efficient than S-MAC.

5. RESULTS

In this section, we investigate the performance of the proposed IH-MAC protocol. In the simulation setup, we take 100 nodes distributed in a 10 m×10 m area grid. The nodes are static and the radio range is chosen so that all the non-edge nodes have eight neighbors. The sink node is chosen on the bottom right corner of the network grid. The duty cycle is chosen 15 percent both for IH-MAC and S-MAC. The results are averaged over several simulation runs. The performance metrics used in the evaluation of IH-MAC is energy consumption per bit. We compare the performance of our proposed IH-MAC protocol with the standard S-MAC protocol. The parameters used for simulation are listed in Table I.

Energy efficiency of sensor nodes for IH-MAC and S-MAC is shown in figure 6. We vary the packet generation interval

from 1 to 10 seconds. We see that energy consumption per bit of IH-MAC is less than the energy consumption of S-MAC when traffic is heavy.

Table 1

Parameters for the MAC protocol

Parameter value	Name
Channel bandwidth	20 kbps
Data packet length	20 bytes
Transmission power	36 mW
Receive power	14.4 mW
Idle power	14.4 mW
Sleep state	15 μ W
Frame length	1 sec
Threshold value for the buffer size (for IH-MAC)	3 packet
Duty cycle	15 %

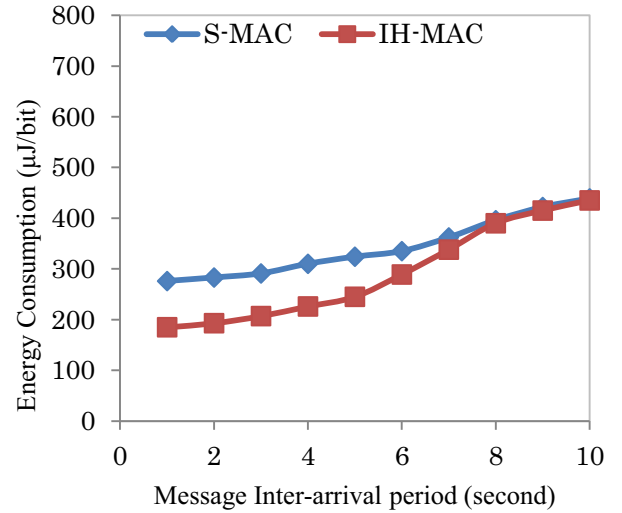


Figure 6. Energy consumed by IH-MAC and S-MAC with varying traffic rate.

It is because during heavy traffic IH-MAC protocol makes some rendezvous slots. Energy consumption in rendezvous slot is less than the energy consumption of a slot of S-MAC as explained in section 3.3 of this paper. But as traffic declines IH-MAC cannot create frequent rendezvous slot, hence its energy efficiency deteriorates. From the figure 6, we see that when message inter arrival period is 1 to 5 second, energy efficiency of IH-MAC is better than the energy efficiency of S-MAC but when the message inter arrival period increases, the performance of IH-MAC and S-MAC become almost equal. In the worst case scenario, while very light traffic or no traffic at all, the energy consumption of IH-MAC will be at least equal to the S-MAC. For simplicity, we have not considered all features of IH-MAC, like transmission power adjustment.

Nevertheless IH-MAC shows its energy efficiency evidently.

6. CONCLUSION AND FUTURE WORK

This paper presents IH-MAC; a novel energy efficient MAC protocol for WSNs. IH-MAC introduces the idea to combine the strength of contention based and schedule based approach of medium access control to achieve significant amount of energy savings. The transmission power adjustment feature of IH-MAC is very promising and will certainly contribute to enhance sensor nodes lifetime.

Since, we could not implement yet all features of IH-MAC, as a future work, we expect more detail result about energy efficiency, throughput and fairness issue of our proposed IH-MAC protocol.

At present there is no unique standard of MAC protocol of WSNs. Therefore, an early initiatives of ITU will significantly contribute in the maturation process of the of WSNs technology.

REFERENCES

- [1] I. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "A survey on sensor networks," *IEEE Commun. Mag.* pp.102-114, August,2002.
- [2] W. Ye, J. Heidemann, and D. Estrin, "Medium Access Control with Coordinated Adaptive Sleeping for wireless sensor networks," in *Proc. IEEE/ACM Transaction on Networking*, vol.12, No.3, June 2004.
- [3] A. Keshavarzian, H. Lee, and L. Venkatraman, "Wake up scheduling in wireless sensor networks," in *Proc.Of ACM, MobiHoc2006*.
- [4] E. Arıkan, "Some complexity results about packet radio networks," *IEEE Transaction on Information Theory*, vol. 30. no.4, pp.681-685,1984.
- [5] T. Van Dam and K. Langendoen, "An adaptive energy-efficient MAC protocol for wireless sensor networks," in *Proc. ACM SenSys*, New York, 2003, pp.171-180.
- [6] Sang Hoon Lee, Joon Ho Park, Choi.L, "AMAC : Traffic-Adaptive Sensor Network MAC protocol through Variable Duty-Cycle Operations ," in the *Proc. ICC2007*, pp.3259-3264.
- [7] R. Yadav, S. Varma, and N. Malaviya, "Performance Analysis of Optimized Medium Access Control for Wireless Sensor Networks" *IEEE Sensors Journal*, Vol.10, No.12, December.2002.
- [8] J. Polstre, J. Hill, and D. Culler, "Versatile low power media access for wireless sensor networks," in *Proc. ACM SenSys*, New York, 2004,pp.95-107.
- [9] I. Rhee, A. Warriar, M. Aia, and J. Min, "Z-MAC: A hybrid MAC for wireless sensor networks," in *SenSys'05*, Nov 2-4, 2005.
- [10] I.Rhee, Warriar, J.Min, and L.Xu, "DRAND: Distributed randomized TDMA scheduling for wireless ad hoc networks," in *Proc. ACM MobiHoc*, New York, 2006, pp.190-201.
- [11] S. Ramanathan, "A unified framework and algorithm for (T/F/C) DMA channel assignment in wireless networks," in *Proc. IEEE INFOCOM*, 1997, pp.900-907.
- [12] S. Li, D. Qian, Y. Liu, J. Tong, "Adaptive distributed randomized TDMA scheduling for clustered wireless sensor networks," in *Proc. of IEEE WiCOM*, pp.2688-2691,2007.
- [13] LAN MAN Standards Committee of the IEEE Computer Society. *IEEE Std 802.11-1999, Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications*.IEEE,1999.
- [14] J.Gomez, A.T.Campbell, M.Naghshineh, and C.Bisdikian, "Conserving Transmission Power in Wireless Ad Hoc Networks," *Proceedings of ICNP2001*. pp.11-14,November,2001.

ROUTE OPTIMIZATION BASED ON THE DETECTION OF TRIANGLE INEQUALITY VIOLATIONS

Papa Ousmane Sangharé and Bamba Gueye and Ibrahima Niang

Université Cheikh Anta Diop de Dakar, Senegal
papa.sanghare@ucad.edu.sn, bamba.gueye@ucad.edu.sn, iniang@ucad.sn

ABSTRACT

During the last decade, new services networks and distributed applications have emerged. These systems are flexible insofar as they can choose their ways of communication among so much of others. However, this choice of routing is based on a large number of measurements of times (Round Time Trip (RTT)) which are sources of overload in the network. Network Coordinate Systems (NCS) allow to reduce measurements overhead by mitigating direct measurements. However, NCS encounter inaccuracies with respect to distance prediction, when the measured distances violate the principle of the triangular inequality (TIV-Triangle Inequality Violation).

Firstly, we propose a new metric, called “RPMO”, which is based on the Ratio of Prediction and the Average Oscillations of the estimated distances, to detect the potential TIVs. The obtained results show that the “RPMO” metric gives better performance compared to metrics presented in former work. Secondly, we propose to use the existence of TIVs to optimize the routing in Overlay Network. To achieve this goal, we present a new approach that enables to detect the best shortened paths offered by the existence of potential TIVs.

Keywords— Network Coordinate Systems, Triangular Inequality Violation, Overlay Routing

1. INTRODUCTION

Nowadays, Network Coordinate Systems are widely used in network applications and services on a large scale and globally distributed applications such as file sharing peer to peer [1], nearest server selection [2], online games [3] etc.

Indeed, Network Coordinate Systems (NCS) [4, 5, 6, 7] allow hosts on the network to estimate the time between them without making measurements, and thus reduce resource consumption and particularly the number of measures on demand.

The main idea of NCS is to model the Internet as a geometric space, and characterize the position of each node in the network by a set of coordinates. Therefore, the latency between two nodes in the network is thus estimated as the geometric distance between their coordinates in this geometric space. Indeed, explicit measures are no longer needed.

Nevertheless, network policies routing [8] can break down the principle of triangle inequality. These violations are the cause of distortions and prediction errors for coordinate systems [9]. Let's assume three nodes A , B , and C such that $d(A, B)$ is $36ms$, $d(B, C)$ is $16ms$ and $d(A, C)$ is $9ms$, where $d(XY)$ denotes the delay between node X and node Y . In this case, the principle of triangle inequality is violated because $d(A, B) > d(A, C) + d(C, B)$.

In such case, the triangle ABC is a TIV (*Triangle Inequality Violation*) and AB (the longest side) is a **TIV-base**. TIV-base means that it exists a potential shortcut that can be used to reduce the distance between these two nodes that form this link considered as TIV-base. Since the principle of triangle inequality should be respected in any metric space, finding “good” coordinates in order to obtain an accurate estimation of delay between each pair of nodes will be impossible. In the presence of TIVs, node's coordinates will tend to alternate between sub-estimates and over-estimates the actual distance, without ever managing to position themselves in the metric space so perfect [9, 10, 11].

In order to exploit the coordinate systems for various operations of prediction distances it is mandatory that NCS give accurate and stable coordinates. Since the presence of TIV leads to inaccuracies with respect to the prediction delays, some researchers in this field have proposed TIVs detection techniques [10, 12] to allow nodes located in the systems to avoid links that are TIV-bases. In so doing, these nodes enhance their accuracy following the distance estimates.

However, the presence of TIVs in the Internet offers an opportunity that can be exploited to improve routing distribution applications online games, file sharing, or VoIP [3]. These applications can potentially improve their performance routing, using the shortcuts provided by TIVs [13].

Firstly, we introduce our TIV detection metric called *Ratio of Average Prediction on the oscillations* (“RPMO”), which detects TIVs accurately, while solving the shortcomings of previous works [12, 10].

Secondly, since TIVs are inherent to Internet, based on the their existence, we propose to optimize overlay network routing by using the *MDGD* metric (*Metric for Detecting Good Detours*) to detect the best shortcuts offered by TIVs.

The paper is organized as follows. Section 2 describes the different Network Coordinate Systems proposed in the related work ; in addition we present the previous metrics

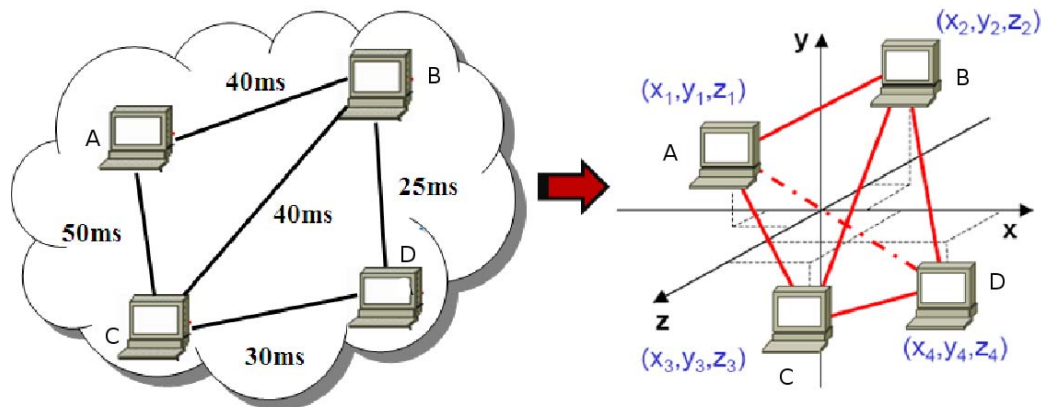


Figure 1. Geometric space model of the Internet.

used for detecting TIVs, i.e. the ratio of prediction [10] and OREE [12]. In section 3, we present and evaluate our proposed metric RPMO. In Section 4, we propose a new approach for optimizing Overlay routing by the use of triangle inequality violation in NCS. Finally, we conclude and present some research perspectives in Section 5.

2. RELATED WORK

In order to achieve the objectives of optimizing performance and scalability of network applications, several approaches for predicting network distances (propagation delay and transmission round trip “RTT”) based on coordinates have been proposed [6, 4, 5, 14, 7]. The main idea of such systems is to model the Internet as a geometric space. Consider the example shown in Figure 1, where we have four nodes (A, B, C, D) illustrated in a three-dimensional geometric space after an embedding from a given network.

Therefore, the distance between two nodes in the network is predicted as the distance between their coordinates, without making explicit measurements. In other words, if node A knows the coordinates (x, y, z) of node D , A will not need to make an explicit measure to determine the RTT towards D , instead, A computes the distance between itself and the node D in the coordinate space. The obtained distance represents the prediction of the RTT between A and D . It should be noted that until a precise location and reasonable for a node can be obtained with little overhead, much of the cost of distance measurements by sampling can be eliminated.

From the literature, Network Coordinate Systems can be splitted in two categories :

- *Centralized Coordinate Systems* : they involve a central component (a set of hosts called either Landmarks, or beacons, or Lighthouses) [6], from which other nodes calculate their own coordinates, according to the fixed infrastructure of measurements. We can give as examples *GNP* [4], *NPS* [5].
- *Distributed Coordinate Systems* : these systems general-

ize the role of landmarks in all nodes in the system, or eliminate the landmarks infrastructure. Decentralized Coordinate Systems can be seen as a peer-to-peer system. For instance we can cite *BBS* [14], *Vivaldi* [7].

2.1. Vivaldi overview

Vivaldi [7] is a decentralized coordinates system in which each node computes its own coordinates by making measurements with a small number of other nodes called its neighbors that are heterogeneous (half close and half away). Each time a node takes a measurement with one of his neighbors ; it compares the estimated delay measured by using their coordinates and modifies its position in space so as to move toward or away from its neighbor.

If a given node i wants to update its own coordinates towards a given neighbor j it needs a sample. This sample is formed by RTT_{ij} which is the RTT measured between i and j , neighbor’s coordinates x_j , and the confidence error e_j [7]. Let assume that $EST_{ij} = \|x_j - x_i\|$ represents the estimated RTT between nodes i and j based on their coordinates.

Following the algorithm proposed in [7], firstly node i computes the weight w of its sample. We have $w = \frac{e_i}{(e_i + e_j)}$. Afterwards, it uses this weight to update its local error $e_i = e_j \times w + e_i \times (1 - w)$ and then computes a value of $\delta = C_c \times w$ (with $0 < C_c < 1$). The goal of δ is to evaluate the amplitude of the displacement of the node. Finally, node i updates its coordinates as follows :

$$x_i = x_i + \delta \times (RTT_{ij} - EST_{ij}) \times u(x_i - x_j) \quad (1)$$

where $u(x_i - x_j)$ is a unit vector that indicates the direction of node i with respect to its replacement. For more details about equation 1 please refer to [7].

2.2. Metrics for detecting TIV

Previous works have proposed two metrics for detecting TIV. The first one is called “*Ratio of prediction*” [10] and the later

“OREE” [12]. The metric OREE is based on the oscillations of a given node and the relative error estimation, whereas the ratio of prediction represents the relationship between the estimated distance and the measured (actual) distance.

The authors of [10] have shown that the sides of the triangle that have a small ratio of prediction, *i.e.*, the narrowed sides according to the Euclidean space, tend to cause severe TIVs. However, the Ratio of prediction presents some issues with respect to the node’s neighbors update. Indeed, each node belonging to the network periodically chooses 32 other neighbors at random, it adds to its 32 neighbors already available. The 64 neighbors are sorted according to the value of their prediction ratio.

If the ratio of prediction of a link is very small, this implies that the link is probably underestimated due to the existence of severe TIVs [10]. Subsequently, the node removes from its list of neighbors, the 32 nodes with the smallest ratio of prediction. Quite often, these neighbors are those that are generally far from this given node.

After having removed these 32 neighbors, the given node keeps as neighbors the remaining 32 nodes (the nearest) as neighbors for the next iteration. This set of neighbors is not suitable for Vivaldi algorithm according to [7].

The metric OREE involves the variance of the estimated distances, the distance measured and the mean estimated distances. The authors of OREE [12] have shown that when OREE’s value is small the link can be considered as a TIV-base, and vice versa. This means that the probability that a link is a TIV-base increases when the value of OREE decreases.

The main drawback of OREE is that it uses a huge amount of information for detecting TIVs. In fact, we should keep node’s coordinates of previous rounds of measurement. Therefore, OREE is not scalable in large network such as Internet. It causes a considerable computing time, leading poor performance of peer-to-peer hosts that aim to determine the best path as quickly as possible (e.g. online applications gaming or VoIP [3]).

Note that, Kawahara *et al.* in [15] propose to find quality overlay routes between node pairs based on TIV optimization according to the latency and packet loss ratio metrics. It is worth noticing that they do not propose a mechanism to detect TIVs.

3. TIV’S DETECTION BASED ON RPMO METRIC

To overcome the limitation of previous works [12, 10], we propose in this section a new metric that allows us to take into account the ratio of prediction as well node’s oscillations in the network.

3.1. RPMO (Ratio of Prediction on Average Oscillations)

Our goal is to find a metric that allows us to detect TIVs accurately without altering the heterogeneous selection of neighbors according to Vivaldi, and using less computation over-

head. Our proposed metric, called RPMO, takes into account three parameters (the oscillations, the estimated distance and the actual distance) in order to detect if a link can be considered as a potential TIV-base.

$$RPMO = \frac{Estimated\ distance}{RTT} \times \frac{1}{Average\ oscillations} \quad (2)$$

By definition, a tick represents a round where a given node update its own coordinates once. An oscillation is the difference of estimated distances of two successive ticks.

For instance, let assume that d_1 is the estimated distance of AB during the first tick (tick 1), d_2 is the estimated distance of AB during the second tick (tick 2), and d_3 is the estimated distance of AB during the third tick (tick 3). Therefore, the average oscillations between these three rounds can be computed as follows :

$$Average\ oscillations = \frac{(|d_1 - d_2| + |d_2 - d_3|)}{2}$$

Therefore, the RPMO value is obtained by

$$RPMO = \frac{d_n}{RTT} \times \frac{(n-1)}{\sum_{i=1}^n |d_i - d_{i+1}|} \quad (3)$$

3.2. Experimental setup

To evaluate the RPMO metric, we used the *P2Psim* discrete-event simulator [16] which provides an implementation of Vivaldi. During our simulations, each Vivaldi node has 32 neighbors and the results are obtained for a 9-dimensional Euclidean space. The constant C_c is set at 0.25 as recommended in [7].

In order to evaluate the RPMO metric, we used three matrices delays as datasets : *P2Psim King* dataset (1740 nodes) [16], *Meridian* dataset (2500 nodes) [17] and the *PlanetLab* dataset (180 nodes) [18].

King and *Meridian* dataset are obtained following the *King* measurement technique [19] which is similar to ping in the sense that it estimates the latency between arbitrary end nodes using recursive DNS queries. The third matrix, which we call *PlanetLab* dataset, is a matrix delay constructed by performing ping measurements between 180 *PlanetLab* nodes [18] distributed around the world.

To study the characteristics of TIV, two criteria have been defined to indicate the severity of TIV : the *absolute severity* and the *relative severity*.

The absolute severity is computed as follows :

$$Ga = d(A, C) - (d(A, B) + d(B, C)) \quad (4)$$

The relative severity is obtained by

$$Gr = \frac{d(A, C) - (d(A, B) + d(B, C))}{d(A, B)} \quad (5)$$

These criteria reflect the potential gain that can be achieved by detecting the existing TIVs in the network. A gain equals

to $G_a = 10ms$ illustrates that instead of going through the direct path from A to B , going through the path via node C allows us to gain $10ms$.

However, a large G_a and G_r do not show only severe violations, but also a possible gain. In our work, we are interested in TIVs that meet both criteria, namely $G_a > 10ms$ and $G_r > 0.1$. Indeed, TIVs offering shortcuts that allow a gain less than $10ms$ are not very interesting.

3.3. Evaluation and Results

To study the performance of these different TIV detection metrics (RPMO, Prediction Ratio, and OREE), we take into account a comparison of their Receiver Operating Characteristic (ROC) curves. Therefore, we use the classical false/true positive/negative indicators. A *true positive* (TPR - True Positive Rate) is a TIV-base, which should therefore be suspected by the test. A *false positive* (FPR - False Positive Rate) is a non TIV-base that has been wrongly suspected by the test.

Figure 2 illustrates ROC curves obtained following different TIV detection metrics such as Ratio Prediction, OREE, and RPMO by considering the King dataset. It should be noted that the *Ratio-Pred* as depicted in Figure 2 refers to the "Ratio of Prediction" metric. Each point on the ROC curves (Figure 2) determines the TPR along the y-axis and the FPR along the x-axis obtained with a given detection threshold. During our simulations, we take different threshold values that range from 0.5 to 9 by step of 0.5. It should be noted that for a ROC curve, more the curve is near to the top left the corner of the graph, better is the detection.

The value 0.3 labelled in Figure 2 represents a given threshold value that gives better results among the different threshold values (0.5 to 9) that we used during our simulations. For instance, according to RPMO metric in Figure 2, a percentage detection of 59% of TIV-base with 17% of FPR corresponds to a given threshold value fixed to 0.3. This threshold value gives a better tradeoff.

In fact, the RPMO and Ratio-Pred metric have the same trend. For FPR smaller than 0.2, the RPMO metric outperforms the Ratio-Pred metric. Nevertheless, OREE is less efficient with respect to both metrics RPMO and Ratio-Pred.

Figure 3 shows the ROC curves obtained following Meridian dataset. The general trend one can observe, compared to Figure 2, is the fact that we have higher TPR detection with respect to same FPR (eg., 11%). According to RPMO metric, the threshold value that gives high TPR (88%) with low FPR (11%) is 0.3 (Figure 3).

The main reason is due to the fact that we have more links that are TIV-bases in Meridian dataset with respect to King and PlanetLab datasets. Following our three datasets, the computed values of links that are TIV-bases are estimated to 23%, 42%, and 9% for King, Meridian, and PlanetLab datasets respectively. We recall that a TIV-base is a link where it is possible to find a shortcut in the overall system.

Figure 4 also illustrates the ROC curves according to Planet-

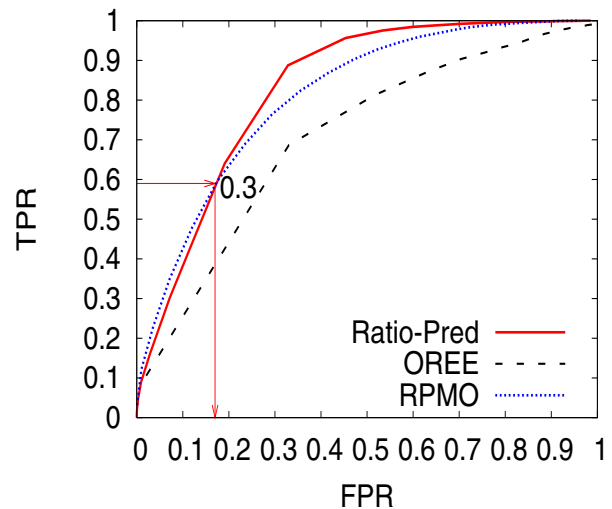


Figure 2. King dataset : Comparison between RPMO, Prediction Ratio and OREE metrics.

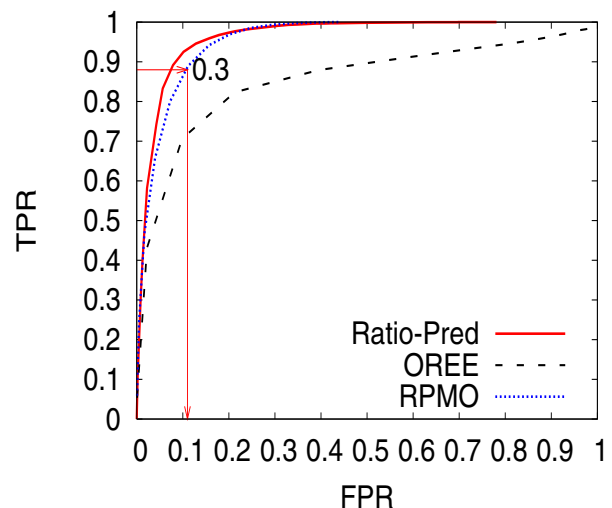


Figure 3. Meridian dataset : Comparison between RPMO, Prediction Ratio and OREE metrics.

Lab dataset with respect to our three studied metrics.

Following the PlanetLab dataset, the best threshold value with respect to RPMO metric is 0.65 with a TPR and a FPR equals to 55% and 22% respectively.

As summary, based on Figure 2 we remark that for a TPR values up to 60%, the RPMO metric is better compared to OREE and the ratio of prediction ; on the other hand for TPR values upper than 60% , the ratio of prediction becomes a little bit better than RPMO with a FPR upper than 30%. Following the Meridian dataset as illustrated on Figure 3, the gap is reduced between the ratio of prediction and RPMO. The same trend is also observed according to PlanetLab dataset (Figure 4).

It appears clearly that our TIV detection metric, called

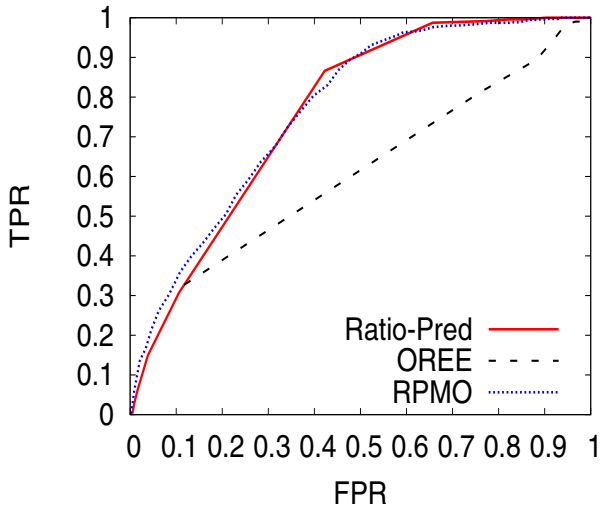


Figure 4. PlanetLab dataset : Comparison between RPMO, Prediction Ratio and OREE metrics.

RPMO, is more efficient compared to OREE metric by considering all datasets (see Figure 2, Figure 3 and Figure 4). It should be noted that with respect to the ratio of prediction metric the gap is reduced, and roughly we observe the same trend.

Nevertheless, the ratio of prediction presents several drawbacks according to the selection mechanism of node’s neighbors. The prediction ratio tends to select only nearest neighbors that is not suitable for Vivaldi algorithm [7, 6]. We argue that the metric RPMO is most suitable for detection TIV when we use a distributed coordinate system like Vivaldi.

4. OPTIMIZATION OF ROUTING IN THE OVERLAY NETWORK THROUGH TIVS DETECTION

As TIVs are inherent to Internet, they represent an opportunity that can be exploited for routing in overlay networks. In fact, multimedia, peer-to-peer file sharing, online games, distribution applications, or VoIP [3] require quality of service guarantees in term of delay. Therefore, these applications can potentially improve their performance by exploiting a TIV-based routing approach.

By definition, we recall that if the side AB of a “bad triangle” ABC (triangle where the triangle inequality is not respected) is a TIV-base, it exists a shortcut, for instance via node C , to get towards B from A instead of using the direct path (AB). In such case, applications can use the shortcut to gain more time. Our goal is to detect for each link TIV-base, for instance AB , the best C_i points that allow to gain more time from A towards B (i.e., $d(A, C_i) + d(C_i, B) < d(A, B)$).

4.1. Clustering approach

Clustering is a technique used to group elements with similar characteristics. Therefore, the idea is for each link TIV-base (eg., AB), to cluster potential nodes from a given diameter that can be considered as shortcuts with respect to the link TIV-base. In so doing, we reduce the number of shortcuts that will be evaluated in order to find the best one. As well, we can remove those shortcuts that are not clustered (*outliers*) in the set of shortcuts where we should seek good shortcuts.

To achieve this clustering, we used the “*QT_Clustering*” algorithm [20]. This algorithm has been initially proposed by Heyer et al. for genetic sequence clustering. It is based on the unique constraint of the cluster diameter, as a user-defined parameter. The cluster diameter represents the maximal distance existing among any two members of the cluster.

The main idea behind this clustering algorithm is to find the best shortcuts as well the shortcuts that are able to give the same gain of time. Indeed, these shortcuts should share the same cluster. We hope that these best shortcuts will be given by shortcuts that are not clustered (*outliers*) by the use of *QT_Clustering* algorithm.

For our simulations, we choose a diameter of $30ms$ for clustering the potential “good detour” (shortcut) obtained with respect to different links that are TIV-base. The result is shown in Figure 5.

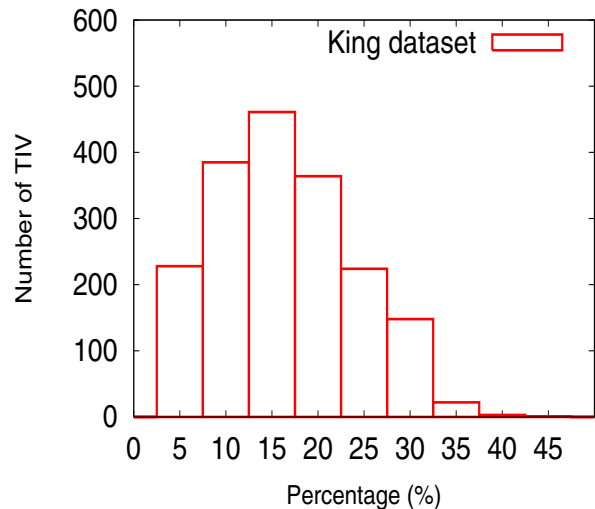


Figure 5. King dataset : Proportion of outliers (shortcuts) that represent the best shortcuts for a given TIV-base.

Figure 5 shows the percentage of shortcuts that are among the best shortcuts following a set of shortcuts obtained with respect to a given TIV-base. Note that, here the considered shortcuts are those that are not clustered by the *QT_Clustering* algorithm. The y-axis represents the number of TIV-base considered and the x-axis represents the percentage of shortcuts that are among the best shortcuts. It is worth noticing that the best shortcuts are those that enable to gain more time with respect to a given link considered as

TIV-base.

In Figure 5 we remark that the percentage of outliers that are among the best shortcuts varies between 5 and 45% with respect to all shortcuts for a given TIV-base. Note that a given box (Figure 5) can be seen as a bin where all TIV-bases give the possibility to find the same percentage of shortcuts that are among the best shortcuts. For instance, we can see that less than 10 TIV-bases have 35% of their shortcuts, considered as outliers, that are among the best shortcuts for each fixed TIV-base. Here, we recall that the considered shortcuts are those that do not belong to any cluster after having executed the QT_clustering algorithm.

The obtained results, by considering the shortcuts that are characterized as outliers, do not give high detection of best shortcuts.

Since we can find the center of each cluster, called *Cluster Head* (CH), we would like to seek if a cluster head can be considered as the best shortcut according to all other cluster heads that own the remaining clusters. In so doing, we should rank the different cluster head following the gain that they can offer as shortcut. The first one, after their ranking, is considered as the best cluster head. By definition, the “*Best Cluster*” is the one that is owned by the best cluster head. Put simply, we hope that all shortcuts that belong to this cluster offer a good shortcut.

With the QT_Clustering algorithm, each cluster has a cluster head that represents the center of the cluster. For a given TIV-base (eg., (AB)), we rank the different cluster heads with respect to the amount of time that they can offer. In other words we sort the different CH_i following their gain $1 \leq i \leq \text{number of cluster}$. Furthermore, we consider the members of the best cluster and seek their percentage among the best shortcuts that exist for this given TIV-base.

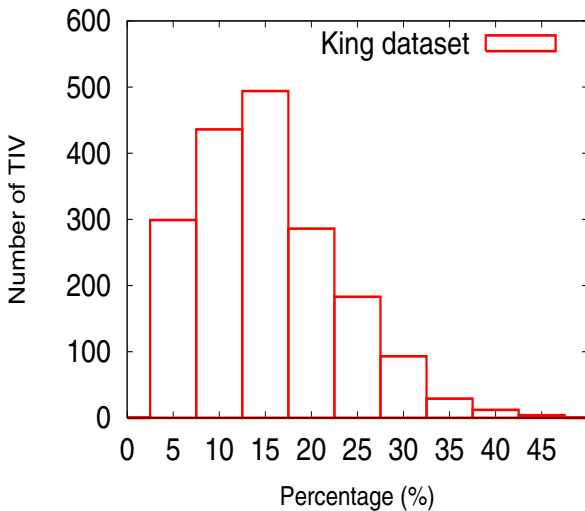


Figure 6. King dataset : Proportion of shortcuts that represent the best shortcuts following the “Best Cluster”.

The obtained results are shown on Figure 6. The y-axis represents the number of TIV-base considered and the x-axis rep-

resents the percentage of shortcuts that are among the best shortcuts with respect to the Best Cluster. This second approach, gives the same trend i.e. the percentage of shortcuts which are located in the Best Cluster and are considered as best shortcuts varies between 5 and 45%.

Based on the results illustrated in Figure 5 and Figure 6, we can conclude that the clustering approach does not allow a good detection of best shortcuts.

4.2. MDGD (Metric for Detecting Good Detours) approach

Since the previous method (clustering) does not help to find the best shortcuts, we focus on a new approach. The goal is to find a metric that enables to say whether any potential shortcut is part of the best shortcut (i.e, a shortcut with a gain of time upper than 10ms).

Therefore, we investigate the possible relation between the distance D' which is equal to $d(AB) - (d(AC) + d(CB))$, where $d(AB)$ represents the RTT between A and B . Note that D' is obtained based on actual distance (RTT delay). The *pseudo gain* for a triangle ABC represents the difference between the RTT distance ($d(AB)$) and the sum of estimated distances of links AC and CB , namely : $d(AB) - (Estimate(AC) + Estimate(CB))$.

We put the triangles in bin of 10ms based on their pseudo gain. In each bin, we calculate the minimum, the median, and the maximum distance of the distance D' of triangles present in the bin. We illustrate these three metrics in Figure 7, where on the x-axis we have the pseudo gain in milli second (ms) and on the y-axis the severity of the TIVs . The curve of the median distance D' of triangles shows that more and more that the pseudo gain increases, we are dealing with triangles TIV-bases (i.e a triangle that violates the principle of triangle inequality), which increasingly are becoming more severe (offering gains increasingly large).

It is worth noticing the negative values along the y-axis means that the triangle is not a TIV. In so doing, when we consider the minimum distance of D' , we can see that all triangles do not violate the principle of triangle inequality violation.

Furthermore, we consider these metrics in order to figure out our MDGD approach :

- The relative estimation error (Er) : $d(AB) / Estimate(AB)$
- The absolute estimation error (Ea) : $d(AB) - Estimate(AB)$
- Pseudo gain (PG) : $d(AB) - (Estimate(AC) + Estimate(CB))$

The pseudo gain can help us to find severe links that are TIV-base. Based on the metrics Er , Ea , and PG we propose the MDGD metric that allows to find the best shortcuts (gain upper than 10ms). The MDGD metric is described as follows :

$$MDGD = \frac{(Er \times Ea)}{PG} \quad (6)$$

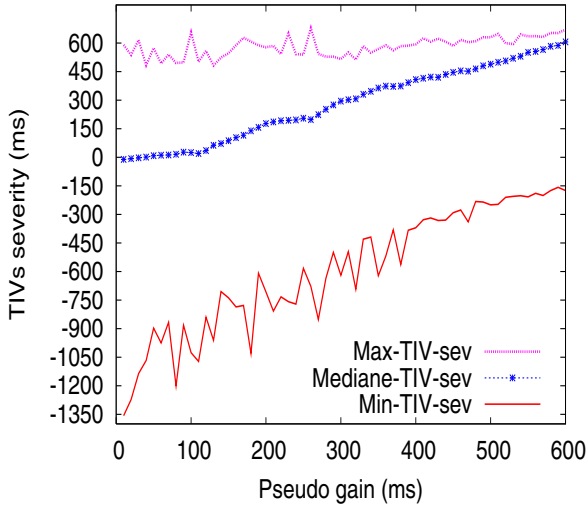


Figure 7. King dataset : Relationship between pseudo gain and TIVs severity.

4.3. Evaluation of MDGD approach

To study the effectiveness of this metric, our goal is to find the threshold value that allows to find the maximum number of “good shortcuts”. In such case, we rely on the TPR (True Positive Rate) and the FPR (False Positive Rate) according to each threshold value.

The TPR represents the percentage of shortcuts that are detected as well provide a gain of time upper than 10ms. The FPR represents the percentage of shortcuts that are wrongly detected as giving a shortcut upper than 10ms. Figure 8 illustrates the obtained results.

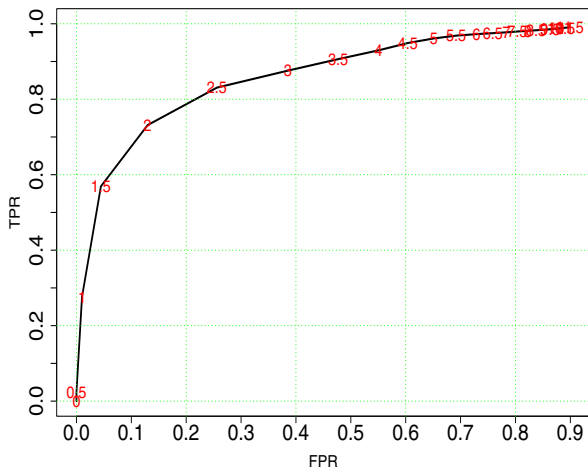


Figure 8. King dataset :ROC curve of the MDGD metric.

It should be noted that for a ROC curve, more the curve is near to the top left the corner of the graph, better is the detection. Based on Figure 8 we can notice that these following

thresholds (1.5, 2, 2.5, 3) exhibit this propriety.

To determine the best threshold that give us the best shortcuts with good accuracy, we compute the accuracy of the following thresholds 1.5, 2, 2.5, 3. By definition, the accuracy (ACC) (Table 1) represents the veracity of the classification and it is estimated as follows :

$$ACC = \frac{TP + TN}{P + N} \quad (7)$$

where TP and TN represents the number of true positive and true negative respectively. It is worth noticing that P and N represents the number of positive and negative respectively. Therefore, P expresses the total number of detours that give a gain upper than 10ms. In contrast, N represents the total number of detour that give a gain lower than 10ms.

Table 1. Evaluation of MDGD metric

Threshold	TPR	FPR	Accuracy (ACC)
1.5	0.57	0.04	0.68
2	0.73	0.13	0.77
2.5	0.83	0.26	0.81
3	0.88	0.38	0.80

Table 1 shows that a threshold value equals to 2.5 gives 83% of true positive whereas we have 26% of false positive. The threshold value equals to 2.5 gives the best accuracy.

It should be noted that it is very difficult to detect the best shortcuts with the use of clustering approach. The fact that potential shortcuts are clustered or are outliers, does not justify that they share the same characteristic. Nevertheless, with our MDGD approach, we could find the good nodes that offer shortcuts with a gain of time upper than 10ms, and with an accuracy of 81% (Table 1).

5. CONCLUSION

In this paper, we proposed a new metric called RPMO that enables Network Coordinate Systems to avoid the existence of TIV. We have shown that the RPMO outperforms OREE metric and presents the same trend with respect to the Prediction Ratio metric.

Although the TIVs are harmful to Network Coordinate Systems, they present opportunities to improve routing in overlay networks. In such case, the existence of TIV can lead to overlay networks that are TIV-aware. We can reduce consequently the delay between nodes by using the shortcut that TIVs can offer.

Therefore, we propose a metric called MDGD, to detect the best shortcuts of any triangle ABC that violate the principle of triangle inequality. The obtained results obtained show that with a threshold value equal to 2.5, MDGD, has a detection accuracy of 81%.

This result present a nice opportunity for peer-to-peer applications, online games, distributed applications, and VoIP

that require quality of service guarantees in terms of delay to maintain a certain level of performance.

Note that it is difficult to find a same RPMO's threshold value that can be applied in all studied datasets. As future work, we plan to find a metric that can enable to use a same threshold for all the used datasets. We plan also to investigate other clustering algorithms.

REFERENCES

- [1] Gnutella, "Gnutella, a distributed peer-to-peer data-sharing system," <http://www9.limewire.com/developer/gnutella/protocol/0.4.pdf>.
- [2] Sylvia Ratnasamy, Mark Handley, Richard M. Karp, and Scott Shenker, "Topologically-aware overlay construction and server selection," in *INFOCOM*, 2002.
- [3] Wookyun Kho, Salman Abdul Baset, and Henning Schulzrinne, "Skype relay calls : Measurements and experiments," in *in Proc. INFOCOM'08*, 2008.
- [4] T. S. E. Ng and Hui Zhang, "Predicting Internet network distance with coordinates-based approaches," in *Proceedings.Twenty-First Annual Joint Conference of the IEEE Computer and Communications Societies*. 2002, vol. 1, pp. 170–179, IEEE.
- [5] T. S. Eugene Ng and Hui Zhang, "A network positioning system for the internet," in *Proceedings of the annual conference on USENIX Annual Technical Conference*, Berkeley, CA, USA, 2004, ATEC '04, pp. 11–11, USENIX Association.
- [6] Benoit Donnet, Bamba Gueye, and Mohamed Ali Kaafar, "A survey on network coordinates systems, design, and security," *IEEE Communications Surveys and Tutorials*, vol. 12, no. 4, pp. 488–503, December 2010.
- [7] Frank Dabek, Russ Cox, Frans Kaashoek, and Robert Morris, "Vivaldi : a decentralized network coordinate system," *SIGCOMM Comput. Commun. Rev.*, vol. 34, pp. 15–26, August 2004.
- [8] Han Zheng, Eng Keong Lua, Marcelo Pias, and Timothy G. Griffin, "Internet routing policies and round-trip-times," in *In PAM*, 2005.
- [9] Mohamed Ali Kaafar, Bamba Gueye, Francois Cantin, Guy Leduc, and Laurent Mathy, "Towards a two-tier internet coordinate system to mitigate the impact of triangle inequality violations," in *Proceedings of the 7th international IFIP-TC6 Networking Conference, Lectures Notes in Computer Science 4982*, Singapore, May 2008, pp. 397–408.
- [10] Guohui Wang, Bo Zhang, and T. S. Eugene Ng, "Towards network triangle inequality violation aware distributed systems," in *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, New York, NY, USA, 2007, IMC '07, pp. 175–188, ACM.
- [11] Francois Cantin, Bamba Gueye, Mohamed Ali Kaafar, and Guy Leduc, "A self-organized clustering scheme for overlay networks," in *IWSOS, Lectures Notes in Computer Science 5343*, Dec. 2008, pp. 59–70.
- [12] Yongjun Liao, Mohamed Ali Kaafar, Bamba Gueye, Francois Cantin, Pierre Geurts, and Guy Leduc, "Detecting triangle inequality violations in internet coordinate systems by supervised learning," in *Proceedings of the 8th International IFIP-TC6 Networking Conference, Lectures Notes in Computer Science 5550*, Aachen, Germany, May 2009, pp. 352–363.
- [13] Cristian Lumezanu, Randy Baden, Dave Levin, Neil Spring, and Bobby Bhattacharjee, "Symbiotic relationships in internet routing overlays," in *Proceedings of the 6th USENIX symposium on Networked systems design and implementation*, Berkeley, CA, USA, 2009, pp. 467–480, USENIX Association.
- [14] Yuval Shavitt and Tomer Tankel, "Big-bang simulation for embedding network distances in euclidean space," *IEEE/ACM Trans. Netw.*, vol. 12, pp. 993–1006, December 2004.
- [15] Ryoichi Kawahara, Eng Keong Lua, Masato Uchida, Satoshi Kamei, and Hideaki Yoshino, "On the quality of triangle inequality violation aware routing overlay architecture," in *INFOCOM*, 2009, pp. 2761–2765.
- [16] P2Psim, "A simulator for peer-to-peer protocols," <http://www.pdos.lcs.mit.edu/p2psim/index.html>.
- [17] Bernard Wong, Aleksandrs Slivkins, and Emin Sirer, "Meridian : A lightweight network location service without virtual coordinates," *IN SIGCOMM*, pp. 85–96, 2005.
- [18] PlanetLab, "An open platform for developing, deploying, and accessing planetary-scale services," 2002, <http://www.planet-lab.org>.
- [19] Krishna P. Gummadi, Stefan Saroiu, and Steven D. Gribble, "King : Estimating latency between arbitrary internet end hosts," in *SIGCOMM Internet Measurement Workshop 2002*, 2002.
- [20] Laurie J. Heyer, Semyon Kruglyak, and Shibu Yooseph, "Exploring Expression Data : Identification and Analysis of Coexpressed Genes," *Genome Research*, vol. 9, no. 11, pp. 1106–1115, Nov. 1999.

SESSION 6

ARCHITECTURES TO SUPPORT A FULLY NETWORKED SOCIETY

- S6.1 Invited Paper: Effective Collaborative Monitoring In Smart Cities: Converging Manet And Wsn For Fast Data Collection
- S6.2 SOA Driven Architectures for Service Creation through Enablers in an IMS Testbed
- S6.3 A Virtualized Infrastructure for IVR Applications as Services
- S6.4 Seamless Cloud Abstraction, Model and Interfaces

EFFECTIVE COLLABORATIVE MONITORING IN SMART CITIES: CONVERGING MANET AND WSN FOR FAST DATA COLLECTION

Giuseppe Cardone, Paolo Bellavista, Antonio Corradi, and Luca Foschini

Dipartimento di Elettronica, Informatica e Sistemistica (DEIS) - University of Bologna (Italy)

ABSTRACT

Ubiquitous smart environments, equipped with low-cost and easy-deployable Wireless Sensor Networks (WSNs) and with widespread Mobile Ad-hoc NETWORKS (MANETs), are opening brand new opportunities in urban monitoring. Urban data collection, i.e., the harvesting of monitoring data sensed by a large number of collaborating sensors in a wide-scale city, is still a challenging task due to typical WSN limitations (limited bandwidth and energy, long delivery time, ...). In particular, effective data collection is crucial for classes of services that require a timely delivery of urgent data, such as environmental monitoring, homeland security, and city surveillance. This paper proposes an original solution to integrate and to opportunistically exploit MANET overlays that are impromptu and collaboratively formed over WSNs in order to boost data collection: overlays are used to dynamically differentiate and fasten the delivery of urgent sensed data over low-latency MANET paths. The reported experimental results show the feasibility and effectiveness (e.g., limited coordination overhead) of our solution for MANET overlays over WSNs. In addition, our proposal can easily integrate with the latest emergent WSN data collection standards/specifications, thus allowing immediate deployability over existing smart city environments.

Keywords— MANET, WSN, collaborative monitoring, smart cities, opportunistic networks.

1. INTRODUCTION

Smart cities are usually envisioned as intelligent environments able to facilitate everyday citizens' life by increasing their wellbeing in many advanced and comforting ways. Toward the realization of such a smart city vision, there is the need to enable dynamic and continuous collection, elaboration, and presentation of data, possibly deriving from collaborative participatory sensing [1]. Moreover, those functions should be supported by small (disappearing) wireless sensors and by computing devices that are typically available because either embedded in the environment or offered by collaborating users. A relevant goal is to devise new, autonomic, and adaptable services for smart cities, which may span several different application domains, from environmental and habitability monitoring (noise/light pollution, vehicle traffic, etc.), to security monitoring (anti-

theft protection, structural monitoring to prevent collapses of old buildings and bridges, etc.), and to assist citizens' urban living and roaming (elderly assistance services, emergency response, etc.).

Recent advances in wireless communications and mobile devices are enabling these new service opportunities through novel integration possibilities. Wireless Sensor Networks (WSNs), namely networks composed of tiny and inexpensive autonomous devices equipped with sensors, can take measurements, locally store and handle sensed data, and communicate to each other [2]. At the same time, last-decade progresses in ad-hoc wireless technologies (for instance, IEEE 802.11 and Bluetooth) have made viable and widely diffused Mobile Ad-hoc NETWORKS (MANETs), where it is possible to build impromptu connections without predefined fixed infrastructures.

The possibility of integrating WSNs and MANETs enables brand new cross-network routing opportunities to overcome the typical limitations of WSN data collection, thus enabling novel and cost-effective applications for a large audience of smart city users. In fact, many WSN applications can benefit from low-latency delivery of some specific subsets of sensed data, such as for urgent data measurements associated with security alarms and critical conditions. Most solutions for WSN data collection are unable to grant timely delivery of urgent data with low-latency requirements because, especially in city-wide WSN deployments, sensed data have to traverse a high number of potentially congested WSN nodes before reaching their ultimate gateway toward the fixed infrastructure [3, 4]. In addition, existing solutions typically tend to mix together in-band urgent with normal non-urgent data. Some research activities have started to explore the possibility of using special nodes, immersed in WSNs and equipped with both low-power and powerful ad-hoc wireless interfaces, that act as relays to accelerate data collection. However, these solutions have not been widely deployed yet, primarily because they typically require special-purpose dedicated hardware, thus preventing from the exploitation of the already existing large base of installed sensors/devices [5, 6].

Very recently, mobile phones, already equipped with multiple wireless interfaces (IEEE 802.11, Bluetooth, and cellular 3G), have started hosting onboard also low-power connectivity solutions, such as IEEE 802.15.4; moreover, they are expected to become more and more available also on consumer devices in the near future [7, 8]. Stimulated by these recent technological advances, we propose to opportunistically exploit these new mobile devices to speed up

WSN urgent data collection while they roam in a smart city WSN. Let us stress that, differently from other approaches in the literature, our MANET relays are neither mobile harvesters that relay sensed data to the Internet, nor mobile users of WSN only, as better detailed in the following.

The paper presents the core design guidelines of our novel architectural approach toward fast, cross-network, and opportunistic WSN data collection by focusing on several original contributions. First, we do not assume any previous knowledge about MANET node mobility. We adopt a cross-network and cross-layer design that limits the communication over the WSN by activating interactions between MANET and WSN only when necessary, mainly for urgent data delivery. Second, we follow a locality principle to carefully control and limit the MANET coordination overhead: we form localized clusters of a limited number of MANET nodes in an impromptu way only when and where needed, i.e., only for the delivery of the most valuable urgent data. Third, we finely and locally tune the proposed protocol for MANET cluster formation in order to find the best balance between the benefit of enabling MANET-layer routing and the cost of coordinating MANETs and WSNs. Last but very relevant, the proposed solution is fully compliant with the Collection Tree Protocol (CTP), which recently informed the standardization process of the soon-to-be-standardized IETF IPv6 Routing Protocol for Low power and Lossy Networks (RPL) [4, 9].

We have designed and implemented a prototype of our proposal, described in-depth in the following; the prototype has been tested through extensive simulations, publicly available for WSN practitioners¹. The reported experimental results show that our solution can effectively differentiate and speed-up urgent data delivery in the wide-scale smart city environments envisioned for the near future, with controlled, predictable, and very limited MANET coordination overhead and with easy deployability thanks to compliance with emerging standards.

2. MOTIVATIONS AND REFERENCE MODEL

Our main motivation is to support, in an easily deployable and cost-effective way, the realization of wide-scale urban monitoring applications, e.g., targeted to structural monitoring, in smart cities. The case study of structural monitoring applications is characterized by properties, requirements, and features that are common to any application environment where alarm situations with differentiated urgency levels may occur; our proposal can easily apply to any of these environments.

To practically exemplify our proposal in a concrete application domain, let us consider a monitoring application that targets the structural integrity of buildings, with WSN nodes deployed for data collection purposes over different city areas, namely busy roads, old buildings, and bridges. In smart cities, mobile devices are expected to roam through densely populated areas, by enabling the possibility to consider the impromptu formation of MANETs. When a criti-

cal event is detected, the considered structural monitoring application should trigger an alert that has to be delivered as fast as possible to WSN data collection points. For instance, if a critical flexure of a column is detected, the application should dispatch the alert faster than other normal flexure readings. It is important to note that packet latency over WSNs typically depends on the number of routing hops and on duty cycling of sensor nodes, which may periodically turn off their radio transceivers to save energy at the cost of higher latency [10]. Packet latency is usually considered a system parameter that has to be tuned properly, also because it has non-negligible effects on battery lifetime. In common deployment scenarios where about one year of battery lifetime is required, radio duty cycling is forced to be around one second per hop, with even a longer delay when energy constraints are stricter [11]. Our primary idea is to reduce the delivery time of only most relevant urgent data without sacrificing battery lifetime, by dynamically pushing urgent alerts from WSN atop a MANET-based overlay whenever possible. Of course, depending on the targeted urban deployment, the involved WSNs are exposed to MANET nodes with different mobility patterns and density; our solution aims at addressing any execution environment, without imposing any strong assumption on the static knowledge of MANET characterization parameters.

Given the targeted application domain sketched above, our proposal tries to effectively satisfy some primary requirements. First of all, there is the need for the WSN to relay urgent data and to flow them over the more rapid MANET overlay trunks as soon as the WSN can locally have connectivity with a mobile MANET node. That alleviates two main WSN communication issues: scarcity of energy and low communication bit rate. In fact, even if MANET devices often have limited computing power, their constraints are of orders of magnitude weaker than the WSN ones. Let us also briefly note that incentive-based frameworks can be employed to avoid non-cooperative, selfish, and malicious behaviors of MANET nodes, which may endanger the effective realization of MANET-WSN integration. In addition, data tampering can be prevented by adopting authentication systems tailored for low-power devices [12, 13]. In addition, since communication between MANETs and WSNs could drain precious energy resources from WSN nodes, it is crucial to design solutions and protocols that minimize MANET-WSN interactions, by enabling them only for urgent data delivery. While we have already addressed some of the above problems from the WSN perspective in our previous work (please refer to [14] for further information), this article totally focuses on the MANET side. Hence, in the rest of the paper we will detail all main coordination and clustering functions realized at MANET nodes for WSN-MANET integration in smart city deployment environments.

Let us complete this section by sketching our abstract reference model and by providing some needed background material. Our distributed architecture includes two main network layers: at the lower level, WSN (fixed) sensor nodes form an autonomous routing layer that delivers normal/urgent data to one or more roots; at the higher level, multi-homed mobile MANET nodes roam across the WSN-

¹ Additional information, experimental results, and simulation code are available at: <http://lia.deis.unibo.it/Research/WHOO/>

equipped environment. Let us assume a tree-based data collection for WSNs: that leads to organize WSNs in tree-like topologies and to exploit a very general tree formation method based on a gradient function. Tree-like topologies for WSN data routing have been widely employed in both experimental protocols, such as Hyper and CTP [3, 4], and in more recent standardization efforts, such as ZigBee and IETF RPL [9, 15]. Thus, we decided to use a generic tree-based protocol as the reference for our work in order to easily enable its deployment and immediate usage with all emerging collection solutions and standard specifications. Tree formation and data routing work as follows. Generally, data roots start advertising a zero cost, while each internal node advertises a total incremental cost, equal to the cost of its father node plus the cost of the link to the next hop; data packets flow along paths toward lower cost nodes. The WSN level opportunistically exploits its MANET nodes in visibility to create the additional low-latency high-bandwidth overlay for urgent data routing. To glue together WSNs and MANETs, MANET nodes exploit their WSN interfaces to participate to urgent data routing by dynamically discovering WSN nodes during their roaming and by advertising their presence to them.

To overcome mobility and scalability issues typical of large and dense MANET deployments, we claim the need for novel solutions and standards to organize MANET nodes in small local clusters, as we will better detail in the following. For the sake of easy readability and presentation clarity, let us define here some useful terms. *Roots* are sensor nodes that advertise themselves as collection tree roots, typically acting as gateways to the Internet. All other sensor nodes build routing trees to forward collected data toward roots at the WSN layer. A *WSN exit point* is any WSN node in visibility of at least one MANET node and able to jump urgent data over the MANET, while a *WSN entry point* is the WSN node with the lowest gradient cost that the MANET cluster can reach. Finally, *MANET entry/exit points* are MANET nodes that can respectively receive/forward data from/to the WSN.

Our solution is general enough to work with most tree-based sensor data collection standards and related research-oriented protocols, such as IETF RPL and CTP. IETF RPL is a very promising standard specification in the field, but at the current stage there are still a very few examples of its deployment and it suffers from limited testing in realistic in-the-field scenarios. Therefore, in our current prototype of the proposal, we have decided to be fully compliant with CTP because it is robust, thoroughly assessed, and has a strong community of developers working on it. Additional information about CTP is out of the scope of this paper and the interested reader can refer to [4].

3. DESIGN GUIDELINES AND PROTOCOL OVERVIEW

This section overviews the design guidelines of our proposal by focusing on MANET-related aspects. Then, we present the main facilities that we have designed and im-

plemented to facilitate the exploitation of MANET-WSN convergence. Design Guidelines.

Given the goal of seamlessly bridging extremely heterogeneous networks such as WSNs and MANETs, our solution addresses a very challenging cross-network environment and has to follow some original design guidelines. First, our MANET communication protocol should be *opportunistic*. MANET nodes should dynamically self-organize and opportunistically exploit their current neighbors to provide fast data collection even in difficult network conditions. Second, the MANET organization protocol should be *localized*. It is well known that MANETs suffer from severe robustness and bandwidth problems as the end-to-end path length increases, thus MANET nodes should organize themselves in local clusters, without requiring long-distance interactions [16, 17]. Third, it should be *tolerant to node mobility*: MANET nodes are intrinsically mobile, thus it is important to tolerate node mobility while avoiding possible disruptions in the low-latency routing overlay offered to WSNs. Fourth, our solution should be *able to save energy*: it is crucial that our integration protocol avoids wasting resources, especially WSN ones, to save battery lifetime.

3.2 MANET-WSN Convergence Facilities

Our original proposal for MANET overlay follows our design guidelines and includes two core functions: i) MANET-WSN integration to enable MANET-WSN impromptu communications only if that is feasible and beneficial (opportunistic and energy-saving guidelines), and ii) MANET cluster formation to organize MANET nodes in small clusters to avoid the routing issues of large MANETs (localization and tolerance to node mobility guidelines)

About MANET-WSN integration, two facilities are needed to enable a MANET to play the role of WSN backbone: *discovery*, to let MANET nodes explore the WSN topology and select the WSN node with the best gradient, i.e., WSN entry point, and *advertising*, to inform the WSN of the presence of MANET entry points. Let us stress that, given that in many modern low-power radio transceivers sending and receiving packets require comparable amounts of energy, there is the need to minimize both discovery and advertising packets [18]. If our MANET-WSN integration support were always active, regardless of WSN traffic, that would impose an additional traffic load on the WSN, by worsening node power consumption. For this reason, the energy-saving guideline suggests avoiding packet exchanges between MANET and WSN nodes at default, by keeping MANET nodes usually idle (*dormant state*). MANET nodes should only and passively snoop CTP traffic to obtain information about the underlying WSN tree topology. Only upon sniffing a urgent packet, MANET nodes start coordinating and communicating with WSN ones to self-organize as relays for urgent WSN packets (*running state*) [14].

About MANET cluster formation, we recall that MANET relays should avoid all fragilities related to network architectures with centralized coordination and multiple-hops communication. Thus, upon snooping a urgent WSN packet, MANET nodes should organize themselves in local independent clusters, each one with its own MANET entry

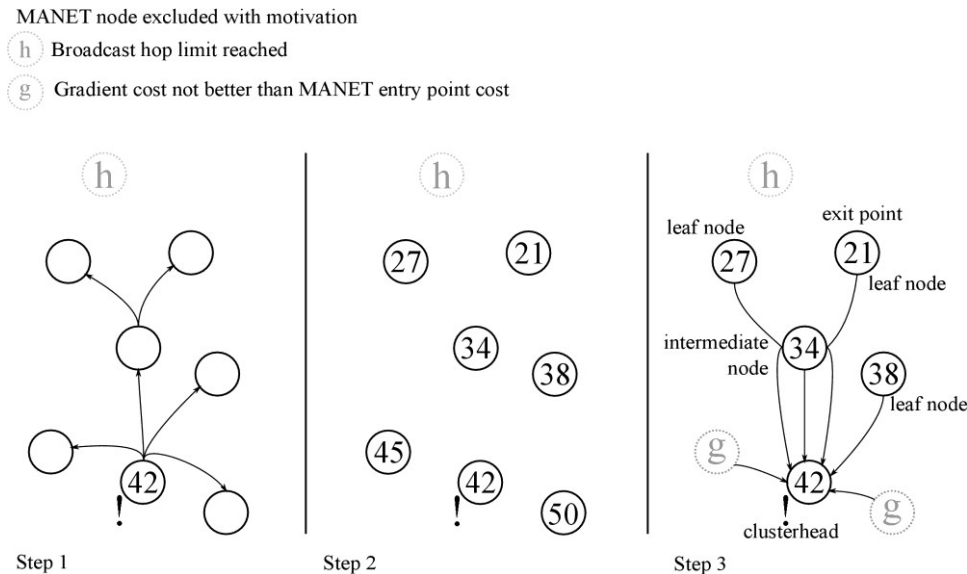


Figure 1. Phases of MANET cluster formation. Exclamation mark identifies the clusterhead. Each MANET node is marked with the best gradient (the lowest) it can reach at the WSN layer. Phase 1: a MANET node snoops a urgent data packet and broadcasts a hop-limited broadcast request. Phase 2: MANET nodes hit by the request send discovery request to WSN, obtain gradient costs of reachable sensor nodes, and choose the best one. Phase 3: all sensor nodes reply to the MANET node that started the process. MANET nodes marked with “g” will not enter the cluster because their gradient is worse than what broadcasted in phase 1.

and exit points. Note that, due to diversity in wireless coverage ranges between IEEE 802.15.4 and IEEE 802.11, even small clusters can significantly improve data collection performance, by making it possible to jump several WSN hops by traversing a few MANET ones, as exemplified in [19, 20]. In addition, small clusters are intrinsically more tolerant with regard to node mobility if compared with fully-connected mobile networks because they have to keep a limited number of routing paths (tolerance to node mobility guideline).

4. MANET PROTOCOLS

This section presents our cluster formation protocol and describes it as a Finite State Machine (FSM). Then, it details the most important packet exchanges performed by our solution at the MANET layer.

4.1 MANET Cluster Formation

Our MANET-WSN integration exploits MANET clusters opportunistically formed in localized areas that need urgent data transmission. It is simple and robust, and relies only on one-hop communications and limited-hop broadcasts. Although our protocol does not intrinsically pose a limit to the hop radius of clusters, we found that a 2-hop limit is a good trade-off between routing improvements and cluster robustness; hence, in the following, we assume the broadcasts to be 2-hop limited.

The cluster formation protocol is reactively started by any MANET node, defined as clusterhead, that snoops a urgent data packet being routed on the WSN. It consists of three phases (Fig. 1). In the first phase, the clusterhead extracts from the sniffed packet the gradient of the WSN node that has routed it (42 in the example in Fig. 1); it broadcasts a 2-

hop limited request to other nodes, by asking them to join the new cluster, i.e., join request. In the second phase, MANET nodes that received the join request send a discovery message to the WSN layer to get the best gradient among the WSN nodes they can communicate with; after that, they compare it with the gradient declared in the join request. Only the MANET nodes that can communicate with WSN nodes having a better gradient will take part to the cluster. In the third and last phase, MANET nodes reply to the clusterhead stating that they will join the new cluster. The clusterhead gathers responses from cluster nodes and chooses as MANET exit point the node that declares the best gradient value.

Fig. 1 clearly shows that the proposed protocol forms clusters with a tree-like structure, having the notable property that the gradient gets better by following any path from the clusterhead to MANET leaves. This property strongly enhances the robustness of urgent data routing within a cluster: in fact, when the clusterhead tries to route a urgent packet to the designated exit node, it is guaranteed that at each hop the urgent packet will be forwarded to a MANET node that has a better gradient than the previous hop. Thus, even if the path to the final MANET exit point is disrupted (e.g., due to node mobility), intermediate nodes can still route the urgent data packet to the WSN, by achieving anyway a (suboptimal) performance boost.

4.2 MANET Coordination Protocol as FSM

To better understand the different roles that MANET nodes can play in a cluster, we model it as an FSM.

Fig. 2 shows the simplified FSM implementing our MANET coordination protocol. MANET nodes start working in the IDLE state: in this state, they only snoop WSN traffic, while waiting for a urgent packet to be routed by the

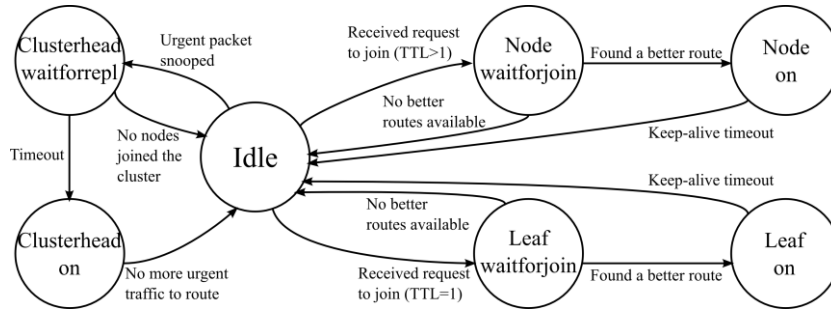


Figure 2. State diagram of the MANET coordination protocol. For the sake of presentation clarity, it does not show the substates necessary to manage asynchronous communication with WSN nodes.

underlying WSN. When a MANET node snoops a urgent packet, it switches to CLUSTERHEAD_WAITFORREPL state: it broadcasts a hop-limited join request to other MANET nodes to ask them to join the new cluster where the sender acts as clusterhead. A MANET node waits in CLUSTERHEAD_WAITFORREPL for a fixed amount of time; after that, the potential clusterhead considers the replies received by nearby MANET nodes: if at least one MANET node joined the cluster, the potential clusterhead really becomes clusterhead and switches to the CLUSTERHEAD_ON state. If no nodes joined, instead, it switches back to the IDLE state.

Upon receiving a join request, IDLE nodes switch to NODE_WAITFORJOIN or LEAF_WAITFORJOIN, respectively if either they have to broadcast it again or not, based on the TTL (time to live as number of hops) in the join request. Then, they broadcast a discovery request to the WSN to obtain the gradient value of the WSN nodes they can communicate with. They decide whether to join the cluster by comparing gradient values of nearby WSN nodes with the gradient value broadcasted by the potential clusterhead. Finally, they send their decision back to their parent.

Nodes and leaves, accepting to be members of the cluster, switch to NODE_ON (internal tree nodes) and LEAF_ON (leaf nodes), instead other nodes revert back to IDLE. A clusterhead in the CLUSTERHEAD_ON state periodically broadcasts hop-limited keep-alive packets to prolong the cluster lifetime by keeping the cluster in a working state. Intermediate nodes forward those packets to other intermediate nodes and leaves. The clusterhead evaluates cluster lifetime by taking into account both the number of MANET nodes and the frequency of discovery/advertising functions [14].

As explained in the previous sections, clusters are formed independently of one another. Thus, a single MANET node can participate with different roles (namely, node or leaf) to different clusters; anyway, it can be clusterhead in one cluster only. From the implementation point of view, that means that each MANET node runs at the same time different FSMs, each one uniquely identified by the address of its clusterhead.

5. EXPERIMENTAL RESULTS

To validate our proposal on a wide scale, we have originally ported CTP to the QualNet network simulator and have implemented our protocol on top of it [4, 21]. In the adopted simulation environment, we made the sensor nodes use the IEEE 802.15.4 physical layer and a Carrier Sense Multiple Access (CSMA) MAC protocol, thus simulating a realistic communication testbed, similar to what used by many widely adopted real-world sensor nodes, such as TelosB and MICAz² [22]. In addition, we modified the used MAC layer to simulate the delays due to radio duty cycling, as often done in WSNs to improve battery lifetime. In particular, we simulated the delays experienced by a TelosB sensor node running on a 2.5% duty cycle that keeps the radio interface active for the 2.5% of its running time. This duty cycling, much less aggressive than what usually employed (1% or less [23]), has been chosen as a worst-case scenario not to favor too much our MANET-enabled urgent data delivery. Finally, MANET nodes use the IEEE 802.11b physical and MAC layers [24].

To evaluate the performance of our system in a smart city environment, in QualNet we have modeled a 1km-long and 10m-wide street, monitored by 50 sensor nodes, 20m apart from each other. The sensor node at the beginning of the street acts as the tree root, while the one at the end of the street alternately generates one normal data packet and one urgent data packet, with a period of 3s.

Our first evaluations focus on packet latencies. In particular, we observed the impact of MANET node density on packet delivery latency. We simulated the reference scenario by constantly increasing the number of randomly placed MANET nodes. We repeated each test for 30 runs; the collected error margins were always under 5%. Fig. 3 reports the experimental results. As expected, our solution vastly reduces delivery latency of urgent packets in every tested case. The reported results show that there is a minimum for MANET node density that gives the best latency improvements (in our scenario, for 30 MANET nodes). For lower MANET densities, latencies are not so good because MANET clusters do not cover all the WSN areas, thus forc-

² By default, TinyOS uses a CSMA MAC instead of the IEEE 802.15.4 MAC on TelosB; MICAz and other sensor nodes that use the same radio transceiver.

ing data packets to hop over slower WSN hops. When there are more than 30 nodes, latency goes up again due to the increased traffic induced by advertising and discovery packets, which cause more packet collisions. Finally, it is worth noting that, since CTP does not natively include any form of traffic differentiation, the reported results for nor-

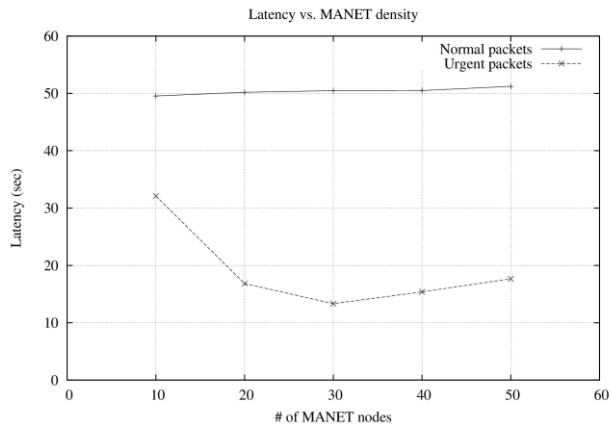


Figure 3. Packet latencies vs. MANET nodes density.

mal packets also exemplify the performance of urgent packets in a WSN-only scenario where MANET nodes are not available to help in speeding up the routing function.

Our second set of experimental results assesses the impact of MANET node mobility on packet latency. We used the same scenario of the previous evaluation and kept the number of MANET nodes fixed at 40. We made MANET nodes move randomly over the simulated road adopting the random waypoint mobility model at various speeds (from 1 m/s to 10 m/s). Figure 4 reports the related evaluation results. Numerical simulation shows that, as MANET speed increases from 0 to 1.0m/s, the average latency goes from 15s to 20s. At higher speeds, latency grows very slowly, reaching 23s at 10m/s. This result shows that there is a non-negligible difference in packet latency only when comparing completely still MANET nodes to mobile MANET nodes, while speed per se has a limited impact on packet latency.

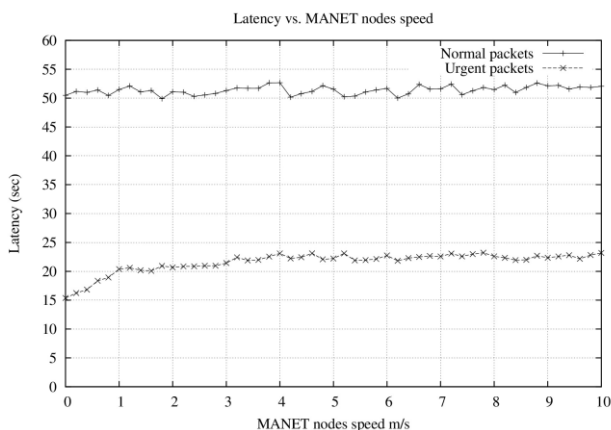


Figure 4. Packet latency vs. MANET nodes speed.

Another important indicator is the packet delivery ratio, i.e., the number of packets successfully dispatched from data source to root. Fig. 5 reports packet delivery ratio under the same conditions of the previous evaluation, i.e., with MANET nodes moving at growing speeds. As expected, the CTP protocol is very reliable and always delivers more than

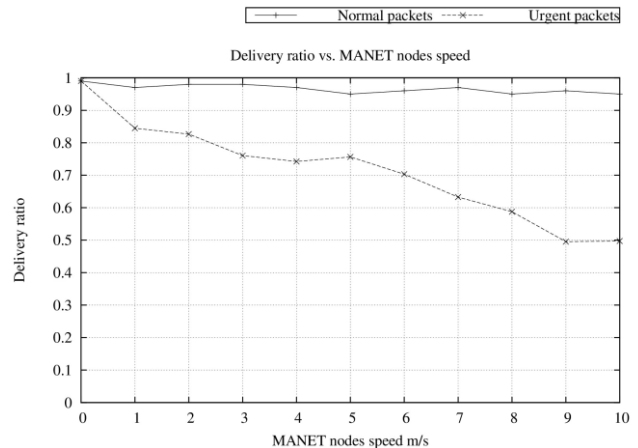


Figure 5. Ratio of packets successfully routed vs. MANET nodes speed.

95% of packets. Our solution achieves a packet delivery ratio comparable to CTP when MANET nodes are not moving, while the ratio sensibly decreases as MANET nodes become more and more mobile. In particular, it delivers more than 80% of urgent packets when MANET nodes move at 1m/s, and drops to about 50% when MANET nodes move at 10m/s [25]. In addition, let us note that the delivery ratio of our solution can be boosted by repeated sending of urgent packets. For example, when MANET nodes move at 1m/s, a urgent packet has 82% probability to be successfully routed to the collection tree root. If the node sends the packet twice, the probability of successful routing raises to 96% (it follows a Bernoulli distribution); another repeated send operation achieves the success probability of 99%. Thus, we claim that our solution can achieve a good trade-off between routing speed and delivery ratio, by allowing to significantly improve data collection in most common use cases for typical smart city environments.

6. RELATED WORK

It is widely recognized that WSNs are a key enabling technology for smart cities. Yovanof and Hazapis describe the crucial role of WSNs in the software and hardware architecture of smart cities [26]. Far to be of pure academic interest, [27] reports that many municipalities in the next decade will adopt WSNs, mainly for public safety, localization, and environmental monitoring. An example of the potential impact of this technology is the Republic of Korea, which in recent years has invested in WSNs and related technologies to provide a wide range of services/applications to citizens, spanning from environmental monitoring to traffic management and entertainment [28].

In a more technical perspective, the use of multi-radio devices in WSNs and the integration of WSNs and mobile nodes have recently started to appear in the literature. Yarris et al demonstrated that a modest number of reliable long-range links can improve WSN delivery ratio and battery lifetime [29]. A practical proof of this effect is the ExScal project, which deployed more than a thousand WSN nodes, by exploiting about two hundreds high-powered dual-radio nodes as an always-on high-speed network backbone [5]. Siphon is similar to ExScal but exploits multi-radio sensor nodes to provide an on-demand traffic management service that relieves congested traffic [6]. In general, the work in the literature assumes that a relevant subset of sensor nodes provides both a low-power radio and an IEEE 802.11 interface; however, this assumption is not realistic for deployment environments that should be cheap and cannot guarantee enough power for IEEE 802.11 interfaces for all the expected lifetime.

The exploitation of mobile nodes as data harvesters and as WSN gateways towards the Internet is a hot research area. [30] employs mobile nodes with a predictable roaming path as data sinks/mules, thus trading lower power consumption for higher latencies. Wang et al showed that one mobile relay that stores/forwards gathered data to a data sink can improve WSN lifetime up to a factor of four [31]. [32] proposes the tiered mWSN architecture, which makes WSN nodes form a cluster around the expected position of mobile nodes that act as statically pre-defined and mobile Internet gateways. A survey of hierarchical WSN architectures enhanced by mobile nodes can be found in [33]. All these papers highlight aspects related to energy saving, but not mobility. To the best of our knowledge, our solution is the first one in this field that specifically focuses on mobility by proposing to opportunistically exploit mobile MANET nodes as mobile relays for fast collection of WSN urgent data in smart city environments.

7. STANDARDIZATION REQUIREMENTS AND CONCLUSIVE REMARKS

This paper proposes a novel solution for opportunistic cross-network collaboration for fast collection of urgent sensed data in smart cities. As also highlighted by other relevant papers in the recent literature, our proposal shows the potential of WSNs for smart cities: ubiquitous and collaborative urban sensing can vastly improve everyday citizens' life by providing an intelligent environment that offers services, prevents emergencies and reacts to them, and enables a fine-grained adaptive control for better and more scalable management of urban environments.

In order to make this vision real and usable (cost-effective and easily deployable), we claim the need of middleware-layer solutions fully compliant with accepted standards (or largely adopted specifications). In this perspective, we claim the suitability of working on additional standardization efforts on two fronts: hardware and software. About hardware standards, our reference scenario shows that it is important to provide smart phones with low power wireless interfaces for WSN connectivity. Currently, there are several competing proposals, the most prominent ones being

ZigBee, Bluetooth Low Energy, DASH7, Wavenis, and Z-Wave [15, 34, 35, 36, 37]. Those specifications, either open or proprietary, offer different and competing features: the recent research work in the area, including our prototyping/development activities, is showing that the convergence towards a dominant single standard could relevantly help in reducing production costs and ensuring maximum interoperability, thus leveraging the mass deployment of smart city solutions on consumer devices. About software, we claim the need for a standard routing solution to enable MANET-WSN integration in an open and highly interoperable way. Some existing standardization efforts have proposed guidelines for such routing protocols: the most relevant efforts in this field are ZigBee and RPL [9, 15]. However, even though routing protocols can be changed via firmware upgrade, our opinion is that it would be far more desirable to reach an agreement about a single and flexible routing solution (dynamically configurable for different use cases) or, at least, a single network access abstraction to enable the communication between mobile devices and WSNs. It is hard to decide which standards organization would be more suitable to work on the above standardization issues. Given the addressed areas, hopefully, a joint effort by IEEE and IETF would be able to reach the widest audience, by accelerating the wide-scale adoption of interoperable low-power wireless interfaces.

In conclusion, we claim that industry and academic standardization efforts in the near future could significantly contribute to the speed-up of the research and the real-world deployment of smart city solutions coming from the dynamic collaboration of WSNs and MANETs, thus enabling them to finally express their full potential. Hence, we believe our proposal represents an important step forward toward the development of a novel standard aimed to truly enable MANET and WSN convergence in order to support fast collection of urban sensing data.

This paper also provides a full assessment and quantitative evaluation of the feasibility and performance of our proposal, by reporting an extensive set of simulation results that are indicative of its benefits and costs. All results show the viability of our proposal in terms of induced overhead and confirm the suitability of employing mobile MANET nodes equipped with low power interfaces to increase data collection performance.

The encouraging results already obtained are stimulating us toward the extension and refinement of our prototype. On the WSN side, we are considering the addition of more expressive management operations, such as dynamic, fine, and application-aware differentiated tuning of the used levels of transmission power. On the MANET side, we are working on lightweight solutions for predicting node mobility in order to further improve data delivery ratios, especially useful for scenarios with high node speeds.

REFERENCES

- [1] P. Bellavista, A. Corradi, M. Fanelli, L. Foschini, "A Survey of Context Data Distribution for Mobile Ubiquitous Systems," to appear in *ACM Computing Surveys (CSUR)*, pp. 1-

49. Available online at http://www-lia.deis.unibo.it/Staff/LucaFoschini/pdfDocs/context_survey_CSUR.pdf.
- [2] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, E. Cayirci, "Wireless sensor networks: a survey," *Computer Networks*, Vol. 38, No. 4, pp. 393-422, 2002.
 - [3] D.E.T. Schoellhammer, B. Greenstein, D. Estrin, "Hyper: a Routing Protocol to Support Mobile Users of Sensor Networks," Center for Embedded Network Sensing (CENS) Technical Report, 2006.
 - [4] O. Gnawali, et al., "Collection Tree Protocol," in *SenSys'09: Proc. 7th ACM Conf. Embedded Networked Sensor Systems*, Berkeley, USA, pp. 1-14, 2009.
 - [5] A. Arora, et al., "ExScal: Elements of an Extreme Scale Wireless Sensor Network," in *RTCSA'05: Proc. 11th IEEE Int. Conf. Embedded and Real-Time Computing Systems and Applications*, Hong Kong, pp. 102-108, 2005.
 - [6] C.-Y. Wan, S.B. Eisenman, A.T. Campbell, J. Crowcroft, "Siphon: Overload Traffic Management Using Multi-Radio Virtual Sinks in Sensor Networks," in *SenSys'05: Proc. 3rd Int. Conf. Embedded Networked Sensor Systems*, San Diego, USA, pp. 116-129, 2005.
 - [7] P. Trevor, et al., "The PSI Board: Realizing a Phone-centric Body Sensor Network," in *BSN'07: Proc. 4th Int. Workshop on Wearable and Implantable Body Sensor Networks*, Aachen, Germany, 2007.
 - [8] ABI Research, "Short Range Wireless ICs: Bluetooth, NFC, UWB, 802.15.4 and Wi-Fi Market Forecasts," 2010.
 - [9] T. Winter, et al., "RPL: IPv6 Routing Protocol for Low power and Lossy Networks. Available online at <http://tools.ietf.org/html/draft-ietf-roll-rpl-19>, 2011.
 - [10] L. Sha, F. Kai-Wei, P. Sinha, "CMAC: an Energy Efficient MAC Layer Protocol Using Convergent Packet Forwarding for Wireless Sensor Networks," in *SECON'07: Proc. 4th IEEE Communications Society Conf. Sensor, Mesh and Ad Hoc Communications and Networks*, San Diego, USA, pp. 11-20, 2007.
 - [11] P. Bellavista, G. Cardone, A. Corradi, L. Foschini, "The Future Internet Convergence of IMS and Ubiquitous Smart Environments: an IMS-based Solution for Energy Efficiency," *Journal of Network and Computer Applications (JNCA)*. In Press.
 - [12] P. Hui Chia, "Analyzing the Incentives in Community-based Security Systems," in *PERCOMW'11: Proc. IEEE Pervasive Computing and Communications Workshops*, Seattle, USA, pp. 270-275, 2011.
 - [13] Y. Lili, P. Daiyuan, G. Yuexiang, "The Study of Mutual Authentication and Key Exchange Protocols for Low Power Wireless Communications," in *WICOM'10: Proc. 6th Int. Conf. Wireless Communications Networking and Mobile Computing*, Wuhan, China, 2010.
 - [14] G. Cardone, A. Corradi, L. Foschini, "Cross-Network Opportunistic Collection of Urgent Data in Wireless Sensor Networks," *The Computer Journal*, May 2011.
 - [15] ZigBee Alliance, "ZigBee Specification," 2005.
 - [16] J. Li, et al., "Capacity of Ad hoc Wireless Networks," in *MobiCom'01: Proc. 7th Int. Conf. Mobile Computing and Networking*, Rome, Italy, pp. 61-69, 2001.
 - [17] R. Ramanathan, J. Redi, "A Brief Overview of Ad Hoc Networks: Challenges and Directions," *IEEE Communications Magazine*, Vol. 40, No. 5, pp. 20-22, 2002.
 - [18] W. Qin, M. Hempstead, W. Yang, "A Realistic Power Consumption Model for Wireless Sensor Network Devices," in *SECON'06: Proc. 3rd Annual IEEE Communications Society Conf. Sensor and Ad Hoc Communications and Networks*, pp. 286-295, 2006.
 - [19] E. Callaway, et al., "Home Networking with IEEE 802.15.4: a Developing Standard for Low-Rate Wireless Personal Area Networks," *IEEE Communications Magazine*, Vol. 40, No. 8, pp. 70-77, 2002.
 - [20] J. Zheng, M. J. Lee, "A Comprehensive Performance Study of IEEE 802.15.4," in *Sensor Network Operations: Wiley-IEEE Press*, pp. 218-237, 2006.
 - [21] Scalable Network Technologies. QualNet Simulator. Available online at <http://www.scalable-networks.com/products/qualnet/>.
 - [22] IEEE 802.15 Working Group, "IEEE standard 802.15.4-2006," 2006.
 - [23] P. Dutta, D. Culler, S. Shenker, "Procrastination Might Lead to a Longer and More Useful Life," in *HotNets VI: Proc. 6th Workshop on Hot Topics in Networks*, Atlanta, USA, 2007.
 - [24] IEEE 802.11 Working Group, "IEEE Standard 802.11b-1999," 1999.
 - [25] R.L. Knoblauch, M. Pietrucha, M. Nitzburg, "Field Studies of Pedestrian Walking Speed and Start-up Time," Transportation Research Board Records No. 1538, 1996.
 - [26] G. Yovanof, G. Hazapis, "An Architectural Framework and Enabling Wireless Technologies for Digital Cities & Intelligent Urban Environments," *Wireless Personal Communications*, Vol. 49, No. 3, pp. 445-463, 2009.
 - [27] ON World Research, "WSN for Smart Cities: a Market Study," 2007.
 - [28] S. Lee, H. Han, Y. Leem, T. Yigitcanlar, "Towards Ubiquitous City: Concepts, Planning and Experiences in the Republic of Korea," in *Knowledge-Based Urban Development: Planning and Applications in the Information Era*, T. Yigitcanlar, K. Velibeyoglu, S. Baum (Eds.) London: Information Science Reference, pp. 148-170, 2008.
 - [29] M. Yarvis, et al., "Exploiting Heterogeneity in Sensor Networks," in *INFOCOM'05: Proc. 24th Annual Joint Conf. of the IEEE Computer and Communications Societies*, Miami, USA, pp. 878-890, 2005.
 - [30] A. Chakrabarti, A. Sabharwal, B. Aazhang, "Using Predictable Observer Mobility for Power Efficient Design of Sensor Networks," in *Information Processing in Sensor Networks*, Vol. 2634, F. Zhao and L. Guibas (Eds.): Springer Berlin Heidelberg, pp. 552-552, 2003.
 - [31] W. Wang, V. Srinivasan, K.-C. Chua, "Using Mobile Relays to Prolong the Lifetime of Wireless Sensor Networks," in *MobiCom'05: Proc. 11th Int. Conf. Mobile Computing and Networking*, Cologne, Germany, pp. 270-283, 2005.
 - [32] J. Ma, C. Chen, J. Salomaa, "mWSN for Large Scale Mobile Sensing," *Journal of Signal Processing Systems*, Vol. 51, No. 2, pp. 195-206, 2008.
 - [33] S.A. Munir, et al., "Mobile Wireless Sensor Network: Architecture and Enabling Technologies for Ubiquitous Computing," in *AINAW'07: Proc. 21st Int. Conf. Advanced Information Networking and Applications Workshops*, Niagara Falls, Canada, pp. 113-120, 2007.
 - [34] Bluetooth SIG, "Bluetooth Core Specifications v. 4.0," 2010.
 - [35] International Organization for Standardization, "ISO/IEC 18000-7:2009: Information Technology - Radio Frequency Identification for Item Management," 2009.
 - [36] C. Dugas, "Wavenis ULP Long Range Wireless Platforms, Sensing, and M2M Monitoring Solutions," in *Proc. 1st ETSI Workshop on Machine to Machine (M2M) Standardization*, 2008.
 - [37] Zensys, "Z-Wave Protocol Overview," 2007.

SOA DRIVEN ARCHITECTURES FOR SERVICE CREATION THROUGH ENABLERS IN AN IMS TESTBED

*Mosiuo Tsietsi, Alfredo Terzoli and George Wells**

Department of Computer Science, Rhodes University
Grahamstown, 6140
South Africa

ABSTRACT

Standards development organisations have long been in agreement that the most appropriate and cost effective way of developing services for the IP Multimedia Subsystem (IMS) is through the use — and re-use — of service capabilities, which are the building blocks for developing complex services. IMS specifications provide a theoretical framework for how service capabilities can be aggregated into large service applications. However, there is little evidence that mainstream IMS service development is capability-based, and many services are still designed in a monolithic way, with no re-use of existing functionality. Telecommunication networks are well positioned to stimulate the Internet services market by exposing these service enablers to third parties. In this paper, we marry the two issues by defining an extended IMS service layer (EISL) that provides a service broker that is the central agent in both service interaction management and the execution of external requests from third parties. A prototypical implementation of the service broker is described that was developed using a converged SIP servlet container, and a discussion is also provided that details how third party developers could use HTTP APIs to interact with a service broker in order to gain access to network capabilities.

Keywords— IMS, standardisation, service capabilities, SOA, service broker, SCIM

1. INTRODUCTION

With the advent of the Internet, a multitude of services have become available to everyday consumers. Static forms of communication such as email and FTP have given way to newer technologies such as instant messaging, voice over IP, video on demand, location-based services and various forms of multimedia group communications. Their popularity and demand have grown substantially, spurred on by the increased availability of high bandwidth connections such as

DSL, WiMax and others. These services are typically referred to as over-the-top (OTT) Internet services since they are offered “over” an operator’s existing network.

In response to the influx of these new OTT Internet players, network operators have begun looking at the IP Multimedia Subsystem (IMS) as an architecture that will change the services scene. IMS is an Internet Protocol (IP) based middleware architecture that realises the convergence of the Internet with 3rd Generation (3G) mobile networks. It is anticipated that consumers will be drawn to operator-based services instead of those served on the open Internet for three main reasons: quality of service (QoS) guarantees, flexible billing options and diverse service options through service integration with third parties [1].

When a developer has created an application server that hosts a service and wishes to deploy it in an IMS network, details must be provisioned that allow the serving CSCF (S-CSCF) to know what conditions must be met (initial filter criteria, or iFC) for that application server to be invoked. When the developer wishes to deploy another application server, say a messaging service, a conferencing service or a multi-party virtual classroom for e-learning, the same procedure must be followed each time. IMS is very good at facilitating this, but there is a major drawback with this deployment strategy. Some of the functions that are provided in the messaging service may already be present in the virtual classroom application server. Similarly, the virtual classroom will likely implement some of the functions that are provided by the conferencing application server as well. This strategy is wasteful and introduces significant code redundancy.

The second challenge is that service developers must be well acquainted with the underlying protocols that are used to implement these services, such as how to manipulate a streaming media server using the Realtime Streaming Protocol (RTSP) or how to signal chat messages in session or page mode using the Session Initiation Protocol (SIP).

In order to solve these problems, the standards development organisations that are behind the formulation of IMS specifications have introduced a new paradigm in service creation. This new move involves standardising the basic building blocks for the development of services, known as service capabilities. By doing so, complex services could be dynam-

*The authors wish to thank the following sponsors of the Telkom Centre of Excellence in Distributed Multimedia at Rhodes University for their financial support: Telkom South Africa, Tellabs, Stortech, Easttel, Bright Ideas 39 and THRIP.

ically created by aggregating the service capabilities that are needed to realise them. However given the currently defined structure of the IMS, the interactions that must occur between the service capabilities cannot easily be facilitated. To solve this, a node known as a service capability interaction manager (SCIM) has been defined, but its functional structure has not yet been standardised.

Once enablers are introduced into an IMS network, it would be beneficial to allow third party service developers to explore underlying enablers and use them to add value to existing Internet solutions. This is reminiscent of the SOA (service oriented architecture) approach which allows a service requester (third party) to discover a service provider (service enabler) through a service broker (SCIM). It is anticipated that the SCIM node, from a logical point of view, given that it has direct access to service enablers, could form part of a more complex node known as the IMS service broker.

In this paper we present an extended view of the IMS services architecture called EISL (Extended IMS Service Layer). EISL supports service creation through enablers and service exposure through a service broker. In order to support this architecture, new roles and repositories are introduced, but standard IMS interfaces and protocols are used in the solution. It is our opinion that this architecture will ease the development of services, and through mediation via the service broker, will support the creation of services in a manner that does not require a third party developer to be intimately aware of the semantics of IMS protocols. We also describe the development of a SCIM prototype using the open source Mobicents SIP Servlet container which supports the creation of converged SIP/HTTP servlet applications. Lastly, we describe how service creation looks like in the EISL system from the perspective of a third party developer.

2. RELATED WORK

2.1. Services and Service capabilities

The term service is a highly overused term which has often been used to refer to different things at different times. To address this problem, a work group of the standards development organisation ETSI (European Telecommunications Standards Institute) known as TISPAN (Telecommunications and Internet converged Services and Protocols for Advanced Networking) put forth terminology in [2] to clarify the different uses of the term.

According to TISPAN, a service is composed of service applications. A service application implements a portion of the technical functionality of an overall service. It is composed of elements called service capabilities that can be re-used by other service applications. This approach is desirable since it is the service capabilities and not the service applications themselves that are standardised, which leaves room for service differentiation by vendors without sacrificing innova-

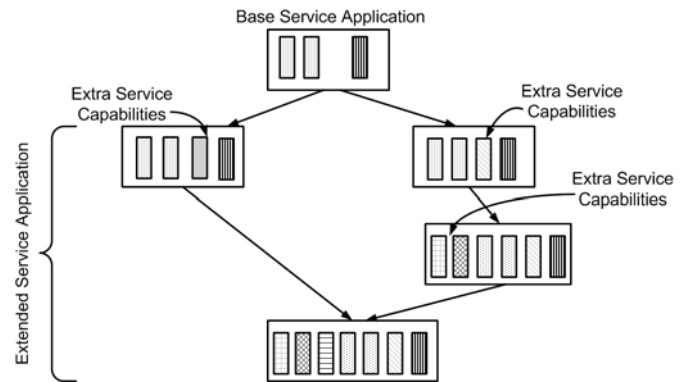


Figure 1. Services and service capabilities. Source: [2]

tion. Figure 1 illustrates the relationship between services, service applications and service capabilities.

This view is reiterated by the standards development organisation 3GPP (3rd Generation Partnership Project) in [3] wherein a classification of services is given. According to this technical specification by the 3GPP, the most basic telecommunication services are divided into two types: bearer services and teleservices. Bearer services involve only low layer functions and are responsible for the transmission of signals between access points. Teleservices provide complete capability including terminal equipment functions for communication between users such as a multimedia conference or multimedia messaging services. Next in the hierarchy are supplementary services which supplement, modify or personalise a basic telecommunication service. This taxonomy forms the basis for defining service capabilities.

2.2. SOA Models of Development

In order to fully realise the benefits of a modular deployment strategy through service capabilities, further interventions are needed. Developers must be able to discover the service capabilities that are embedded in the network in order to develop their services. This implies that service capabilities must somehow be registered with the network. The registry that holds information about the service capabilities must also provide facilities through which it can be interrogated. Also, when a service application interacts with a service capability, it must be able to access that service capability and invoke its functionality somehow. This implies that the service capability (or the network) defines application interfaces that are usable by service applications through which they can interact with each other.

2.2.1. OSA

Open Service Access (OSA) is a framework that provides access to network functions through a standard interface. OSA is often referred to as OSA Parlay since it was originally

specified by the Parlay group (though currently there is joint standardisation of OSA by both 3GPP and ETSI). OSA APIs (Application Programming Interface) are designed to be used by both home network service applications as well as by third party service providers, and can be developed using practically any programming language. The API is able to expose network features such as call control, presence functions and terminal location [4]. A node called a Parlay gateway or service capability server (SCS) sits at the border of the operator domain so as to hide network protocols from the outside world. Web services have become a popular implementation for OSA APIs. A newer version of the Web service API called Parlay X has been developed as a simpler alternative to OSA Parlay.

2.2.2. OMA OSE

The Open Mobile Alliance (OMA) is an international organisation made up of various stakeholders in the mobile services market whose purpose is to develop specifications for interoperable mobile services [5]. OMA is the biggest single contributor to specifications on service capabilities (called service enablers by OMA) and to date has over 100 different technical specifications for enablers such as presence, messaging and device management. Like IMS, OMA also has the objective of allowing secure and policy-driven access to enablers and allowing service providers who use OMA enablers to expose network capabilities to other parties.

To facilitate this, OMA has developed the OMA Service Environment (OSE) with the objective of creating a single consistent architecture that can host multiple service enablers [6]. OSE ensures that silo architectures are avoided, and that integration and deployment complexities are reduced. Figure 2 shows the architecture of the OSE.

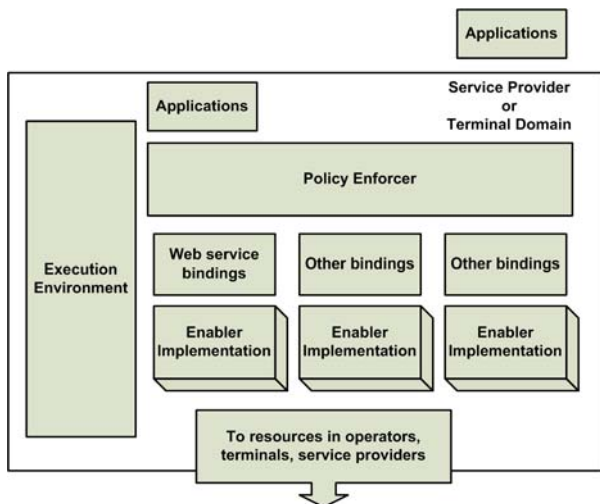


Figure 2. The OMA Service Environment. Adapted from: [6].

As the figure shows, the OSE can incorporate web services through bindings so that the OSE can provide external access

to enabler implementations. Central to the architecture is the role of a policy enforcer that is responsible for policy evaluation and enforcement. Policy evaluation refers to the evaluation of conditions specified in a repository whereas policy enforcement refers to the process of executing actions which may be performed as a consequence of the output of the policy evaluation process. OMA has never provided a specific enabler that should be used to implement this entity, but has suggested that the Policy Evaluation and Enforcement Management (PEEM) enabler or a derivative of it could be tasked with the job [6].

2.2.3. SOA Designs in NGNs

The arrival of Parlay APIs represents a significant step towards the opening up of telecommunication networks. By using web services through Parlay X, programmers do not need to know about the inner workings of complicated telephony protocols such as SIP or intricate details about the networks they are deploying their services to. Web services are not the only way of realising OSA, but they do represent a widely popular service tool for IMS. Figure 3 shows an illustration of the Open SOA Telco Playground that is operated by the Fokus Fraunhofer Institute. The system makes extensive use of web services.

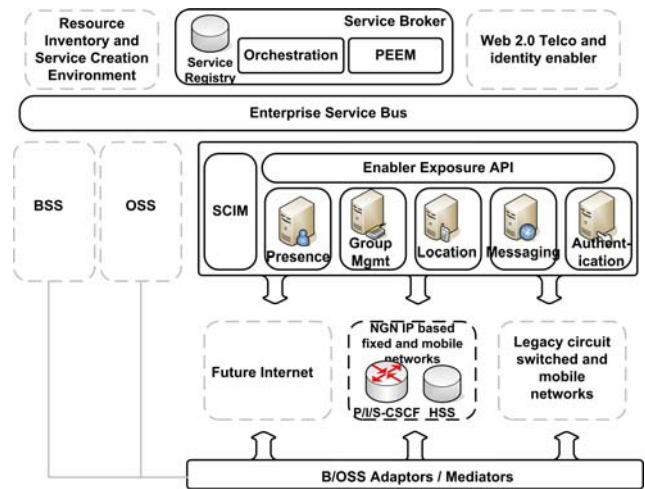


Figure 3. OpenSOA Telco Playground. Adapted from: [7].

OpenSOA is designed to be an open, secure, OSA Parlay based testbed for academia and industry players to experiment with APIs for service integration in IMS, legacy systems and the so-called Future Internet [8]. This initiative by Fokus summarises the evolution of APIs for service integration because it brings together concepts from IMS, service capabilities, OMA enablers and policies, enabler exposure and Parlay X. The Open IMS Core project provides the IMS functions while service capabilities such as presence and location are also provided. The management of the interactions between service capabilities is provided by the service capability interaction manager, or SCIM. Web services are utilised to provide service exposure. A PEEM-like function

exists, in addition to a service registry which can be interrogated by 3rd party service applications.

3. DISCUSSION

IMS implementers are awakening to the importance of “enablerising” their testbeds, that is, to deploy enablers in their network as a core part of service creation. This move will allow them to avoid costly silo architectures and will promote the re-use of provisioned network functionality. This change however, is not without difficulties, since it requires careful attention as to how the management of service capabilities will be performed. This is further compounded by the fact that this function is not clearly specified in the IMS technical specifications. It requires the creation of a service registry that will keep track of the service capabilities that are available in the network and provide the means for network operators to define safe and open APIs that will provide access to them. Additionally, policy handling is required so that access to those service capabilities can be individually evaluated and enforced.

Web services have a long history in telecommunication environments, however, there are alternatives. JAIN (Java APIs for Intelligent Networks) is a set of Java APIs that provide developers with the tools needed to interact with the IMS. When coupled with a container that implements the JAIN SLEE (Service Logic and Execution Environment) Java standard, a consistent environment is created that allows for the use of multiple JAIN stacks along with an execution environment for the execution of service building blocks.

Though it is difficult to say which platform is superior to the other, JAIN SLEE/JAIN does have certain advantages over Web services as mentioned in [9]. For instance, the article states that Parlay APIs have been known to be too detailed or complex in certain cases and not specific enough in other cases. Thus developers have had to use lower level APIs such as JAIN to complete certain tasks. Also, Parlay APIs by themselves do not provide a service execution environment, unlike SLEE. Thus, with SLEE, developers have the ability to create and execute services using a single container.

The Mobicents communication platform, which provides an implementation of the JAIN SLEE standard, has been proposed in previous work as a suitable platform for the creation of a service platform for IMS [10]. Mobicents is an umbrella term for a set of enablers that are developed using JAIN SLEE principles and delivers services such as presence, group management, media services, accounting and billing [11]. The inclusion of a service broker into a network testbed that makes extensive use of Mobicents enablers provides a suitable environment for the development of services that interact with underlying enablers.

In the next section, the architecture of an extended IMS service layer called EISL that extends the standard service layer

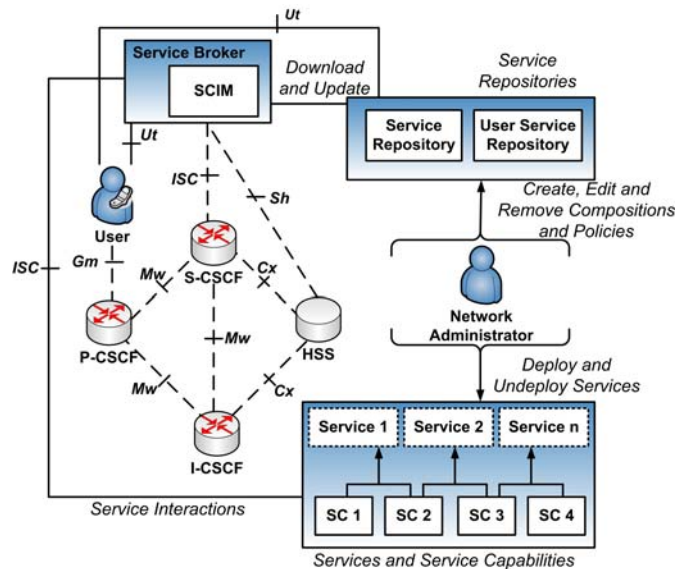


Figure 4. The EISL System.

is presented that satisfies the requirements that have been outlined.

4. THE EISL SYSTEM

Figure 4 shows the structure of the EISL system. The broken lines represent the standard IMS interfaces while the full lines represent new interfaces that are required by the design. The two main constraints defined by EISL are that no or few changes to the standard IMS entities should be effected and that, as far as possible, EISL must re-use existing IMS protocols and procedures for those interfaces that it introduces. A complete examination of the EISL system, including a discussion on a prototypical implementation of a SCIM using Mobicents components is described in [12]. The next few sections describe some of the important parts of the EISL system that help to create an enablerised testbed for IMS.

4.1. Network Administrator

The network administrator is responsible for carrying out network operator tasks. This includes those that would normally be done such as the deployment, monitoring and configuration of application servers. The administrator is also responsible for executing tasks that are necessitated by the EISL design such as provisioning information that relates services to those enablers that are required to compose them (service composition) and the specification of application-level operator policies that relate to services.

4.2. Service Broker

The service broker has interfaces similar to a SIP application server, which are the ISC (IP multimedia Service Control)

interface and the Sh (Diameter) interface. Unlike in [7], the service broker embeds the SCIM as the authors believe that the SCIM functions are a subset of the functions of the service broker and it makes sense from an architectural point of view to co-locate them. The service broker also has an XCAP (XML Configuration Access Protocol) interface with service repositories that are implemented using an XDMS (XML Document Management Servers). XCAP and the XDMS are pervasive technologies in existing IMS services such as presence.

The service broker is responsible for orchestrating services on behalf of the user. During standby mode, when the service broker is not servicing any user requests, it uses offline mechanisms to ensure that it obtains the service compositions and user preferences from the service repositories so it can handle a user request appropriately. During the online stage, the broker queries its internal storage for those compositions and rules and uses them to orchestrate the services. User personalisation is an important service aspect that is catered for in EISL since it allows the customisation of services by users, so long as those customisations do not conflict with operator policies defined by the administrator.

4.3. Service Repositories

The service repositories are storage spaces that contain information regarding the full set of services that an operator provides to its users, including how those services are composed from enablers. In [12], a lengthy comparison is made between the XDMS and the HSS for the purpose of implementing the service repositories. The conclusion that results from this discussion is that the XDMS better realises the two design goals previously mentioned. This is because the data model of the HSS would have to be fundamentally changed since service associations would need to be stored as part of the data stored in the HSS. In particular, the structure of the service profile would need to be changed. XDMS better caters for such extensions since all that is needed is the inclusion of a new application usage (appusage) for XCAP.

There are two types of repositories in EISL. The first is called the service repository which contains the global set of service information. Included in this data are the policies that are specified by the network operator which address the legal interactions between service capabilities and any other constraints on their use.

The second repository is called the user service repository which is introduced in order to cater for the personalisation of services by individual users. The information contained in this repository is similar in structure to that which is found in the service repository. In fact, the process of personalisation involves copying some information from the service repository into the user service repository. The difference is that the user service repository is compartmentalised into user directories that pertain to individual users. This means that operator policies are copied into these directories and must

be adhered to irrespective of user customisations. Users can specify additional rules in the form of their own preferences that detail how they would like their services to be composed or to behave during runtime.

4.4. Implementation using Mobicents and open Source Software

The combined use of the Open IMS Core coupled with Mobicents as a service platform, provided the features needed to implement a prototype of the EISL system.

Mobicents provides a presence service which comprises a presence server and an XDMS. The XDMS can be deployed in standalone mode. The Mobicents SIP Presence 1.0.0.BETA 6 version of the XDMS was used to implement the service repositories. To provide an XML schema for the storage and management of XML documents to be stored in the XDMS, the ETSI-defined sirmservs appusage was used [13]. The sirmservs appusage supports the definition of complex services that are composed of simpler, supplementary services that can be used to create complex multimedia telephony applications. Due to their similarity with service capabilities, the appusage provided a useful template in the absence of a standard appusage that is explicitly defined for service capabilities. The sirmservs appusage supports the inclusion of IETF (Internet Engineering Task Force) common policy and OMA common policy schemas which allow the definition of operator and user policies by allowing developers to insert elements that define conditions and actions.

The implementation of the service broker was limited to the functions of the SCIM. For that, the Mobicents SIP Servlet container provides a suitable platform since it supports the interfaces that are defined for the SCIM which are SIP, HTTP and Diameter. The Mobicents SIP Servlets 1.4.0.FINAL version was used in the prototype. A servlet application was developed that consisted of two servlets that implemented online and offline interaction management functions. A separate servlet was used to mimic the actions of a user interacting with the service repositories since no IMS client could be found that supported the sirmservs appusage.

A servlet container by definition must provide a servlet application with what is called a servlet context which stores state that can be shared between servlets in that servlet application. The servlet context was used to maintain state that is needed by the SCIM to execute online interaction management after the offline stage had completed. The offline stage provided the online stage with service compositions, operator rules and user preferences. The service capabilities were implemented using instances of a Kamailio (OpenSER) server that mimicked simulation services. Figure 5 shows the architecture of the SCIM portion of the prototype.

In an experiment, the `ClientSimulator` servlet was used to create the document shown in Listing 1. The

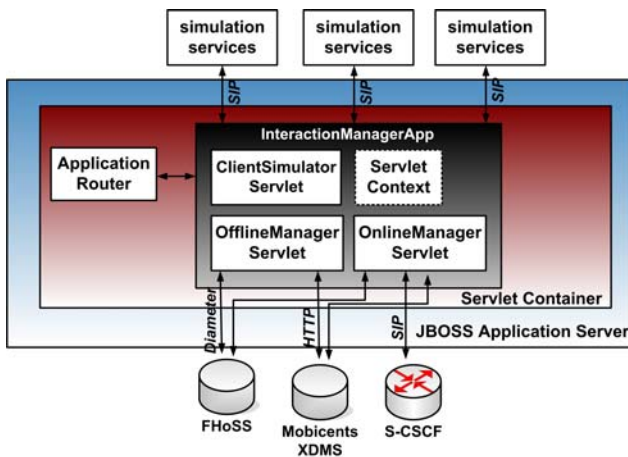


Figure 5. Architecture of the SCIM.

document corresponds to the configurations of a complex multi-component telephony service, for which the user has subscriptions to an originating identity presentation restriction service as well as an outgoing communication barring service. Both these services are simulation services. The priorities indicated in the service descriptions show the invocation order of those services. These are set by the operator. The ruleset description that is part of the outgoing communication barring service description shows how the user has specified their own preferences for how the telephony service should behave.

5. SERVICE DEVELOPMENT IN EISL

The SCIM shown in Figure 5 provides an interface between applications and the enablers that are embedded in the network. Provisioned functionality in the form of enablers simplifies the job of the service integrator since all that is required of them is to write a minimal amount of code.

In order to interact with the servlet-based SCIM function, there are two possible methods. One of them is to use the SIP protocol itself. Developers would be able to use SIP messages targeted at the SCIM to make requests that would be handled by it to deliver certain functions. The challenge with this approach is that developers still need to know much about the SIP protocol in order to issue service requests. SIP is more suited as a signaling protocol and not as an API.

As a converged container, Mobicents SIP Servlets also supports the use of HTTP. HTTP has advantages over SIP in this usage scenario. For instance, it is better known and understood than SIP and can be used by the many developers who have already been exposed to the HTTP protocol.

In terms of the existing architecture, the same repositories that are currently defined by EISL could be used in order to capture policies that relate to access by third party developers to the service enablers through their applications. Common policy extensions that define conditions and actions can be

```
<?xml version="1.0" encoding="UTF-8"?>
<simservs
  xmlns="http://uri.etsi.org/ngn/params/xml/
  simservs/xcap"
  xmlns:cp="urn:ietf:params:xml:ns:common-policy"
  xmlns:ocp="urn:oma:xml:xdm:common-policy">

  <originating-identity-presentation-restriction
    active="true" priority="1">
    <default-behaviour>presentation-not-restricted
    </default-behaviour>
  </originating-identity-presentation-restriction>

  <outgoing-communication-barring active="true"
    priority="2">
    <cp:ruleset>
      <cp:rule id="rule66">
        <cp:conditions>
          <cp:identity>
            <cp:one id="sip:mallory@open-ims.test"
              />
          </cp:identity>
        </cp:conditions>
        <cp:actions>
          <cp:allow>false</cp:allow>
        </cp:actions>
      </cp:rule>
    </cp:ruleset>
  </outgoing-communication-barring>
</simservs>
```

Listing 1. A simservs user document.

applied to guide the service broker on how to handle a request that originates from an application server.

Figure 6 shows the nature of the interactions that occur in EISL for this service brokering to be performed. The administrator publishes information to the XDMS about the services hosted on the network, including the policies that are to be applied to different application servers, or sets of application servers. A request is received by an interceptor module in the servlet container from an external application server. If the message has not been processed yet, it is passed to a policy manager that examines the policies that need to be applied. If the policies have not yet been downloaded for that application server, the policy manager queries the XDMS. Once downloaded, the policy manager is able to determine the rules that should apply, which will then influence the behaviour of the SCIM when it performs interaction management during an online service request. Work on extending the existing implementation is still ongoing.

6. CONCLUSION AND FUTURE WORK

This paper has described the EISL system as part of an IMS network and has shown how it can be used as a suitable platform for supporting the deployment of services that make extensive use of service enablers. The paper has also married this new paradigm in service creation with the need to expose service enablers to third parties, using the service broker as

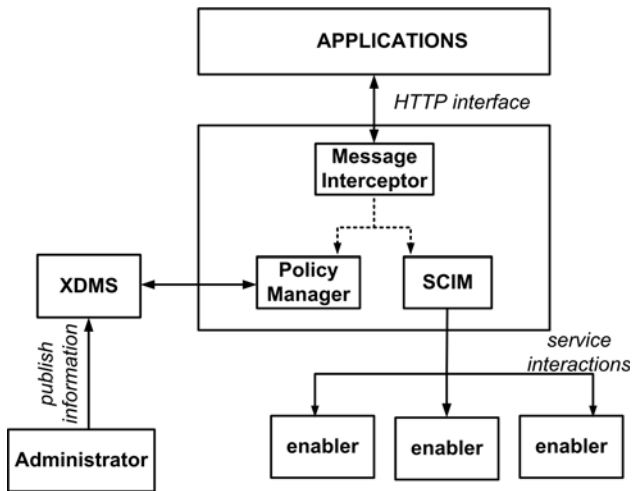


Figure 6. Brokering interactions with application requests.

the enabling agent for this. The service broker is the centerpiece of EISL and is a multi-protocol application server. A prototype was described that took the form of a converged servlet application consisting of both SIP and HTTP components. HTTP can be used as a method through which developers can interact with the service broker using any platform available to them, and through this, the service broker can act upon application requests, applying necessary policies, and expose services to developers in that manner.

This paper does not outline the structure of the API that can be used as an interface between the service broker and the application servers. This will be for future work and will address features that are currently offered in web service APIs such as call control, messaging and presence.

REFERENCES

- [1] G Camarillo and M Garcia-Martin. *The 3G IP Multimedia Subsystem (IMS)*. John Wiley and Sons, Ltd, Third edition, 2008.
- [2] ETSI. TR 181.004: NGN Generic Capabilities and their use to Develop Services. European Telecommunication Standards Institute, March 2006.
- [3] 3GPP. TS 22.105: Services and Service Capabilities. Third Generation Partnership Project, December 2008.
- [4] N. Ajam. Privacy based access to parlay x location services. In *Networking and Services, 2008. ICNS 2008. Fourth International Conference on*, pages 204 –210, March 2008.
- [5] OMA. About OMA. Available Online, June 2010. URL: <http://www.openmobilealliance.org/AboutOMA/Default.aspx>.
- [6] OMA. OMA Service Environment Version 1.0.5. Open Mobile Alliance, October 2009.
- [7] N Blum, T Magedanz, F Schreiner, and S Wahle. Service Oriented Testbed Infrastructures: a CrossLayer Approach for NGNs. *Mobile Networks and Applications*, 15(3):413–424, June 2010.
- [8] T Magedanz, K Knuttel, and D Witaszek. Service Delivery Platform Options for Next Generation Networks within the National German 3G Beyond testbed. In *SATNAC '04: Proceedings of the 7th South African Telecommunications Networks and Applications Conference*, September 2004.
- [9] OpenCloud. OSA Parlay and Parlay-X. Available Online, May 2011. <https://developer.opencloud.com/devportal/display/RD-2v0/2.3+OSA+Parlay+and+Parlay-X>.
- [10] M Tsietsi, A Terzoli, and G Wells. Mobicents as a Service Deployment Environment for OpenIMSCore. In *SATNAC '09: 12th Southern African Telecommunications Networks and Applications Conference*, September 2009.
- [11] Red Hat. Mobicents - The Open Source SLEE and SIP Server. Available Online, May 2011. URL: <http://www.mobicents.org>.
- [12] M Tsietsi. *A Structural and Functional Specification of a SCIM for Service Interaction Management and Personalisation in the IMS*. PhD thesis, Rhodes University, South Africa, August 2011.
- [13] 3GPP. TS 24.623: Extensible Markup Language (XML) Configuration Access Protocol (XCAP) over the Ut interface for Manipulation Supplementary Services. Third Generation Partnership Project, December 2009.

A VIRTUALIZED INFRASTRUCTURE FOR IVR APPLICATIONS AS SERVICES

Fatna Belqasmi^{#1}, Christian Azar^{#2}, Mbarka Soualhia^{*3}, Nadja Kara^{*4}, Roch Glitho^{#5}

[#]Concordia University, Canada

¹fbelqasmi@alumni.concordia.ca

²ch_azar@encs.concordia.ca

⁵glitho@ece.concordia.ca

^{*}ETS, University of Quebec, Canada

³mbarka.soualhia.1@ens.etsmtl.ca

⁴Nadja.Kara@etsmtl.ca

ABSTRACT

Interactive Voice Response (IVR) applications (e.g. automated attendant) are ubiquitous nowadays. Cloud computing is an emerging multi-faceted paradigm (Infrastructure as a Service – IaaS, Platform as a Service – PaaS, and Software as a Service – SaaS) with several inherent benefits (e.g. resource efficiency). Very few, if any, IVR applications are offered today in cloud settings despite all the potential benefits. This paper introduces a novel architecture for a virtualized IVR infrastructure and demonstrates its potential with a case study. The architecture proposes IVR substrates that are virtualized, composed, and assembled on the fly to build IVR applications. It relies on a business model which introduces the IVR substrate provider as a new role in the cloud business model. In the case study, which includes a prototype, IVR service providers develop and manage IVR applications using a simplified platform that adds a level of abstraction to the substrates available in the virtualized infrastructure. The applications are offered as SaaS to end-users.

Keywords— automated attendant, cloud computing, Everything as a Service, IVR, Infrastructure as a Service, virtualization

1. INTRODUCTION

Interactive Voice Response (IVR) enables interactions with automated information systems. Its applications are numerous. One example is automated attendants, which replace live receptionists by transferring callers to the extensions they dial. Another example is automated meter readers that provide fully automated dialogs, which enable utilities customers to remotely enter their meter readings. There are several other examples including automated surveys, bank tellers and clinical trials.

Cloud computing is a promising paradigm with many inherent advantages, such as the easy introduction of new applications, resource efficiency, and scalability. There is

not yet a standard technical definition for cloud computing. However, a consensus is emerging around the most critical facets it encompasses: Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS) [1].

Service providers use platforms (offered as PaaS by platform providers) to develop and manage applications. The applications are provisioned to end-users (or other applications) as SaaS on a pay-per-use basis. Platforms add one or more levels of abstraction to the infrastructures offered as IaaS by infrastructure providers. They ease application development and management. Infrastructures are the actual dynamic pool of virtualized resources used by applications.

Virtualization enables the co-existence of entities in general on the same substrates. These entities may be operating systems co-existing on the same hardware, applications co-existing on the same operating system, or even full-blown networks co-existing on the same routers. The key benefit is efficiency through the sharing of physical resources.

Several applications are offered today in cloud settings (e.g. enterprise databases, IT help desks). However, these offerings very rarely include IVR applications, despite the obvious potential benefits, especially efficiency and scalability. For example, an automated dialog manager offered as a substrate could be virtualized and shared by automated attendants, meter readers, and several other applications. It could also be dynamically allocated to these applications for scalability purposes. An announcement player is another example of a potential IVR substrate. It is finer-grained than the automated dialog manager. It could also be virtualized, shared and dynamically allocated to applications.

To the best of our knowledge, there is no full-fledged cloud environment that enables the development, management and offering of the full range of IVR applications. This paper proposes a novel architecture for a virtualized IVR infrastructure as a first step towards the deployment of full-fledged IVR applications in cloud settings. The architecture relies on a business model which introduces the IVR

substrate provider at the infrastructure layer as a new role in the cloud business model.

The architecture's potential is demonstrated by a case study, focused on a virtualized IVR infrastructure with a selected set of substrates. It shows how IVR service providers can develop and manage simple IVR applications that rely on the substrates provided by the infrastructure. The next section introduces the proposed architecture. The case study is discussed in the third section. The fourth section is devoted to related work. We conclude in the last section by a summary.

2. PROPOSED ARCHITECTURE

Figure 1 depicts our vision. The bottom layer shows a simplified IVR IaaS layer with the following substrates: announcement player, voice recorder, key detector, extension detector, call transfer. At the top layer, we have a simplified SaaS layer with applications such as automated attendant, automated meter reader, automated survey, and IVR banking that share the substrates. The middle layer is the platform layer. It includes graphical user interfaces (GUIs) and application programming interfaces (APIs) that may ease the development and management of the applications in the top layer. The design goals of our architecture and the proposed business model are presented next. Architectural components and interfaces are then discussed. We end the subsection with a brief introduction to RESTful Web services, a key technology of our architecture.

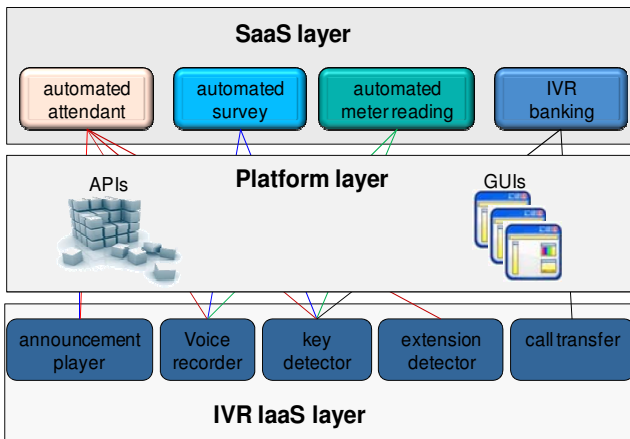


Figure 1. Different services can share the same substrates

2.1. Design goals and business model

One of the first design goals we have in mind is that different IVR applications in different domains should be able to share substrates, as illustrated by figure 1. Conversely, an IVR application should also be able to use many instances of a same substrate, for scalability. Another design goal is that it should be possible to publish and discover substrates and substrate instances. Yet another goal is that IVR service providers should be able to compose the substrates available in the infrastructure into

powerful IVR applications, using appropriate platforms. Figure 2 shows the proposed business model.

The IVR service provider offers IVR applications as SaaS, accessible by end-users and other applications. It develops and manages these applications using the IVR platform offered by IVR platform providers. The platform adds levels of abstractions to IVR infrastructures to ease IVR application development and management by IVR service providers.

IVR substrates are offered by IVR substrate providers to IVR infrastructure providers. A given IVR substrate provider may interact with several IVR infrastructure providers and offer them the same substrates, since these substrates are sharable. A given IVR infrastructure provider may also interact with several IVR substrate providers. The broker enables the publication and discovery of substrates. The connectivity provider enables connectivity between the different actors.

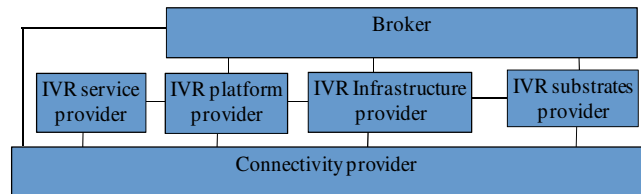


Figure 2. The business model

2.2. Architectural components and interfaces

Figure 3 shows the proposed architecture. It comprises two layers, three planes architecture and a repository. The first layer contains the functional entities that realize the infrastructure provider role. The second layer is comprised of entities that realize the IVR substrate provider role. The interactions between the two layers are organized via three planes: service, management and composition. The repository realizes the role of the broker.

The key functional entity of the first layer is the virtual IVR engine; the key entity of the second layer is the substrate IVR engine. The virtual IVR engine coordinates the activities of the virtual service engine, the virtual management engine and the virtual composition engine. The substrate IVR engine coordinates the activities of the substrate service engine, the substrate management engine, and the substrate composition engine. The virtual IVR engine interacts with several substrate IVR engines--or more precisely, it interacts with the engines of all the substrates that make up a given composed service.

The main functionality handled in the service plane is mediation. IVR infrastructure providers may decide to make substrates available to platform providers using interfaces other than the original interfaces with which they were made available by substrate providers. In addition to mediation, the service plane also coordinates the execution of services that involve several substrates.

The management plane handles the actual control and management of substrate resources. It enables the instantiation of IVR applications and related substrates, and

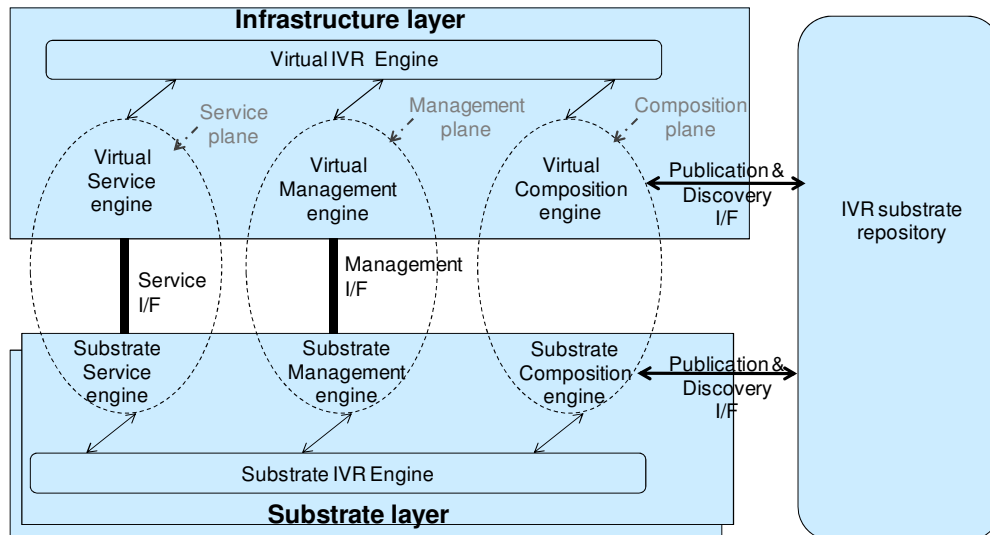


Figure 3. The overall architecture

their configuration. It also enables fault monitoring, performance monitoring, and the accounting for charging purposes. The composition plane interacts with the repository and enables the publication and discovery of the substrates and substrate instances that are used in composition.

The virtual service engine and the substrate service engine communicate via the interface supported by the substrate service engine. This is motivated by the fact that existing potential IVR substrates come with a plurality of interfaces (e.g. VoiceXML, Session Initiation Protocol-SIP, Media Server Control Markup Language-MSMCL). They communicate with the virtual service engine via mediators that are incorporated in the virtual service engine.

A key requirement for the management interface is to accommodate a plurality of resource description mechanisms (e.g. XML, plain text). Another requirement is that the interface should be supported by commonly used virtualization servers such as XEN to ease the creation of instances. These requirements have led to our selection of RESTful Web services as the natural choice.

We have also decided to use RESTful Web services for the publication/discovery interfaces to minimize the number of interfaces in our proposed architecture. The fact that RESTful services support a wide range of resource description mechanisms is also an advantage when it comes to publication and discovery. The next sub-section provides more information on RESTful Web services, since it is a key technology of our architecture.

2.3. A brief overview of RESTful Web services

RESTful Web services follow the Representational State Transfer (REST) design paradigm. REST uses the Web's basic technologies (e.g. HTML, XML, HTTP, URIs) as a platform to build and provision distributed services. It is one of the players of Web 2.0, a concept that promotes interactive information sharing and collaboration over the

Web, as well as Web application consumption by software programs. REST adopts the client-server architecture. REST does not restrict client-server communication to a particular protocol, but more work has been done on using REST with HTTP, as HTTP is the primary transfer protocol of the Web.

RESTful Web services can be described using the Web Application Description Language (WADL [3]). A WADL file describes the requests that can legitimately be addressed to a service, including the service's URI and the data the service expects and serves.

REST supports a wide range of representation formats, including plain text, HTML, XML and JavaScript Object Notation (JSON). JSON is an open standard data interchange format for representing simple data structures (e.g. linked lists) and associative arrays (i.e. a collection of pairs (keys, values)).

RESTful Web services are perceived to be simple and easy for clients to use because REST leverages existing well-known Web standards (e.g. HTTP, XML) and the necessary infrastructure has already become pervasive. RESTful Web services' clients (i.e. HTTP clients) are simple and HTTP clients and servers are available for all major programming languages and operating system/hardware platforms.

3. A CASE STUDY

We start off this section with the scope, assumptions and features of our case study, followed by the software architecture. The last sub-section describes the prototype.

3.1. Scope, assumptions and key features

The infrastructure of our case study is composed of the five substrates shown in figure 1 (i.e. announcement player, voice recorder, key detector, extension detector, call transfer). We assume that the substrates are supplied by substrate providers SubP1, SubP2, SubP3, SubP4, and

SubP5, respectively. We further assume that we have one infrastructure provider (InfP), one platform provider (PP), two service providers (ServP1, ServP2) and one end-user with a subscription to one of the two service providers.

The substrates are described using WADL (as implied by our choice of RESTful Web services as technology), and also with Donkey State Machine (DSM) [4]. The DSM description represents the substrate behaviour as a state machine and is included under the <doc> element of the WADL description. The use of DSM is motivated by the fact that it is used by the SIP Express Media Server (SEMS) [5] used in our prototyping environment, and because it makes the substrates' composition easier. DSM enables a textual description of applications (substrates in our case) that can be directly executed by interpreters hosted by the SEMS. Figure 4 shows a simplified WADL description of the 'Announcement Player' substrate.

The case study covers application development (including the pre-required publication and discovery of substrates), management and execution. The management phase is restricted to activation. Figure 5 depicts these three phases. In figure 5.a, the substrate composition engine of each substrate provider publishes its substrate to the repository using a put request, with the WADL description of the substrate as argument. Next, the virtual composition engine of the infrastructure provider sends a get request to get the list of available substrates with their descriptions. It should be noted that the get request could have been sent anytime during the process to get the list of substrates available at that point in time. The list is then made available to the platform engine that adds a level of abstraction to the

substrates by making them visible through a GUI. The first service provider uses the GUI to develop an automated attendant by composing the substrates, while the second one develops an automated survey by the same process.

```
<?xml version="1.0"?>
<application
  xmlns:xsi:shemaLocation="http://wadl.dev.java.net/2009/02">
  <doc xml:lang="DSM" title="Play announcement substrate service">
    state Play enter{playFile(/home/user/welcome.wav); };
  </doc>
  <resources base="http://substrateProvider1.com/">
    <resource path="playAnnouncement">
      <method name="POST" id="instantiate">
        <request>
          <representation mediaType="application/xml" >
            <param name="dsm_description" type="xsd:string"
              required="true"/>
            <param name="serviceProvider" type="xsd:string"
              required="true"/>
          </representation>
        </request>
        <response status="200">
          <representation mediaType="application/xml" >
            <param name="resourceURI" type="xsd:anyURI"
              required="true"/>
          </representation>
        </response>
      </method>
    </resource>
  </resources>
</application>
```

Figure 4. WADL description of the 'Announcement Player'

Figure 5.b. describes the flow for application activation. The infrastructure provider's virtual composition engine

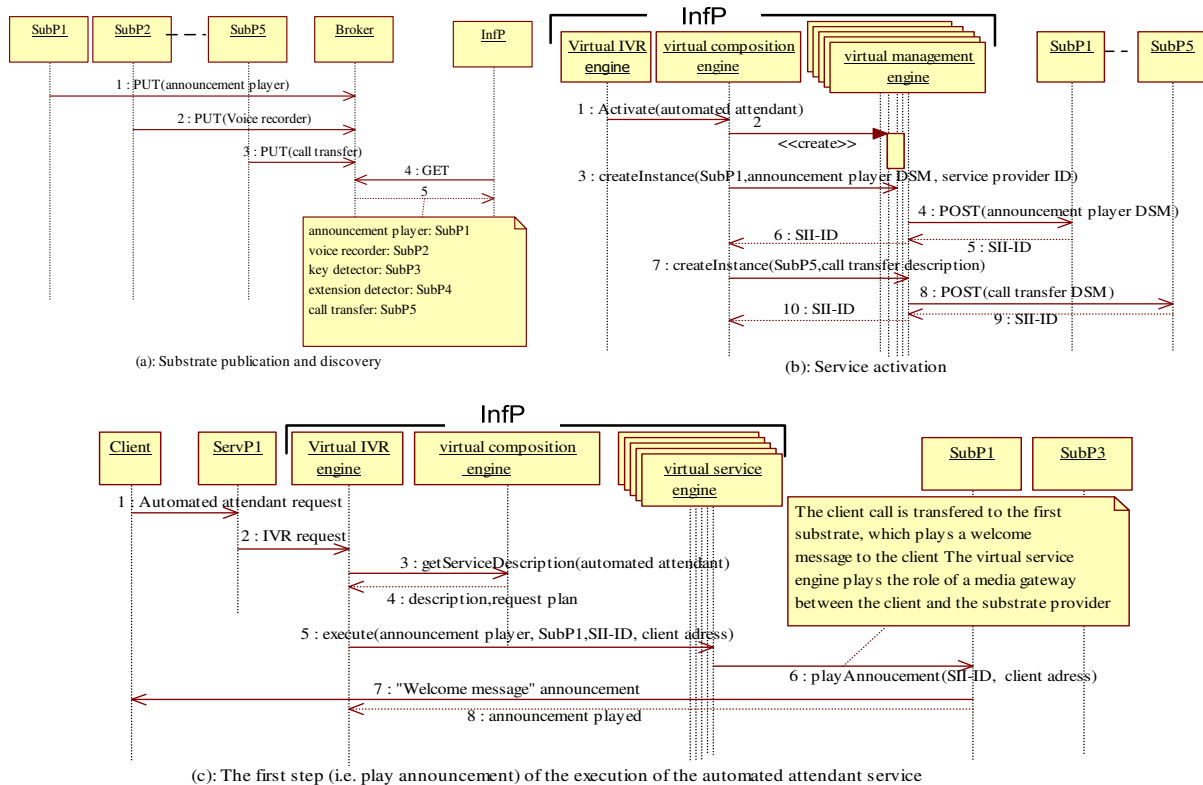


Figure 5. Application development, management and execution phases

uses the description of the composed service (created in the previous step) and creates a different instance of the virtual management engine to communicate with each of the substrates involved in the composition. It then instructs these instances to create a substrate IVR instance (SII) at each substrate. The instantiation is done by sending a post request to the appropriate substrate, along with the DSM description of the service instance to create and the identifier of the IVR service provider. When a substrate management engine receives an instantiation request, it verifies resource availability and then creates a new SII and allocates the necessary resources.

Figure 5.c. presents the execution flow for the automated attendant service. The virtual IVR engine of the infrastructure provider receives an IVR request, gets the service description from the virtual composition engine (including the SII's to use), creates a virtual service engine instance to communicate with each of the substrates, and instructs the different instances to execute the appropriate sub-requests, following the request plan. A request plan is a set of sub-requests and their execution sequence, along with the relevant substrates/SII's that are required to answer an IVR request. The request plan is created by the virtual composition engine during the service creation phase.

When a substrate service engine receives a service execution request, it forwards the request to the appropriate SII, which then executes the request and replies back to the virtual service engine.

3.2. Software architecture

Figure 6 presents the software architecture, with a focus on the IVR substrate repository and the composition and management planes.

The IVR substrate repository includes a publication manager and a discovery manager that handles the publication and discovery requests, respectively. The published service descriptions are stored in a local database.

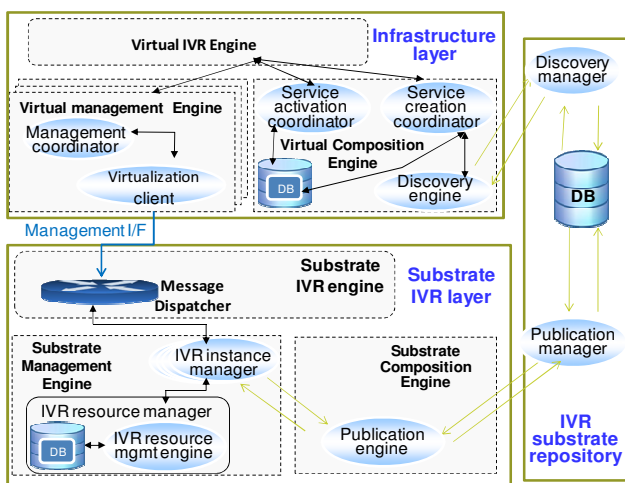


Figure 6. Software architecture

The virtual composition engine includes three functional entities: the service creation coordinator, the discovery

engine and the service activation coordinator. The service creation coordinator gets the list of available substrates from the broker (using the discovery engine) and sends them to the platform provider. It also takes the inputs from the platform provider GUI for the service composition, creates the description file for the composed service, and stores the description in a local database. The service activation coordinator manages and coordinates the instantiation of new SII's.

The virtual management engine includes two entities: the management coordinator and the virtualization client. The management coordinator translates the requests received from the service activation coordinator into requests that the virtualization client should then send to the target substrate.

At the substrate IVR layer, each SII is managed by a separate IVR instance manager. The message dispatcher dispatches the received messages to the appropriate IVR instance manager. The management messages for a given SII are forwarded to the IVR instance manager that created it. The relationships between service instances and their managers are saved when the service instances are created. The IVR resource manager maintains and monitors the current states of the resources and allocates resources for new SII's.

3.3. Prototype

As a prototype, we implemented the scenario where the first service provider creates and provisions an automated attendant service. For the prototype, we assume that the five substrate services are offered by the same substrate provider. They are implemented using DSM and deployed on a SEMS server. The REST interface of the substrate management engine is implemented using Jersey, an open source reference implementation of JSR 311. The interface module is deployed on a Glassfish server and it communicates with the SEMS using sockets. The implementation of the IVR substrate repository is also based on Jersey and deployed on a Glassfish server.

Figure 7 presents the GUI offered by the platform provider. The GUI shows the existing (substrate) services, discovered by interrogating the substrate repository after the service provider pushes the "Discover" button on the GUI. The GUI allows the service provider to create its composed service by choosing the substrates to compose and then ordering them graphically. When the "Compose" button is pushed, a request is sent to the service creation coordinator in order to generate the DSM description of the composed service. The service creation coordinator also generates the execution plan for the composed service and stores the DSM description and the execution plane in the local database. The figure shows the creation of the automated attendant service. The client calls a company's generic number, gets an announcement inviting him/her to enter the extension of the person to reach, the client provides the extension and is then connected to that person. If that person does not answer, the client can leave a voice message. After the composed service is created, the service

provider can publish it to the substrate repository, so that other service providers or clients can use it.

Since the five substrate services are offered by the same provider, the composed service is also offered by the same provider and the execution plan is the same as executing a simple (i.e. non-composed) service.

At the service plane, the composed IVR service is accessed via SIP. A free SIP client, X-Lite, is used as the IVR client. To initiate the service execution, X-Lite sends a SIP INVITE to its service provider, which proxies the call to the substrate service engine via the virtual service engine.

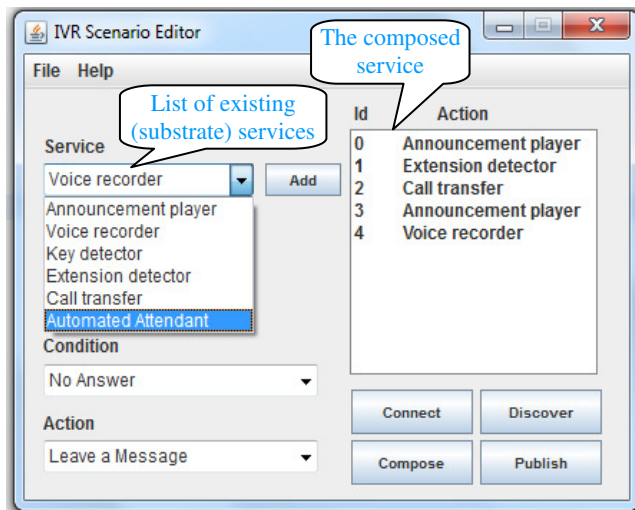


Figure 7. The platform provider GUI

4. RELATED WORK

Several commercial hosted IVRs such as call centres allow users to remotely access IVR applications as services. However, to the best of our knowledge these applications do not rely on virtualized infrastructures. Embryonic architectures have been proposed that could offer a few applications, which may include selected IVR features, as services in clouds. One example is conferencing that can include announcement player and digit collector.

Reference [6] proposes a video conferencing application as SaaS, along with the required supporting virtualized infrastructure. The goal is to ease the integration of video conferencing with other applications. A previously developed Web-based conferencing application is used as the base. The infrastructure described is limited to one virtual full-blown video conferencing server.

The key drawback of the proposal is that it follows a very coarse-grained approach. It does not provide an infrastructure with substrates that can be published, discovered and dynamically shared as we envision in the architecture proposed in this paper. Furthermore it does not

include any IVR feature. Reference [7] also discusses audio/video conferencing as SaaS, but shares the drawbacks of the above proposal in terms of the infrastructure. No fine-grain substrate is proposed and the entire conferencing server is virtualized.

Substrates have been proposed for virtualized infrastructures that support applications with no bearing on IVR. One example is the infrastructure we proposed for presence [8]. That proposed architecture includes a business model. However, the work is still at a preliminary stage and thus far only one substrate has been identified for the application. Publication, discovery and composition have therefore not yet been addressed.

Several virtualized infrastructures have been proposed for Future Internet [9, 10, 11]. These rely on virtualized nodes and links as substrates. However, they focus on the core infrastructure of the Internet and do not address issues related to the virtualization of the edge infrastructure of the Internet, where applications such as those of IVR reside.

A few architectures do deal with some of the challenges related to application development by composition in a cloud environment. However, none of them addresses the specific case of IVR applications. Reference [12] provides an example. It targets applications at large and proposes a virtualized mash up container. Application substrates, called service components in the proposal, are described using existing technologies such as REST, JSON, and RSSFeed. A service proxy maps these descriptions into the internal description technology supported by the container.

Unfortunately, little detail is provided on this internal format. Furthermore, the specific publication and discovery mechanisms are not discussed. Reference [13] provides another example. It proposes an infrastructure where music content providers federate with value-added service providers (e.g. security, audit) to build virtual music stores on the fly. Several registries (e.g. client, capability registry) are mentioned in the paper. However, no detailed information is provided about how the content of these registries is described, published and discovered.

5. CONCLUSIONS

This paper proposes a novel architecture for a virtualized IVR infrastructure. The architecture allows different IVR service providers to share the same IVR substrates, and enables easy development and management of new IVR-based applications via a simplified platform. The paper also includes a case study that shows how a new service (i.e. automated attendant) can be created by composing a number of existing simple IVR services provided by the infrastructure. A proof of concept prototype was also implemented.

REFERENCES

- [1] L. M. Vaquero et al., “A Break in the Clouds: Towards a Cloud Definition”, *ACM SIGCOMM Computer Communication Review*, Vol. 39, No1, January 2009
- [2] L. Richardson and S. Ruby, “RESTful Web Services”, *O’Reilly & Associates*, ISBN 10: 0-596-52926-0, May 2007
- [3] W3C Member Submission, “Web Application Description Language”, 31 August 2009
- [4] Donkey State Machine (DSM), http://ftp.iptel.org/pub/sems/doc/current/ModuleDoc_dsm.html
- [5] SIP Express Media Server, <http://www.iptel.org/sems>
- [6] P. Rodriguez et al., VaaS: “Videoconferencing as a Service”, *5th International Conference on Collaborative Computing: Networking, Application and Worksharing*, 2009.
- [7] J Li, R Guo and X. Zhang, “Study on Service Oriented Cloud Conferencing”, *Third IEEE International Conference on Computer Science and Information Technology*, pp. 21–25, July 2010
- [8] Fatna Belqasmi, Nadjia Kara, Roch Glitho, “A novel virtualized presence service for future Internet”, *Workshop on Future Networks, IEEE International Conference on Communications(ICC2011)*, pp. 1-6, June 2011
- [9] T. Aoyama, “Overview of the new generation network R&D in Japan”, *Proceedings of the 4th International Conference on Future Internet Technologies (CFI’09)*, Seoul, Korea 2009
- [10] K. Tutshuku et al., “Network Virtualization: Implementation Steps Towards the Future Internet”, *WowKIVS 2009*, 2009
- [11] J. Carapinha; J. Jiménez, “Network virtualization: a view from the bottom”, *Proceedings of the 1st ACM workshop on Virtualized infrastructure systems and architectures (VISA-09)*, pp. 73-80, 2009
- [12] M. Stecca and M. Maresca, “An Architecture for a Mashup Container in Virtualized Environments”, *IEEE 3rd International Conference on Cloud Computing*, pp.386-393, 2010
- [13] P. de Leusse et al, “Secure and Rapid Composition of Infrastructure Services in the Cloud”, *Proceedings of the Second International Conference on Sensor Technologies and Applications (SENSORCOMM ’08)*, pp. 770-775, 2008

SEAMLESS CLOUD ABSTRACTION, MODEL AND INTERFACES

Masum Z. Hasan, Monique Morrow, Lew Tucker

Cisco Systems
San Jose, CA USA

Sree Lakshmi D. Gudreddi, Silvia Figueira

Dept. of Computer Engineering
Santa Clara University
Santa Clara, CA USA

ABSTRACT

An enterprise, as a Cloud Service Consumer (E-CSC), may acquire and consume (off-premises) resources in one or more Public or Community Clouds owned and operated by one or more Cloud Service Providers (CSP). A CSP (as a CSC: S-CSC) may itself consume resources from other CSPs on behalf of an E-CSC. For seamless manageability an E-CSC may want to combine a select set of on-premises (intranet or private Cloud) resources with off-premises Cloud resources to create a Seamless Cloud (SCL). The E-CSC may also include in the SCL a select set of its (branch and DC) sites. Based on the definition an SCL subsumes various categories of Cloud, such as private, public, community, hybrid and inter-Cloud. A CSP can offer a service, which we call the Seamless Cloud service that will facilitate creation, deletion and update of an SCL on-demand. In a multitenant Cloud environment SCLs of each tenant should be isolated from each other end-to-end (from CSC enterprise to on-demand acquired resources in CSP DC). The SCL service will facilitate such isolation. By adding proper QoS capability to the SCL service, a CSP will be able to offer (what we call) Differentiated Quality of Seamless Cloud Services (DQSCS). In this paper we describe abstraction, model and interfaces (CSC to CSP) for SCL. It is expected that the interfaces will be standardized.

Keywords— Seamless Cloud, Model, Interfaces

1. INTRODUCTION

Cloud Computing (Cloud) has emerged as a technology that is changing the IT, datacenter (DC), and networking landscape in a major way. A Cloud Service Provider (CSP) offers Cloud services out of one or more DCs, where compute, storage, and network resources are offered on-demand to Cloud Service Consumers (CSC). In other words, the DC infrastructure is not a static entity anymore as in a traditional DC, rather an entity that is offered as a service called the Infrastructure as a Service (IaaS). With the IaaS a CSC can acquire and release resources on-demand and elastically (grow or shrink at will). The Cloud resources are offered under a pay-as-you-go pricing model. In addition to the infrastructure resources, the software resources are also offered via Platform as a Service (PaaS)

and Software as a Service (SaaS) in the same elastic, on-demand and pricing models. In the case of PaaS software development and testing platforms and software middleware components are offered as on-demand resources. In the case of SaaS, full-fledged application products are offered as on-demand resources.

As indicated above, two of the major actors in a Cloud are CSC and CSP. Before we proceed to describe the Seamless Cloud concept, let us look at Cloud from the CSC and CSP perspectives.

1.1. Cloud Service Provider (CSP)

A CSP is an entity that offers Cloud services out of one or more Cloud Data Centers. A CSP may also own or operate a private MAN or WAN via which an enterprise CSC may connect to the CSP operated Cloud in addition to connecting to the Cloud via the public Internet. A CSP publishes Cloud Service Interfaces (CSI) so that CSCs can access IaaS, PaaS and SaaS services via those interfaces. Typically these interfaces are web services interfaces, such as REST-based [1] interfaces or API.

1.2. Cloud Service Consumer (CSC)

In this paper we consider following issues with respect to a CSC:

- C1. A CSC should be able to make use of the resources it acquires in the Cloud as if they are the CSC *intranet* resources. But this is challenging since the CSC does not have any control over the CSP Cloud infrastructure in the same way it has over its intranet infrastructure. A CSP can facilitate this via the SCL service described in this paper.
- C2. There are many types and instances of servers or applications (we use the term resources to refer to servers or applications) on CSC intranet (on-premises) that function as integrated whole or distributed systems. For example, web, application

and DB servers or Hadoop/Mapreduce [2] (H/MR) master server (name node) and slave nodes. A CSC may decide to move some of these resources or instances of a resource, such as H/MR slaves or certain application servers to Cloud, while keeping the others, such as H/MR master or DB servers on-premises. It is obvious that on-premises and off-premises resources should be able to communicate with each other seamlessly and securely in both directions.

- C3. For security and isolation a CSC may want to control the following:
- Which on-premises resources or which enterprise sites may communicate with which off-premises resources and vice versa.
 - E2E isolation: in a multitenant environment isolate a CSC traffic from other tenants (we use the term CSC and tenant interchangeably) at various segments of the network:
 - From the edges of CSC enterprise sites via the MAN/WAN to CSP Cloud DC edges.
 - Various segments within the CSP Cloud DCs.
 - A CSC may want to apply typical enterprise policies to off-premises resources. For example, preventing certain off-premises resources to communicate directly with the public Internet or preventing a set of off-premises resources communicating with the enterprise intranet (how this is done is beyond the scope of the SCL framework).

- C4. A CSC may want to manage Cloud resources in the same way it manages its intranet IT resources. As soon as the off-premises Cloud resources are acquired, they should be manageable via CSC (IT) management systems.

From the Cloud Service Providers (CSP) perspective following are the issues that have to be considered:

- A CSP, while it cannot provide full control to a CSC over its infrastructure, should provide certain (Cloud service) capabilities that will allow a CSC to incorporate off-premises resources into its intranet seamlessly and securely and do so on-demand.
- The CSP should provide proper connectivity and isolation transparently so that the on-premises and

off-premises resources can communicate with each other in both ways seamlessly and securely.

- A CSP has to support multiple tenants (CSCs or multiple groups or departments within a particular CSC) over a shared infrastructure. Hence the CSP has to provide multitenant isolation at various levels (compute, storage and network) and end/edge-to-end-edge (E2E) by providing proper capabilities as part of the SCL service.

The Seamless Cloud Abstraction and framework cover the above issues. In order to support S3 above, the SCL will also incorporate an abstraction called the Cloud Isolation Abstraction (CLIA). One of the major aspects of Cloud Computing is acquisition and/or release of Cloud resources on-demand. By considering the SCL as an abstract Cloud resource, a CSC (IaaS admin) should be able to Create, Read, Update and/or Delete it on-demand. In a separate paper we report how distributed applications (such as Hadoop/Mapreduce applications) and distributed monitoring system (such as Ganglia [3]) can run seamlessly and unmodified over an SCL utilizing both on-premises and off-premises Cloud resources.

In this paper we define the SCL abstraction model, and the CSC facing interfaces and relevant parameters that a CSC can use for on-demand CRUD of an SCL.

Note that, we do not define an (yet another) API framework. Rather the SCL abstraction and associated model and interfaces can be incorporated in existing API or Cloud Stack frameworks, such as OCCI [4], vCloud [5], CDMI [6], OpenStack [7], etc. A CSP (via a Cloud management system or Cloud Stack) has to realize the abstraction in the underlying Cloud infrastructure, which also is beyond the scope of this document.

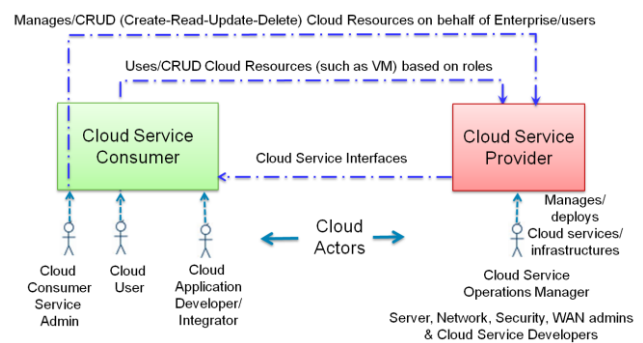


Figure 1. Cloud Actors

The rest of the paper is organized as follows: In section 2, we introduce the Cloud actors who deal with the Seamless Cloud abstraction and its realization. In section 3, we describe the SCL usage scenarios in an enterprise. In section 4, we explain the concept of the Cloud Isolation Abstraction. In section 5, we define the Seamless Cloud characteristics, abstraction model and interfaces. In section

6, Differentiated Quality of Seamless Cloud Services concept is described. Finally, we conclude in section 7.

2. CLOUD ACTORS

Before we proceed to define the SCL abstraction, model and interfaces, we provide an introduction to relevant Cloud actors, as shown in Figure 1.

Following are the actors (actor concept adopted from [8]) who will manipulate an SCL. The actor can also be software executed under relevant privileges.

- Cloud Service Consumer: IaaS Admin (CSC-A). An enterprise IaaS admin will have the authority to CRUD an SCL.
- Cloud Service Consumer: End User (CSC-U): For example, an Employee without admin privileges. A CSC-U will be able to associate or disassociate compute (VM) or storage resources to an SCL, but will not have privilege to create, update or delete an SCL.
- Cloud Service Provider: Ops Manager (CSP-O): CSP-O fulfills or realizes the SCL CRUD requests initiated by a CSC via a Cloud Management and automation framework (CSP Cloud Stack).

Note that, in a Cloud environment a request from a CSC should be realized in the infrastructure with minimum involvement from human actors. It is obvious that the CSP-O or Cloud Stack will have full and E2E visibility and control into the Cloud infrastructure to configure, provision and monitor resources, which the CSC-A will not have. The details of how a Cloud Stack realizes the abstraction in the infrastructure is beyond the scope of this paper.

3. SCL USAGE

Let us consider following Cloud usage scenarios to show where use of SCL makes more sense:

- Cloudburst: An enterprise acquires resources in a Cloud when excess capacity for certain applications is needed. The applications are executed in the Cloud and results moved back to the enterprise demilitarized zone (DMZ). In this case the off- premises resources may be isolated from the intranet and hence SCL use may not be needed.

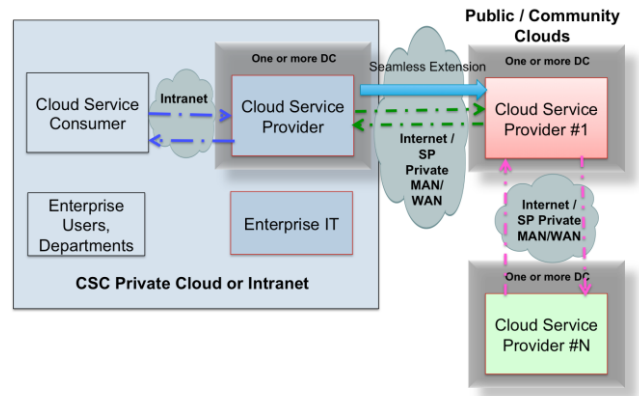


Figure 2. Cloud Deployment Models

- An enterprise acquires resources in a Cloud to be used by enterprise partners, customers or other users, but not by enterprise employees. In this case resources will be isolated from the intranet and SCL use may not be needed.
- Permanent Cloud use and Cloudburst: An enterprise acquires resources in a Cloud to be used by the enterprise itself and its employees on a continued or permanent basis. In this case off-premises resources should be elevated to be part of the enterprise intranet. For example, an application server running on the acquired resources in the Cloud should be able to communicate or write data back to a server (such as an on-premises DB server in the intranet). This is where a CSC benefits from using the SCL. The SCL abstraction together with the CLIA facilitates secure and seamless extension of an enterprise into Cloud(s) as shown in Figure 2.

4. CLOUD ISOLATION ABSTRACTION

One of the major features of Cloud is multitenancy, where multiple tenants share the same physical infrastructure. It is important that tenants are isolated from each other for security and privacy reasons. Typical focus of Cloud isolation has been limited, such as only at virtual machine (VM) and VLAN level. But for comprehensive multitenancy support and to address the issues mentioned in S2 and S3 the scope of isolation has to be extended to covering following:

- ISO/OSI layers 1 through 7.
- End/edge-to-end/edge (from CSC sites via the MAN/WAN to off-premises resources in Cloud DCs).
- Integrated compute, storage and network level isolation.

In this paper we focus only on E2E network level CLIA for multitenancy, which will allow a CSC to specify following features on-demand (we provide examples in later sections):

- The type of isolation, such as MPLS VPN [9], IPSEC VPN [10], GRE [11], VLAN, VSAN [12], etc. The type will be used to indicate only what kind of isolation a CSC desires (not, for example, a specific VLAN number or MPLS VPN VRF elements, which are expected to be assigned by the CSP Cloud Stack automatically). The CSP, via the SCL Cloud service interfaces, will offer options of isolation type to be used. A CSC will not be able to arbitrarily choose a type, since the underlying infrastructure may not support it. For example, a segment of a network may support MPLS VPN only, while the other supporting IPSEC.
- Segments of network where CLIA will be applied. The network itself will be abstracted and presented to a CSC, since a CSC cannot or should not have full visibility into the (physical) Cloud infrastructure. Since a network has multiple segments E2E, multiple different CLIA may be specified E2E.

When a CLIA is created by a CSC IaaS admin, the Cloud infrastructure management framework (or Cloud Stack), depending on the underlying infrastructure capabilities will map or realize the CLIA in the underlying infrastructure (via proper automated orchestration and provisioning), which will involve following:

1. Resource isolation in CSP Cloud DCs, such as VM and virtual storage isolation and firewall between resources.
2. E2E Traffic isolation as they cross various segments of network from on-premises enterprise resources or sites via the MAN/WAN to off-premises Cloud resources. It is obvious that in the absence of a single E2E isolation feature (something similar to a VLAN ID or E2E VRF), packets and frames have to be mapped to proper isolation features when they cross from one network segment to another. For example, from VLAN to VRF/IPSEC to VLAN, etc. The VXLAN [13] proposal could be a solution for this.
3. Routing and switching level isolation (via route and switch table isolation in network, which facilitates option 2 above).

The mechanism and process of how the abstractions are mapped or realized in the underlying infrastructure is outside the scope of this document.

As we have mentioned before in issue C1 a CSC IaaS admin cannot have full visibility and control over a CSP infrastructure. But the CSP can provide a CSC with certain level of control over choosing multitenant isolation capabilities. Hence we define the Cloud Isolation Abstraction that will be exposed to CSC. A CSC IaaS admin will then be able to associate (or disassociate) one or more CLIAs with (or from) an SCL and its elements.

5. SEAMLESS CLOUD ABSTRACTION

5.1. SCL Characteristics

The SCL has following major characteristics:

1. The domain of an SCL spans across LAN, MAN and WAN.
2. It consists of the following elements:
 - a. A set of CSC selected on-premises resources in the enterprise.
 - b. A set of CSC selected off-premises Cloud resources (in CSP Cloud DC).
 - c. A set of CSC selected enterprise sites.
 - d. A set of CSC selected CSP Cloud sites (CSP Cloud DC Sites).
3. On-premises and off- premises Cloud resources working seamlessly together (as in an intranet).
4. In a multitenant Cloud, individual tenant SCL is isolated via proper CLIA E2E and at proper network segments.
5. Any element in 2) can be associated or disassociated from an SCL on-demand and anytime.
6. A CLIA can be created or deleted separately from an SCL and associated to or disassociated from an SCL on-demand and anytime.
7. An enterprise may have multiple instances of SCL (departmental SCL, for example).
8. An SCL may be associated with multiple CLIA in different segments of a Cloud network, since different network segments may support different network isolation technologies (MPLS VPN, VPLS [14], IPSEC, GRE, VLAN, VSAN, LISP multitenancy [15], etc.).
9. Resources and CLIA in multiple CSP can be associated with (or disassociated from) an SCL.

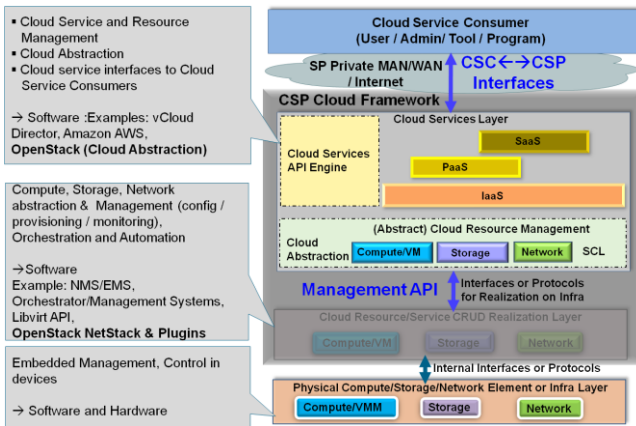


Figure 3, High Level Cloud Framework Architecture

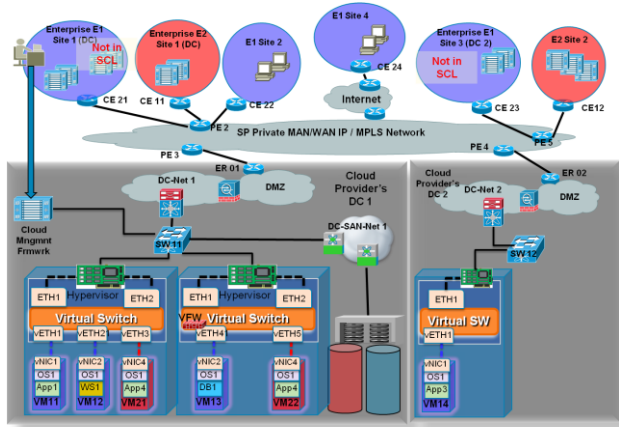


Figure 4. Example Cloud Network

5.2. SCL Cloud Abstraction Model

The SCL is a Cloud abstraction, where an SCL is a logical or abstract object (resource) that is realized in the underlying infrastructure. The SCL, as shown in Figure 3 resides in the Cloud services layer or abstract Cloud resource management layer.

We describe the model using an example of an E2E network, as shown in Figure 4 (a CSP view of network, not CSC user or IaaS admin view), which consists of the following:

- CSC enterprise sites (DC, remote sites, branches).
- A set of CSC intranet sites connected via an SP private MAN/WAN (with MPLS VPN, VPLS, etc.).
- A set of CSC intranet sites connected over the public Internet.

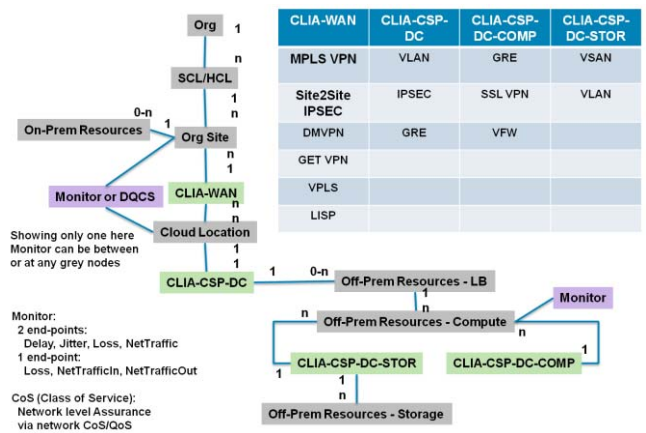


Figure 5. Seamless Cloud Abstraction Model

- CSP Cloud DC's connected to the SP private MAN/WAN and Internet.
- CSP Cloud DC network segments: Server, Server Access, DC LAN, DMZ, DC Edge.

5.3. SCL Interfaces

Below is the high-level description of the relevant (CRUD) interfaces of the model shown in Figure 5. We use the following notation:

<response> ← <abstract resource> : <parameters>.

- <SCL ID> ← SCL : <Tenant or Tenant Dept ID>
 - Possible Tenant ID can be a unique URI, such as Cisco.com.
 - Possible tenant dept ID: <URI/ID or name>, such as Cisco.com/ENGG.
- [<Site ID> ← Site : Location [, {On-premises Resources ID}]]. If not specified then the whole enterprise is considered and tenant or tenant dept ID is used.
 - The standard for Site ID should be defined. A possible ID is <CSC base url>/location.#, such as cisco.com/sanjose.1
 - List of on-premises resources ID, which is optional, and can be IP address or other ID, such as LISP EID or URI.
- <DC ID> ← DC : <location>
- <CLIA ID> ← CLIA-WAN : <SCL ID>, {<Site ID> list} | <Tenant ID> | <Tenant Dept ID>, {<DC ID> list} [, <MAN/WAN Isolation Technology (WIST)>].

This CLIA is used to isolate traffic (including routes) from tenant sites via the MAN/WAN to the Cloud DC edges. The WIST can be any of the

MAN/WAN isolation technologies as shown in Figure 5. If a WIST is not specified, CSP selects a supported technology. A CSP may provide multiple options of WIST a CSC may select from.

- $\langle \text{CLIA ID} \rangle \leftarrow \text{CLIA-CSP-DC} : \langle \text{SCL ID} \rangle, \langle \text{DC ID} \rangle, \{ \text{Off-premises Cloud Resource list} \} [, \langle \text{DC Isolation Technology (DIST)} \rangle]$.

This isolation is used to isolate tenant traffic from Cloud DC edge to the acquired Cloud resources. The DIST can be any of the DC/LAN isolation technologies as shown in Figure 5. If a DIST is not specified, CSP selects a supported technology. A CSP may provide multiple options of DIST a CSC may select from.

- $\langle \text{CLIA ID} \rangle \leftarrow \text{CLIA-CSP-DC-COMP} : \langle \text{SCL ID} \rangle, \langle \text{DC ID} \rangle, \{ \text{Off-premises Cloud Resource list} \} [, \langle \text{DC Isolation Technology (DIST)} \rangle]$.

This isolation is used to isolate inter-compute (server/VM) traffic. It includes also (virtual) firewall between compute resources, such as between web and DB servers. The DIST can be any of the DC/LAN isolation technologies as shown in Figure 5. If a DIST is not specified, CSP selects a supported technology. A CSP may provide multiple options of DIST a CSC may select from.

- $\langle \text{CLIA ID} \rangle \leftarrow \text{CLIA-CSP-DC-STOR} : \langle \text{SCL ID} \rangle, \langle \text{DC ID} \rangle, \{ \text{Off-Prem Cloud Resource (OPCR) list} \} [, \langle \text{Storage Isolation Technology (SIST)} \rangle]$.

This isolation is used to isolate traffic between compute and storage resources. The SIST can be any the storage network isolation technologies as shown in Figure 5. If a SIST is not specified, CSP selects a supported technology. A CSP may provide multiple options of SIST a CSC may select from.

5.4. E2E Network Monitoring

Monitoring can be inserted between and at any end-points (as shown in Figure 5). Following are possible monitoring parameters:

- Two end-points:
 - Delay
 - Jitter
 - Loss
 - NetTraffic. This is useful for monitoring traffic for a “flow” (such as TCP/UDP flow or between an IP prefix to a destination IP address).

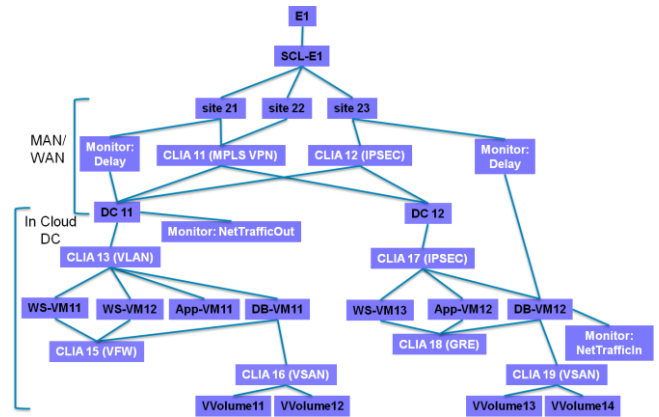


Figure 6. Seamless Cloud Logical Topology – CSC E1 View

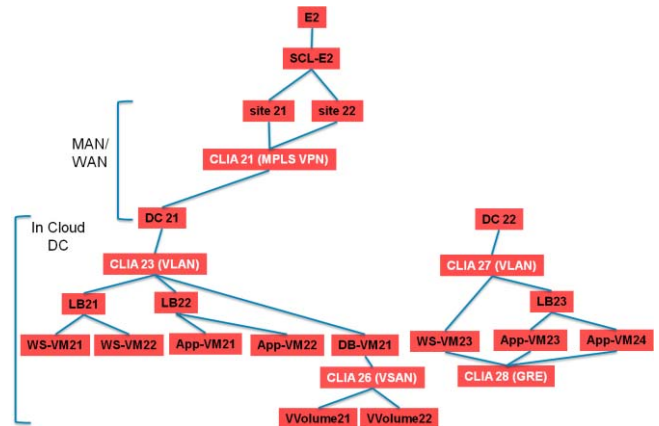


Figure 7. Seamless Cloud Logical Topology – CSC E2 View

- One end-point:
 - NetTrafficIn
 - NetTrafficOut.

5.4.1. Monitoring Interfaces

Similarly as SCL and CLIA, we define abstract monitoring “resources” (at the Cloud abstraction level as shown in Figure 6 and Figure 7), which will be mapped to various monitoring frameworks (such as NetFlow [16], SNMP [17], IPSLA [18], Ganglia, etc.) at the realization layer. The realization mechanism is beyond the scope of this paper. Based on the above we define two types of abstract monitoring “resources”:

- SCLMonitor1 with one end-point only with following interface:
 - $\langle \text{Monitor1 ID} \rangle \leftarrow \text{SCLMonitor1} : \langle \text{SCL ID} \rangle, \text{NETTRAFFICIN} | \text{NETTRAFFICOUT}, \langle \text{end-point 1} \rangle$. The end-point 1 is: $\langle \text{Site ID} \rangle | \langle \text{DC ID} \rangle | \text{Off-premises Cloud resource ID} | \text{On-premises resource ID}$.
- SCLMonitor2 with two end-points with following interface:

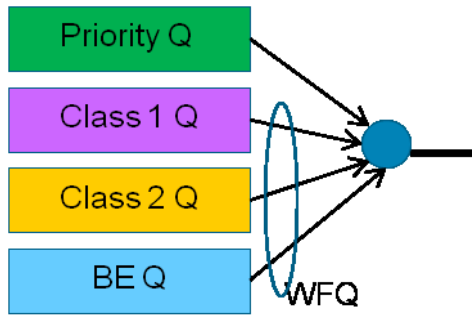


Figure 8. Figure Q

- $\langle \text{Monitor2 ID} \rangle \leftarrow \text{SCLMonitor2} : \langle \text{SCL ID} \rangle, \text{DELAY} | \text{JITTER} | \text{LOSS} | \text{NETTRAFFIC}, \langle \text{end-point 1} \rangle, \langle \text{end-point 2} \rangle$. The end-point 1 or 2 is: $\langle \text{Site ID} \rangle | \langle \text{DC ID} \rangle | \text{Off-premises Cloud resource ID} | \text{On-premises resource ID}$.

6. E2E DIFFERENTIATED QUALITY OF SEAMLESS CLOUD SERVICES

A SCL can be augmented with the capability of E2E network QoS, where traffic is differentiated to provide different level of priority treatments. For example, as shown in Figure 8 traffic in the priority Q is serviced first and the rest are serviced in weighted round robin fashion.

By augmenting E2E network QoS with SCL, Differentiated Quality of Seamless Cloud Services (DQSCS) can be offered by a CSP. The DQSCS can be defined at various levels of granularity as follows:

- Aggregated DQSCS, which is an abstraction of aggregated network QoS, such as Platinum, Gold, Silver services. This option of DQSCS can be applied to a whole SCL. A standard for Aggregate DQSCS is yet to be defined.
- DQSCS abstraction based on application or service classes, such as those defined in RFC 4594 [19]. Since this option is application/service specific, it will apply to only one resource. For example, when a CSC deploys an application (such as a media streaming application or a DB server) in an existing SCL, it will specify the service class, for example, per the RFC 4594.

6.1. DQSCS Interfaces

The interface for the DQSCS Cloud abstraction is as follows:

- $\langle \text{DQSCS ID} \rangle \leftarrow \text{DQSCS} : \langle \text{SCL ID} \rangle, \langle \text{Aggregate DQSCS} \rangle | \langle \text{RFC 4594 App Class} \rangle, \langle \text{end-point 1} \rangle$. The end-point 1 is an Off-premises Cloud resource ID.

The realization layer will map DQSCS to proper network QoS or CoS (such as those specified in RFC 4594 and other mechanisms, such as MPLS EXP [20] and 802.1p [21]) and apply network QoS configuration E2E where applicable and possible. For example, RFC 4594 stipulates that when service class is “low-latency data”, then packets (at the source) should be marked with DSCP AF21.

7. CONCLUSION

Cloud computing has emerged as a major area that is changing the computing, networking and IT landscape in a major way. It is also a nascent area open for many innovations and enhancements. A CSC making use of a Cloud needs services that facilitate seamless and secure integration of its enterprise with the Cloud. These services also have to have certain capabilities that make the Cloud enterprise and service provider (SP) class (as opposed to typical Cloud), such as sophisticated Cloud abstraction that simplifies Cloud usage. In this paper, we described such a Cloud abstraction, called the seamless Cloud that allows a CSC extends its intranet seamlessly, securely, on-demand and in a flexible manner. The SCL abstraction facilitates seamless execution of distributed application as if the SCL associated (on-premises and off-premises) resources are in the same intranet. We also described an abstraction called the Cloud Isolation Abstraction that isolates resources and network traffic of a CSC (tenant) at various segments of the Cloud network from other tenants in a multitenant Cloud environment. In addition, we defined the (API framework agnostic) CSC-CSP interfaces.

For future work, there are a number of possibilities that the concept can be extended to. The potential areas, that we are looking into, are SCL for inter-Cloud or Cloud federation and a Cloud *Language* supporting the SCL and extended concepts. The language, targeted for IaaS and PaaS developers (both admin and typical users), will provide further abstractions over Cloud services and relevant abstractions. Work is also in progress to design an Openstack based framework for SCL. The design includes specifying the SCL interfaces in Openstack API framework. In addition, we are also looking into standardizing the interfaces via a standards organization.

REFERENCES

- [1] Fielding, R. "Architectural Styles and the Design of Network-based Software Architectures." 2000. <http://www.ics.uci.edu/~fielding/pubs/dissertation/top.htm>.
- [2] "Hadoop MapReduce." <http://hadoop.apache.org/mapreduce>.
- [3] Culler, Mathew L. Massie and Brent N. Chun and David E. "The Ganglia Distributed Monitoring System: Design Implementation And Experience." *Parallel Computing*, 2003: 2004.
- [4] "Open Cloud Computing Interface Specification." <http://forge.ogf.org/sf/go/doc16162?nav=1>.
- [5] "vCloud API Specification v1.0." <http://communities.vmware.com/docs/DOC-12464>.
- [6] "SNIA Cloud Data Management Format Specification Version 1.0.0." [snia.org](http://www.snia.org/tech_activities/standards/curr_standards/cdmi). http://www.snia.org/tech_activities/standards/curr_standards/cdmi.
- [7] "Openstack". <http://www.openstack.org>.
- [8] "DSP-IS0103 - Use Cases and Interactions for Managing Clouds." [www.dmtf.org](http://www.dmtf.org/sites/default/files/standards/documents/DSP-IS0103_1.0.0.pdf). http://www.dmtf.org/sites/default/files/standards/documents/DSP-IS0103_1.0.0.pdf.
- [9] "Introduction to Cisco MPLS VPN Technology." http://www.cisco.com/en/US/docs/net_mgmt/vpn_solutions_center/1.1/user/guide/VPN_UG1.html.
- [10] S. Frankel, K. Kent, R. Lewkowski, A. Orebaugh, R. Ritchey, and S. Sharma. "Guide to IPsec VPNs: Guide to IPsec VPNs: Recommendations of the National Institute of Standards and Technology." *NIST Special Publication 800-77*, December 2005.
- [11] "Generic Routing Encapsulation (GRE)." RFC 2784. <http://tools.ietf.org/html/rfc2784>
- [12] "VSAN Configuration". http://www.cisco.com/en/US/docs/storage/san_switches/mds9000/sw/rel_2_x/fm/configuration/guide/vsan.html.
- [13] "VXLAN: A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks." <http://tools.ietf.org/html/draft-mahalingam-dutt-dcops-vxlan-00>.
- [14] "Virtual Private LAN Services (VPLS)." http://www.cisco.com/en/US/products/ps6648/products_ios_protocol_option_home.html
- [15] Farinacci, V. Fuller, D. Meyer, and D. Lewis. "Locator/ID Separation Protocol (LISP)." <http://ietf.org>. March 2, 2009. <http://tools.ietf.org/html/draft-farinacci-lisp-12>.
- [16] "NetFlow Version 9 Flow-Record Format." http://www.cisco.com/en/US/technologies/tk648/tk362/technologies_white_paper09186a00800a3db9_ps6601_Products_White_Paper.html.
- [17] J. Case, M. Fedor, M. Schoffstall, and J. Davin. *A Simple Network Management Protocol (SNMP)*. <http://www.ietf.org/rfc/rfc1157.txt>.
- [18] "Cisco IOS IP Service Level Agreements (SLAs)." http://www.cisco.com/en/US/products/ps6602/products_ios_protocol_group_home.html.
- [19] "Configuration Guidelines for DiffServ Service Classes." RFC 4594. <http://www.ietf.org/rfc/rfc4594.txt>
- [20] "MPLS EXP-bits Definition." <http://tools.ietf.org/html/draft-andersson-mpls-expbits-def-00>.
- [21] "IEEE P802.1p." http://en.wikipedia.org/wiki/IEEE_802.1p.

SESSION 7

SERVICE QUALITY FOR A FULLY NETWORKED SOCIETY

- S7.1 Regulation of Bearer / Service Flow Selection between Network Domains for Voice over Packet Switched Wireless Networks
- S7.2 Accessibility support for persons with disabilities by Total Conversation Service Mobility Management in Next Generation Networks
- S7.3 LabQoS: A platform for network test environments

REGULATION OF BEARER / SERVICE FLOW SELECTION BETWEEN NETWORK DOMAINS FOR VOICE OVER PACKET SWITCHED WIRELESS NETWORKS

Nikesh Nageshar, Rex Van Olst

School of Electrical and Information Engineering
University of the Witwatersrand, Private Bag 3, WITS, 2050
Johannesburg, South Africa

ABSTRACT

With the evolution of wireless systems from traditional circuit switch technology to packet based technology there is a requirement that voice be maintained to an acceptable level of quality such that user experience does not become compromised. All next generation wireless networks have been specified as packet switched radio networks which imply that the flaws of traditional packet based networks now also apply to voice over the wireless medium. This combined with the dynamics of a traditional air interface provides a further challenge to voice over a packet switched wireless network. The following paper proposes the facilitation of regulation that will predefine the handover of Quality of Service (QoS) metrics for voice from one predefined QoS network domain to subsequent network domains for wireless system handover. It is the intention of this paper to highlight the advantages of providing a voice QoS regulated admission control, bearer / service flow selection and mobile transport backhaul so as to ensure the successful transmission of quality voice packets.

Keywords— IP, LTE, QoS, VoIP, WiMAX

1. INTRODUCTION

In contrast to previous standards the current 3.9 and 4th generation wireless network standards have been specified as ‘all-IP networks’ thereby providing an end-to-end packet based connection for all services [1].

Existing circuit switched cellular networks consist of base stations (or base transceiver stations), base station controllers, switching centres and gateways. These base stations conduct fast power control and wireless scheduling and the base station controllers (BSC) execute the majority of the radio resource management. In contrast, the Fourth Generation (4G) network architecture has a simple structure where each BS functions in an integrated intelligent manner to perform radio resource management as well as physical transmission [2] [3].

The radio access solution was of a primary consideration in the development of a Fourth Generation (4G) network strategy as this played a central role in enhancing mobility, service control and the efficient use of network resources. As a result the network architecture called for a ‘flat’, all-IP core network, called the Evolved Packet Core (EPC). The Evolved Packet Core features a simplified architecture with open interfaces, higher throughput and lower latency [2].

In previous cellular telecommunications standards such as Code Division Multiple Access (CDMA) and Global System for Mobile Communications (GSM), voice was inherently the main service offered. This was done over a dedicated circuit switched channel, whereas in the 4G standards video, audio and interactive data services are largely considered [4].

The provisioning of satisfactory quality of service for voice over an Internet Protocol (IP) network is inherently difficult because of the tight delay, jitter and packet loss requirements of voice traffic. For satisfactory voice transmission a network should consist of sufficient bandwidth to carry the coded voice and relevant application, transmission and network protocol overheads. The network should have less than 0.25% packet loss, a maximum jitter of 5 millisecond and less than 150 millisecond packet delay [5]. These parameters have been determined by relating the network quality to objective and subjective voice quality metrics. It has been established that a greater than 0.25% packet loss, 5 millisecond jitter and /or 150 millisecond delay significantly contributed to speech stutter and speech delay [5]. Although efficient Packet Loss Concealment (PLC) algorithms have been created such that voice traffic can withstand a greater than 0.25% packet loss, for the purposes of this paper the above listed parameter is used. The packet based requirements combined with the scarcity of radio resources make the provisioning of voice with a fair to perfect Mean Opinion Score (MOS) a challenge [1].

The purpose of this paper is to present a pointer or label switch approach to QoS implementation for voice across multiple network domains, highlight the QoS frameworks available in 4G packet switched wireless networks, present a recommendation of acceptable service flows / bearers that can be used for voice and introduce a voice network factor that will regulate the mapping between the wireless, transport layers and other network domains.

This work was supported in part by the University of Witwatersrand School of Electrical and Information Engineering. The authors acknowledge the support of the University of the Witwatersrand and Neotel South Africa.

The areas within the wireless network architecture that have been investigated are as follows [6]:

- Voice application to admission control mapping.
- Admission control mapping to bearer / service flow selection.
- Bearer / service flow selection to transport network mapping.

The paper is set out as follows; Section 2 proposes and illustrates the resource management framework for voice; Section 3 highlights admission control management in Long Term Evolution (LTE) wireless networks. Section 4 is a brief examination of the quality of service framework available for 4G networks. Section 5 highlights the admission control to bearer / service flow resource mapping; Section 6 illustrates the bearer / service flow resource to transport resource mapping; Section 7 deals with the introduction of a voice network parameter factor and finally the conclusions are presented in Section 8.

2. RESOURCE MANAGEMENT FRAMEWORK FOR VOICE

Radio networks are required to use highly sophisticated explicit resource management techniques since Radio Frequency (RF) spectrum is a scarce commodity. In many instances these resource management techniques can be adapted to cater for multiple QoS scenarios represented in a framework [7].

In a generic QoS framework, a network is divided into different administrative domains with each network domain consisting of handshake entities (service negotiating entities) such as a negotiating manager and resource manager. The negotiating manager and resource manager are specific to each domain; where the former negotiates with its subscribers and the latter checks the availability of network resources [7].

The main characteristics that a resource negotiating protocol may consist of are listed as follows:

- The ability to negotiate based on the request from the clients equipment.
- Compatible with different QoS architectures across different service provider domains.
- Low signalling overhead.

In the development of a resource reservation and management framework for voice, the voice pointer approach is illustrated in Figure 1.

It is considered important to use simple QoS signalling between differing network elements when negotiating resources for voice, as portrayed in Figure 1 where 1, 2, 3 or 4 represents a pointer. Each pointer would identify to a given traffic handling capability which is relevant to the particular user plane network element. The specification of a traffic handling behaviour provides sufficient information

that allows the realisation of a particular QoS bearer or service flow via that network element.

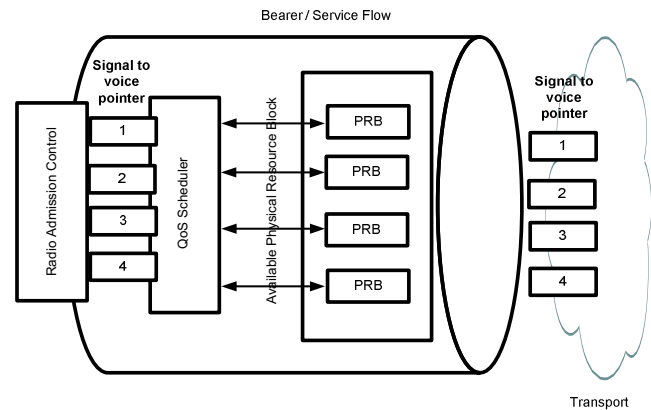


Figure 1. Pointer approach to voice QoS negotiation

The idea is that a set of traffic handling behaviours with pre-defined attributes could be configured inside the User Equipment (UE) as well as at the entry point of each of the individual network domains, such that only the appropriate pointers need to be signalled.

A typical network setup will consist of user equipment signalling the admission control pointer defined for a set type of traffic such as voice or multimedia. The admission control will in turn signal the network bearer or service flow and the network bearer or service flow triggers the transport network pointer. It is anticipated that this session will be held in state until the user equipment hangs up.

The above methodology deviates from the bandwidth / session negotiation model as it does not hold session information but rather trusts the source of the information and applies the relevant forwarding treatment. It is recommended that such a methodology be only applicable to voice traffic.

It is anticipated that bandwidth resources will not be an issue for voice traffic because voice traffic will always occupy the highest order priority QoS available on each of the network domains.

3. ADMISSION CONTROL MANAGEMENT

Radio admission control admits or rejects requests for new connections between the base station and mobile subscribers depending whether it will be able to fulfil the QoS criteria of a new connection request without compromising active sessions [8]. If LTE is considered as an example, in order to provide quality control on the Physical Resource Block (PRB) allocation, the admission control and packet scheduling need to be QoS aware. The QoS aware packet scheduling allocates the dynamically shared data channel to active radio bearers based on predefined QoS requirements [9].

With reference to the admission control algorithm, this decides to admit a new user if the sum of the Guaranteed Bit Rate (GBR) bearer of new and existing users is less than a predetermined value [9].

With reference to packet scheduling, the packet scheduling algorithm consists of priority scheduling given to packets which are farthest below its GBR requirement and the estimated achievable throughput on an available Physical Resource Block (PRB), hence the admission control and packet scheduling give rise to a proportional fair and GBR aware metric that is used for QoS [9].

It is proposed that the standard LTE admission control mechanisms continue as is, but rather a predefined voice pointer be added that can be signalled to allow voice specific bearers on a predefined, high priority QoS class indicator.

4. QUALITY OF SERVICE (QOS) FRAMEWORK IN FOURTH GENERATION NETWORKS

The QoS framework in the standards listed below has been investigated for the purposes of determining its effects on voice quality control:

4.1. IEEE 802.16e (WiMAX)

The QoS framework for the IEEE 802.16e standard is based on Service Flows (SFs). A service flow exists between the Access Service Network Gateway (ASN-GW) and user equipment and is marked with a connection ID illustrating QoS attributes such as packet latency/jitter and throughput [10].

The IEEE 802.16e supports five service flow types [10] [11]. Each of the above service flows has various QoS attributes associated; this is indicated in Table 1 below [11].

Table 1. QoS Parameters for IEEE802.16e

Service Flow Type	MRTR	MSTR	Max Latency	Max Jitter	Traffic Priority
UGS		X	X	X	
ertPS	X	X	X	X	X
rtPS	X	X	X		X
nrtPS	X	X			X
BE		X			X

Where:

- MRTR - Minimum Reserved Traffic Rate;
- MSTR - Maximum Sustained Traffic Rate;
- Max Latency - Maximum packet delay over the air interface;
- Max Jitter - Maximum packet variation delay;
- UGS - Unsolicited Grant Service;
- ertPS - enhanced real time Polling Service;
- rtPS - real time Polling Service;
- nrtPS - non-real time Polling Service, and;
- BE - Best Effort.

4.2. Long Term Evolution (LTE)

The LTE Evolved Packet System (EPS) is based on packet flows established between the Packet Data Network gateway (PDN-GW) and the user terminal. LTE uses separate Service Data Flows (SDFs) mapped to corresponding bearer with a common QoS treatment.

LTE offers two types of bearers [10] [11]:

- a. Guaranteed Bit Rate (GBR): Dedicated network resources related to a GBR value associated with the bearer are permanently allocated and,
- b. Non-Guaranteed Bit Rate (non-GBR): A service utilizing a non-GBR bearer may experience congestion-related packet loss.

A service data flows within a bearer is assigned a QoS Class Identifier (QCI). The QCI refers to a set of packet forwarding treatments (e.g., scheduling weights, admission thresholds, queue management thresholds, and a link layer protocol configuration) preconfigured for each network element [11]. The QCI characteristics are listed in Table 2 [11]. The mapping of a SDF to a dedicated bearer is classified by IP five-tuple based packet filter that is either provisioned in the Policy and Charging Rules Function (PCRF) or defined by the application layer signalling [11].

Table 2. LTE Quality of Service Class Identifier (QCI)

QCI	Resource Type	Priority	Packet Delay Budget	Packet Error Loss Rate
1	GBR	2	100ms	10 ⁻²
2	GBR	4	150ms	10 ⁻³
3	GBR	3	50ms	10 ⁻³
4	GBR	5	300ms	10 ⁻⁶
5	Non-GBR	1	100ms	10 ⁻⁶
6	Non-GBR	6	300ms	10 ⁻⁶
7	Non-GBR	7	100ms	10 ⁻³
8	Non-GBR	8	300ms	10 ⁻⁶
9	Non-GBR	9	300ms	10 ⁻⁶

Where:

- GBR - Guaranteed Bit Rate Bearer, and;
- Non-GBR - Non-Guaranteed Bit Rate Bearer.

In order to enforce QoS for voice the most appropriate characteristic listed above need to be sufficiently negotiated before a call is terminated.

5. ADMISSION CONTROL TO BEARER / SERVICE FLOW RESOURCE MAPPING

In order to sufficiently admit voice traffic in a packet switched network the first step is to ensure accurate identification of voice at the eNodeB or Access Service Network Gateway. The recognition of voice can be done using shallow packet inspection, deep packet inspection, voice codec sniffing or any other methodology.

Considering LTE as an example, the necessary voice Traffic Flow Template (TFT) shall be used to discriminate between different payloads using the IP header, such as IP address or Port numbers etc.

It is anticipated that a repository shall be built relating the IP addresses of the user equipment to the bearer /service flow for voice. At this location the voice pointer is added. This voice pointer shall represent the most relevant bearer or service flow applicable to voice.

As illustrated in Figure 2; when voice traffic is recognised at the eNodeB or Access Service Network Gateway of the network it is expected that the call admission control will signal the bearer / service flow pointer. On analysis of WiMAX and LTE it is recommended that GBR QCI 3 and Extended Real-Time Polling Service (ertPS) are the most preferable bearers / service flows to carry voice on the radio network. This is due to the stringent latency, jitter and error loss requirements of voice traffic.

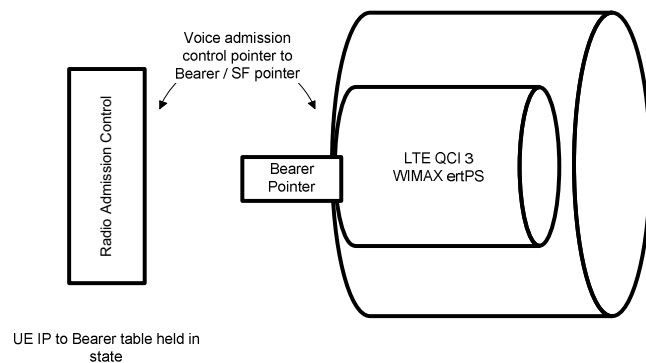


Figure 2. Radio access control to bearer / service flow pointer

6. BEARER RESOURCE TO TRANSPORT MAPPING

Many providers face the situation where the resource constraint does not occur on the radio network layer but rather on the transport or backhaul layer. Taking this into consideration there is a need to provide voice quality control over the transport network. It is anticipated that the transport network will be an all IP network. With the advent of Metro Ethernet systems deployed directly to the eNodeB, the transport limitation issue shifts focus from limited resources to appropriate queuing needed to be applied on the edge of the transport network. In order to provide sufficient voice quality control, route control and packet queuing parameters need to be coordinated with the radio network [6]. Route control refers to the QoS management function that is in charge of selecting the optimum routes in the transport network to guarantee the efficient use of transport resources. Packet queuing is in charge of implementing the appropriate QoS queuing decisions at the IP transport network nodes so that different flows receive the appropriate QoS treatment at every node. Route control also makes reference to link utilization and buffer control on the backhaul network [6].

Considering that QoS is required over the transport network, the Internet Engineering Task Force (IETF) has recommended IntServ and DiffServ for the transport of all-IP 4G networks. IntServ uses Resource Reservation Protocol (RSVP) to reserve bandwidth during a session setup. If the sender receives a resource reservation confirmation returned from the receiver as an indication of QoS guarantee, it proceeds with the initiation of the session.

IntServ ensures strict QoS, but each router in the transport network must implement RSVP and maintain a

per-flow state, which can cause difficulties in a large-scale network [3] [12]. DiffServ, on the other hand, does not require a signalling protocol and cooperation among nodes. As the QoS level of a packet is indicated by the DiffServ (DS) field of the IP header [3] [13].

The RFCs 2474 and 2475 define the fundamental framework of the DiffServ scheme [13] [14]. The DiffServ architectural framework is such that each packet's header is marked with one of the standardised code points. Each packet containing the same code point receives identical forwarding treatment by routers and switches in the path. This obviates the need for state or complex forwarding decisions in core routers based on a per flow bases [13] [14].

The ingress boundary router is normally required to classify traffic based on TCP/IP header fields. DiffServ micro flows are subjected to policing and marking at the ingress boundary router according to a contracted service level. Depending on the particular DiffServ model, out-of-profile packets are either dropped at the boundary or marked with a different priority level, such as best-effort [13] [14].

These functions are termed as traffic conditioning in DiffServ language. A DiffServ flow along with similar DiffServ traffic forms an aggregate. All subsequent forwarding and policing are performed on aggregates by the DiffServ interior nodes. As the interior nodes are not expected to perform an expensive classification function, their ability to process packets at high speeds becomes possible. The enforcement of the aggregate traffic contracts between DiffServ domains is key to providing QoS [13] [14].

However, the admission control modules must ensure that new reservations do not exceed the aggregate traffic capacity. These features make it possible to provide end-to-end services using DiffServ architecture [14]. Table 3 below illustrates the mapping rules that may be applied to 4G traffic flows:

Table 3. DiffServ Code Point (DSCP) to LTE QCI and WiMAX SF mapping

DSCP	QCI	Service Flow
EF	1, 2, 3	UGS, ertPS
AF4	5, 7	rtPS
AF3	4	rtPS
AF2	6	nrtPS
AF1	8	nrtPS
BE	9	BE

Where:

- EF - Expedited Forwarding;
- AF4 - Assured Forwarding Class 4;
- AF3 - Assured Forwarding Class 3;
- AF2 - Assured Forwarding Class 2;
- AF1 - Assured Forwarding Class 1, and;
- BE - Best Effort.

As illustrated in Figure 3 below; when the bearer / service flow pointer has been selected it is expected that the necessary transport pointer will be selected, i.e. the LTE bearer pointer will signal the transport bearer pointer. On the transport side it is expected that the transport bearer will take the form of Diffserv Expedited Forwarding (EF) or Multiprotocol Label Switching Experimental Bit 5 (MPLS EXP BIT 5).

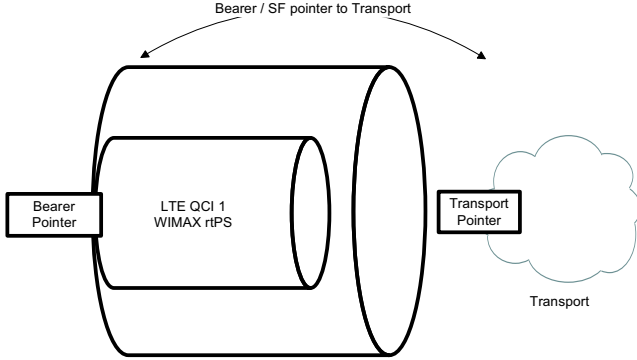


Figure 3. Bearer / service flow pointer to transport pointer

7. VOICE QOS CLASSIFICATION ACROSS DOMAINS

As stated earlier, in order to carry voice of a satisfactory quality a network should consist of sufficient bandwidth to carry the coded voice and relevant application, transmission and network protocol overheads. The network shall also consist of less than 0.25% packet loss, a maximum jitter of 5 millisecond and less than 150 millisecond packet delays [5].

If a voice call has to traverse across multiple network domains then the sum of each of the individual network parameters would be as follows:

$$p_{domain 1} + p_{domain 2} + p_{domain 3} \leq 0.25\% \quad (1)$$

$$j_{domain 1} + j_{domain 2} + j_{domain 3} \leq 5 \text{ ms} \quad (2)$$

$$d_{domain 1} + d_{domain 2} + d_{domain 3} \leq 150 \text{ ms} \quad (3)$$

Where:

- p – Maximum packet loss for an end to end call;
- j – Maximum packet jitter for an end to end call, and;
- d – Maximum packet delay for an end to end call.

Taking into consideration a 3 domain scenario as depicted in (1), (2) and (3), it can be indicated that any of the network domains can occupy the entire reserve network parameter. If this occurrence takes place then the summation of the network parameters of all the network domains will result in an overall degradation of voice quality in the end to end voice system. Hence an equipment vendor can indicate that their individual equipment is fully compliant to carry voice traffic, however when paired with other network domains the final result may be contradictory to that which is stated.

It is proposed that a voice network parameter factor (f) is introduced where each of the QoS flows within a

networking domain are classified by this factor (f). Where, f is classified as the network parameter for an individual network domain in relation to the overall required system parameter, as listed in (4), (5) and (6) below.

$$f_p = \sum_{i=1}^n \frac{p_n}{0.25\%} \quad (4)$$

$$f_j = \sum_{i=1}^n \frac{j_n}{5ms} \quad (5)$$

$$f_d = \sum_{i=1}^n \frac{d_n}{150ms} \quad (6)$$

Where:

- f_p – sum of the maximum packet loss for a network domain in relation to the overall maximum packet loss;
- f_j – sum of the maximum packet jitter for a network domain in relation to the overall maximum packet loss;
- f_d – sum of the maximum packet delay for a network domain in relation to the overall maximum packet loss, and;
- n – number of network domains.

Based on the equations listed above it can be stated that in order for quality voice to be maintained across a networking system, each of the voice network parameters f shall be ≤ 1 .

As an example, considering LTE QCI 1, 2, 3 and 4 from Table 2, the following are the voice network parameter factors (f) for the associated service flows.

Table 4. LTE Quality of Service Class Identifier (QCI) in relation to the Voice Network Parameter Factor (f)

QCI	Packet Delay Budget	Packet Error Loss Rate	f_d	f_p
1	100ms	10^{-2}	0.67	4
2	150ms	10^{-3}	1	0.4
3	50ms	10^{-3}	0.33	0.4
4	300ms	10^{-6}	2	0.0004

As illustrated in Table 4, it can be seen that in respect of f_d , QCI 1 and 3 can be used for voice as they perform below the threshold factor of 1. QCI 3 however outperforms QCI 1 by a factor of 0.33 hence is more robust in terms of being paired with other network domains in a voice traffic system.

In respect of f_p it can be seen that QCI 2 and 3 perform well below the threshold factor of 1 and both have sufficient room to be paired with other network domains.

8. CONCLUSION

It has been indicated that in order to carry good quality voice, physical resources need to be made available at a higher priority than other traffic classes [15]. Voice is an integral part of any operator's business; hence the ability to successfully provide voice over next generation networks becomes vital. Research institutions and standards bodies have taken the liberty to provide efficient data networks

consisting of high data throughputs with the central task of converged services.

The various QoS standards and mechanisms on 4G networks and transport networks have been briefly examined and an associated mapping between the radio network and the transport network illustrated. The latency, jitter and packet loss characteristics of the ePS service flow on WiMAX and the Guaranteed Bit Rate (GBR) bearer with QCI 3 on LTE are compliant with the requirements for voice traffic. The same is also true on the transport network for DiffServ Expedited Forwarding (EF) or MPLS EXP BIT 5. Signalling pointers are proposed so as to represent the relevant QoS attribute depicted for voice in each of the network domains.

This paper investigates the possibility of the standardisation of voice specific QoS structures across network domains, the mapping thereof and proposes a factor (f) that equipment vendors can report on so as to provide operators and system integrators an indication of the preferred QoS service flows that can be used for voice.

Each of the standards bodies such as the ITU, 3GPP etc. have qualified and specified relevant QoS frameworks; however the mapping of voice from one network domain to another is not specified as this requires co-ordination between the standard bodies. The idea behind indicating the voice network parameter factors (f) is so that integration between network domains for quality voice can be quantified and realised, be it via IMS, signalling pointers or some other methodology.

It is recommended that the voice network parameter factors (f) be introduced to assist in the maintenance of quality voice over packet switched wireless networks and across its associated network domains.

REFERENCES

- [1] 3GPP TS 36.300 V9.3.0. "Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Stage 2", March 2010.
- [2] 3GPP TS 36.201 V9.1.0. "Evolved Universal Terrestrial Radio Access (E-UTRA); LTE physical layer; General description (Release 9)", March 2010.
- [3] Choi Y, Lee K, Bahk S. "All-IP 4G Network architecture for efficient mobility and resource management," *IEEE Wireless Communications Journal*, May 2007, pp 42-46
- [4] Rein S, Fitzek F, Reisslein M. "Voice Quality Evaluation in Wireless Packet Communication Systems: A Tutorial and Performance Results for ROHC," *IEEE Wireless Communications*, February 2005, pp. 60-67.
- [5] Schutte J, Helberg A. "A Study of the Effect of MPLS on Quality of Service in Wireless LANS," *South African Institute of Electrical Engineers Journal*, Vol 99, September 2008, pp. 70-76.
- [6] Olmos J, Ferrus R, Sallent O, Perez-Romero J, Casadevall F. "A Functional End-to-End QoS Architecture Enabling Radio and IP Transport Coordination," *WCNC proceedings*, March 2007, pp 4348 – 4353
- [7] AROMA IST-4-027567. "Final Report on AROMA Algorithms and Simulation results," December 2007
- [8] Qian M, Huang Y, Shi J, Yuan Y, Tian L, Dutkiewicz E. "A Novel Radio Admission Control Scheme for Multiclass Services in LTE Systems," *IEEE Global Telecommunications Conference*, December 2009, pp 1-6.
- [9] Anas M, Rosa C, Calabrese F, Pedersen K, Mogensen P. "Combined Admission Control and Scheduling for QoS Differentiation in LTE Uplink," *IEEE 68th Vehicular Technology Conference*, September 2008, pp 1-5
- [10] IEEE 802.16 – 2009. "Part 16: Air Interface for Broadband Wireless Access Systems," May 2009
- [11] Alasti M, Neekzad B, Hui J, Vannithamby R. "Quality of Service in WiMAX and LTE Networks," *IEEE Communications Magazine*, May 2010, pp. 104-111
- [12] IETF RFC 1633. "Integrated Services in the Internet Architecture: an Overview," June 1994
- [13] IETF RFC 2475. "An Architecture for Differentiated Services," December 1998
- [14] IETF RFC 2474. "Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers," December 1998
- [15] Jiang D, Wang H, Malkamaki E, Tuomaala E. "Principle and Performance of Semi-persistent Scheduling for VoIP in LTE System," *WICOM proceedings*, September 2007, pp 2861 – 2864.

ACCESSIBILITY SUPPORT FOR PERSONS WITH DISABILITIES BY TOTAL CONVERSATION SERVICE MOBILITY MANAGEMENT IN NEXT GENERATION NETWORKS

Leo Lehmann

OFCOM, ITU-T Study Group 13 (Future networks including mobile and NGN)
Switzerland

ABSTRACT

This paper describes the principles and concepts necessary to support total conversation service mobility within a fixed/mobile converged telecommunication network. Regarding the network platform, this paper considers the functional architecture of the Next Generation Network (NGN), as the International Telecommunication Union (ITU) standardizes it. The presented procedure shall enable persons, who are disabled by deafness, speech disabilities and/or vision disabilities, to use the advantages of fixed/mobile converged telecommunication networks not only in a stationary situation but also when they are mobile. A strong focus is given on the support of context based service performance adaptation. Different handover scenarios are considered, including devices and network access of different capabilities.

Keywords— Accessibility, NGN, profiles, QoS, service mobility, total conversation

1. INTRODUCTION

“Achieving equitable communication for everyone” is one of the main strategic goals mentioned by the International Telecommunication Union (ITU) on its website. “ITU believes that these people should enjoy the same services and opportunities in life as everyone else”. In 2008 the first resolution on accessibility, Resolution 70 [1], was approved by the highest level meeting of the Standardization Sector, the World Telecommunication Standardization Assembly (WTSA-08). In October 2010, the highest level meeting of the International Telecommunication Union, the Plenipotentiary Conference approved Resolution 175, the first text of ITU in the field of accessibility [2]. Hereby accessibility describes the usability of a product, service, environment or facility by people with the widest range of capabilities.

The implementation of the total conversation service in telecommunication networks, as defined by ITU-T recommendation F.703 [3] fully supports the requirements mentioned above and enables equal possibilities in communication for all. Especially considering persons who have disabilities in hearing, speech or vision, the total conversation service has become an appropriate means to enable them to use telephony services in a more flexible

and equitable manner. According to ITU-T F.703 [3] the term “total conversation service” is defined as “an audiovisual conversation service providing bidirectional symmetric real-time transfer of motion video, text and voice between users in two or more locations”. So far total conversation has been mainly applied in telecommunication networks, which were based on a fixed network access. Meanwhile, these services also have become available for mobile networks. Furthermore the standardization of Fixed/Mobile Convergence (FMC) has proceeded significantly (regarding the definition of FMC see for instance [4]): The development of technical recommendations related to the Next Generation Network (NGN) by ITU (International Telecommunication Union) [5] as well as by ETSI/ 3GPP [6] enables users to make use of subscribed services with different terminals and with a variety of network connections.

Depending on the current context such as used network access or used device, service performance relevant parameters such as effective bandwidth, network latency or delay variation (jitter), may differ significantly. From the perspective of a user, who is using the total conversation service, it is expected that ongoing sessions and media streams may dynamically adapt to the current context (e.g. change of presentation quality of media streams and/or termination of not mandatorily required media streams) but if possible the service in all should not be aborted if the type of network access or the type of device has changed (context based service adaptation).

Thus, the provision of service mobility becomes an important design issue with regard to the deployment of total conversation services in a fixed/mobile environment like NGN. Service mobility defines the ability of a user to access the particular, subscribed, multimedia services during an ongoing session irrespective of the location of the user and the terminal that is used for that purpose [7]. Specific studies on service mobility that consider context based dynamic service adaptation can be found for instance in [8], [9], [10], [11]. The general mobility management framework and architecture for NGN is described in Y.2804 [12]. ITU-T recommendation Y.2111 [13] specifies the resource and admission control functions (RACF), which are required for the QoS support in next generation networks. Furthermore, Y.2018 [14] specifies the architecture of mobility management and control functions (MMCFs) for the NGN transport stratum. However Y.2018

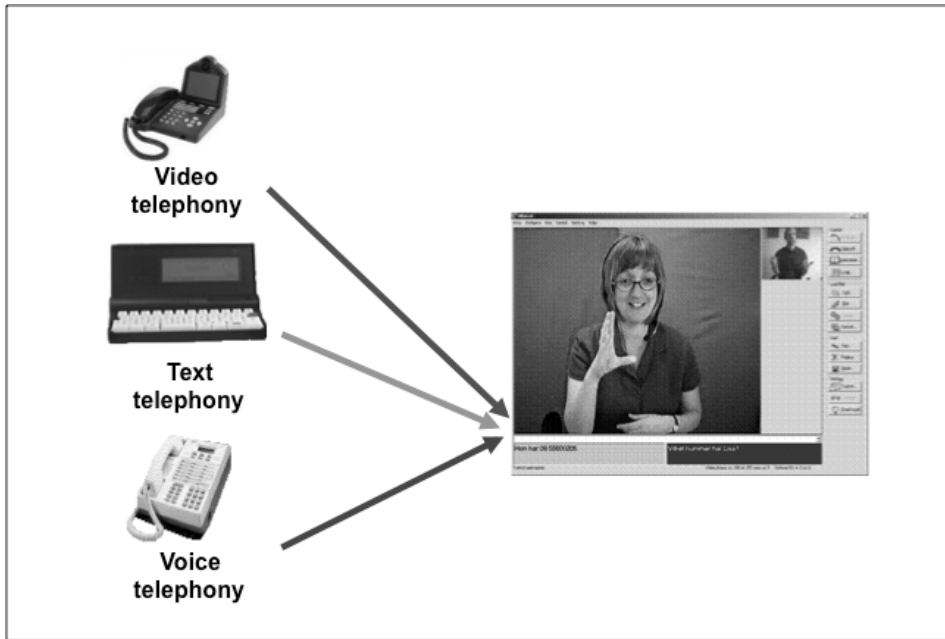


Figure 1. provision of motion video, real time text and voice by total conversation service

mentions that it “does not provide any mechanism to deal with service adaptation if the post-handover quality of service is different from the quality of service before handover”.

Considering the total conversation service this may have the following consequences:

- Inevitable abort of the running service in case of QoS degradation even if the given disability (e.g. deaf, hearing impaired, speech disabled) would allow service continuation with performance degradation (e.g. conversation continuation with text stream only for deaf persons)
- Failing to take advantage of opportunities to enhance post handover service performance in the case of an appropriate QoS increase.

Thus the lack of service adaptation may limit the communication abilities of persons with disabilities to a significant extend.

To alleviate these disadvantages, this paper describes the main features of a service mobility management procedure for NGN that includes particular means for the context based dynamic service adaptation if the post-handover quality of service is different from the quality of service before handover. Furthermore it is explained how this procedure can be applied to total conversation services. Even if inspired by the NGN architecture as defined by ITU-T, the given concept is not restricted to this specific architecture. It also covers other IP based fixed/mobile converged telecommunication architectures, which support the determination of performance relevant parameters like effective bandwidth, network latency or jitter.

The remainder of this paper is organized as follows: Chapter 2 shortly introduces the total conversation service and the associated user profile requirements. In Chapter 3 an overview of the NGN architecture and relevant functional elements for the mobility support is given.

Chapter 4 explains the dynamic determination of the possible variants of the total communication service. Chapter 5 describes the main features of the service mobility management procedure and its application for total conversation services. Chapter 6 finally summarizes the achieved results and notes future directions of work.

2. TOTAL CONVERSATION AND USER REQUIREMENTS

As mentioned in the previous chapter the term total conversation service describes [3] an audiovisual conversation service providing bidirectional symmetric real-time transfer of motion video, text and speech between users in two or more locations (see Figure 1).

Total conversation is standardized by ITU-T but it is also included in standards developed by the Third Generation Partnership Project 3GPP [15] and by the internet standard organization IETF [16]. Total conversation allows, for example, the communication between deaf and hearing persons by text-enhanced video for improved communication. Another application of total conversation could be a call between two deaf persons, using video for sign language and text for exchange of additional information like telephone numbers and addresses. If the used device supports Braille display or if such displays can be connected to the device, blind persons can be supported by additional textual information beside the voice stream in a conversation. Braille displays of different size with different number of character positions in a line are available (often between 12 and 40). In order to ensure a good experience when using total conversation services in telecommunication networks appropriate quality levels concerning text, video and audio have to be ensured. For example video quality for good sign language and lip reading use requires more than 20 pictures per second in

	Voice stream	Video stream	Text stream
Blind	x		x (Braille)
Deafened	x (one way talking)	x (lip reading)	x
Deaf signing		x (sign)	x
Hard-of-hearing	x	x (lip reading)	x
Deafblind speaking	x (one way talking)		x (Braille)
Deafblind signing		x (sign)	x (Braille)
Blind and speech disabled	x (one way listening)		x (Braille)
No communication disability	x	x	x

Figure 2. simplified view of total conversation variants depending on disability

order to support rapid motion in finger spelling, a spatial resolution of 352x288 pixels (CIF) for an appropriate sharpness and a latency below 400msec from the camera to the receivers screen [17] to handle conversational turn-taking well. Further requirements regarding audio and text streams for standard total conversation services can be found in [3]. Total conversation can be used for communication even in situations of service degradation. The kind of degradation that can be acceptable varies widely with the capabilities and preferences of the persons in the call, and with the nature and purpose of the call. Persons without any communication related disability who prefer voice communication can use total conversation services even in the event of a media stream loss (e.g. loss of video stream still allows conversation by voice and text). In case of disabled persons the suitable service variants strongly depend from the type of disability. A simplified view of combination of disability types (deafened, blind, speech disabled and combinations of them) and media usage is shown in figure 2. Deafened persons for example may want to use their voice to communicate but they require an alternative media stream for perception (text and/or video). Whether text or video is considered mandatory depends on personal preferences. Similarly, a person with speech disability can listen to a communication partner but requires text and/or video as an alternative medium to voice for communication. The choice (text, video or both) depends on the personal preferences. Beside the support of text streams, blind people additionally require access to devices that either support Braille, or which allow connecting of Braille displays in order to communicate by visual information. These requirements and related preferences have to be considered appropriately in defining the user profile described in chapter 4.

3. FUNCTIONAL ARCHITECTURE

According to ITU-T recommendation Y.2001 [18] Next Generation Networks are defined as packet-based networks able to provide telecommunication services and able to make use of multiple broadband, QoS-enabled transport technologies and in which service-related functions (service stratum) are independent from underlying transport-related technologies (transport stratum). Following [12], figure 3 gives a simplified overview of the architectural design.

The NGN of an operator consists of an IP core network and several IP-based access networks that may support fixed-wire, fixed-wireless and mobile access (AP). The scope of access ranges from line-based access to wireless and includes relevant technologies like xDSL, EDGE, UMTS/-HSPA, WLAN, WiMAX and LTE.

Functions concerning the network access control (NACF) are located in the access network. The core network interconnects several access networks to each other and interfaces to other networks. It further supports media processing functions as necessary. Regarding the transport stratum network resource and admission control functions (RACF) as well as functions concerning the mobility management control are contained in both the access and the core networks. Considering the service stratum, service convergence is achieved by the service control functions (SCF) (e.g. the IP Multimedia Subsystem, specified by 3GPP [19]). Specifically, service control functions in the core network provide session setup and control as well as connectivity between different user equipment (UE). They also interface as portal functions to service platforms (AS) that are either external (3rd party provider) or internally operated by the NGN network provider.

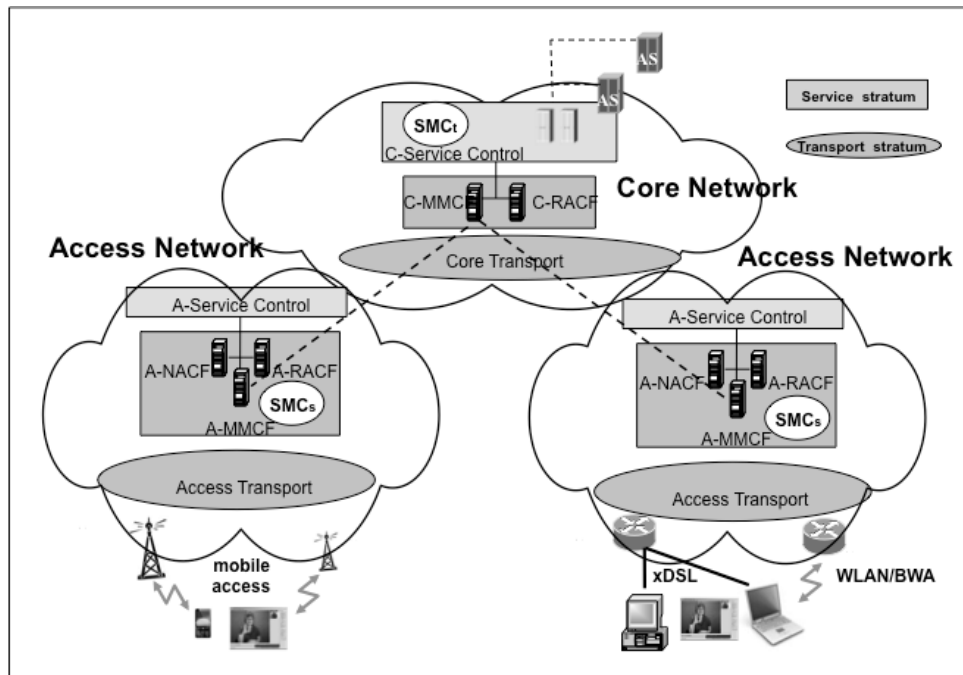


Figure 3. functional architecture

An important task of SCF in the access network is to act as a proxy and to forward messages to and from the user device (UE). Access and core networks are managed as self-contained IP domains. Devices and terminals may change the network point of attachment (terminal mobility [7]) by using the control and support of the mobility management control functions (MMCF), as specified in Y.2018 [14] (in co-operation with the corresponding NACF and RACF). Furthermore, users are enabled by the MMCF to maintain their user identity (e.g. SIP identifiers), irrespective of the terminal used and its network point of attachment (user mobility [7]).

4. DETERMINATION OF POSSIBLE SERVICE VARIANTS

Nevertheless it is evident that specific type of services like total conversation cannot be entirely supported by each transport technology. Mobile networks with UMTS technology can support, for example, video streams but video streams cannot run on pure GSM networks due to bandwidth restrictions of the GSM technology. For the provision of an appropriate service mobility support the transport and service stratum have to provide means to ensure that specific services like total conversation can be started and adapted during execution according:

- To the given capabilities of used devices (device profile),
- To the used network access (network profile),
- To the available service variants (content profile or service profile) specified by the AS and
- To the given user preferences (user profile) from a user profile database (e.g. the Home Subscriber Server as defined by 3GPP) regarding mandatory and optional

media components, preferred quality degrees (e.g. video resolution) and accepted service adaptations.

In order to provide such means a specific functional entity, the Service Mobility Controller (SMC) is foreseen.

The SMC consists of two parts the SMC_t for the transport stratum and the SMC_s for the service stratum.

Both SMC_t and SMC_s can be realized separately or can be integrated in other functional entities (e.g. SMC_t may be a part of the MMCF). While SMC_s resides only in the core network of NGN a SMC_t instance is located in each access network. SMC_s and SMC_t together determine the possible service performance (PSP) of the total conversation service application. First they consider the service variants described in the user profile as well as the content profile. Then they correlate (with the support of the related RACF instances) this information with the capabilities of the used device as well as the considered network access.

Further descriptions of the SMC functionality will be given in the context of the service mobility procedure description in chapter 5.

An illustration of the determination of the current PSP for a total conversation service with a specific service profile is given in figure 4. The access profile, the device profile, the service profile and the user profile describe several audio- and video codec's including also the text codec T.140.

Even if the service profile of the given total conversation service, as well as the used device, supports a variety of video codec's (including the most recent ITU-T video codec H.264) the current access profile only allows H.263-based video transmission.

Furthermore, the user profile of a hearing impaired person shows that text transmission is considered to be mandatory while voice and video streaming remain optional.

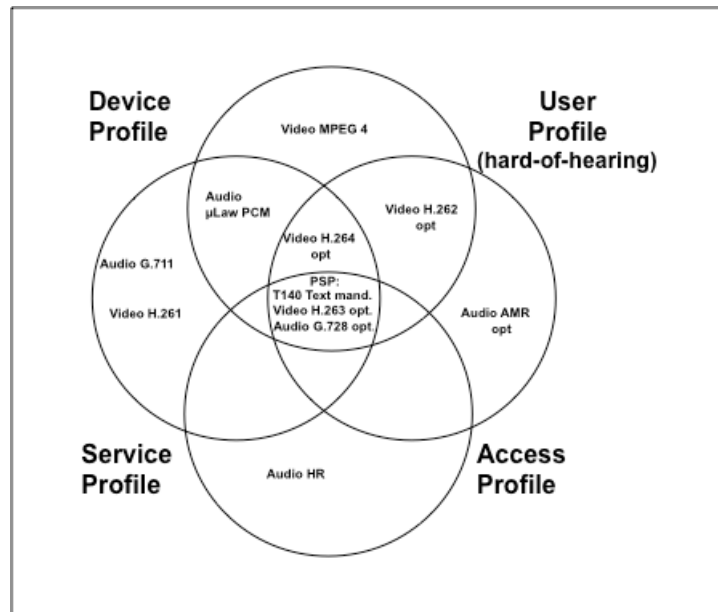


Figure 4. service determination

5. SERVICE MOBILITY PROCEDURE

This chapter gives a description of the main features of the procedure performed by the service and transport stratum of NGN using the SMC_s and SMC_t functional elements. The resulting interworking can be seen in figure 5, which illustrates the most relevant message exchanges between the involved functional elements. Furthermore, for simplicity it is assumed that service registration and service setup have already completed and a total conversation service, using text, voice and video stream has been established. In order to reduce complexity an explanation of conditions for an abnormal termination of the algorithm (e.g. denial from user side for reduced service performance) is omitted.

Details of the procedure

1. A handover is initiated from the current user device UE_0 by sending a handover request message via the current network access point AP_0 to the related mobility management control function A-MMCF₀. The handover request message contains the identities of one or more candidate access points AP (AP_d) and/or the candidate device UE_d to which the session shall be switched.
2. By the support of the C-MMCF (message “destination A-MMCF determination”) of the core network it is determined if AP_d belongs to the same domain as AP_0 . In case of a device change it is also determined whether UE_d is online and if it is connected to AP_d .
3. a) If the candidate AP is in the same domain A-MMCF₀ triggers $SMC_{t,0}$ to determine together with A-RACF₀ the available resources (access profile), associated with the candidate AP as well as to request the device profile from UE_d (respectively UE_0 if there is no device change). After merging both profiles $SMC_{t,0}$ returns the result to A-MMCF₀.

3. b) If the candidate AP belongs to another domain, A-MMCF₀ sends a resource request message to the related A-MMCF_d, requesting A-MMCF_d to determine, together with $SMC_{t,d}$ and A-RACF_d the available resources as mentioned in clause a).

4. The A-MMCF₀ sends a “PSP determination request”-message to the C-MMCF, which includes the merged access and device profile prepared by the related SMC_t as well as the address of the related A-MMCF.

5. Triggered by the C-MMCF the SMC_s , which is located in the service stratum of the core network, determines the new PSP. This is done by the correlation of the input of A-MMCF₀ (the result of the merging of access profile and device profile) with the service profile information located at the application server and the user profile information, located at the application server and/or at the user database of the service stratum.

6. If the new Possible Service Performance allows keeping the current media streams unchanged the procedure will be continued with step 8.

7. If the PSP does not allow the continuation of the current media streams because of decreased service performance (e.g. reduced bandwidth, increased latency) the SMC_s performs together with the SCF the appropriate end-to-end adaptation of the current media streams according to the new QoS performance conditions. Considering the total conversation service application, this includes:

7.a) If the PSP does not guarantee an available bandwidth which allows the continuation of a color video stream it is first checked to determine whether a monochrome video transmission could be continued, otherwise the video stream is terminated. In the case that the user profile mandatorily requires a video stream the current communication session will be terminated too.

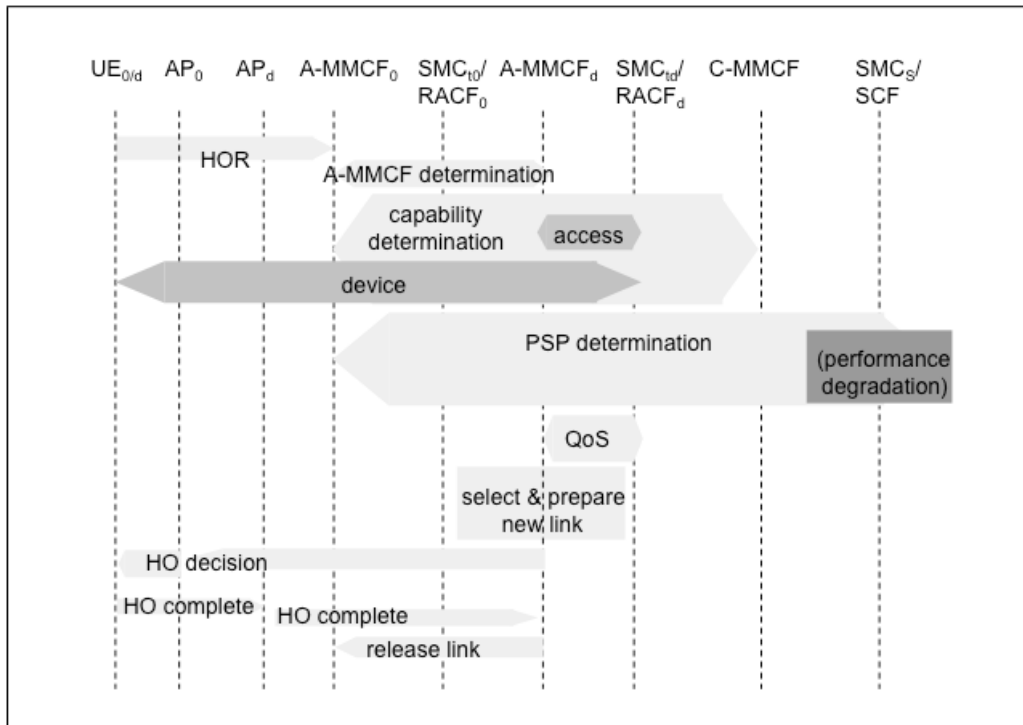


Figure 5. interworking between functional elements

7.b) If the PSP does not guarantee minimum requirements concerning temporal and spatial video resolution as specified by [17], the video stream will be terminated.

In the case of the user profile is mandating a video stream, the current communication session will be terminated too. Further termination conditions of the video stream may be an end-to-end delay bigger than 400ms or an increased packet loss rate related to the new AP.

7.c) If the user profile of a deaf person mandates text streaming and the new network access or the new device to which the ongoing session is handed over does not support text, the service must be terminated.

7.d) Even if voice services are supported by the majority of mobile devices the availability of voice is not a mandatory precondition for executing total conversation. For example if the user profile of a deaf person specifies “text only” total conversation services could be continued by text streaming even if the device does not support voice (e.g. first generation of iPad).

8. The SMC_s returns the calculated PSP to C-MMCF by a PSP determination response message.

9. The C-MMCF sends the PSP determination response via the requesting A-MMCF₀ to A-MMCF_d

10. Triggered by the reception of the PSP determination response message, A-MMCF_d initiates the related A-RACF_d to perform the appropriate resource reservation and prepares the new transport link.

11. A-MMCF_d sends (in case of different A-MMCF instances via A-MMCF₀) the handover decision message (including the identity of the new network access AP_d) via AP₀ to UE.

12. The UE starts handover and tries to connect with AP_d. After successful connection the UE returns via AP_d a handover complete message to A-MMCF_d.

13. If A-MMCF_d is identical to the original requesting AMMCF₀, A-MMCF_d terminates the old transport link associated with AP₀. If A-MMCF_d and A-MMCF₀ are different, A-MMCF_d forwards the received handover complete message to A-MMCF₀, requesting A-MMCF₀ to terminate the old transport link associated with AP₀. In the case where the handover involves also different devices, UE₀ and UE_d, the process starts in the same way as described above. In contrast to the scenario of one device, the handover complete is triggered by a user intervention on UE_d site to indicate the ability to handle the media streams by UE_d.

14. If the determined PSP allows (in accordance with the stored User and Service Profile) an enhanced service performance, the SMC_s will start an appropriate service enhancement after the handover has completed (e.g. switch on further media streams or adaptation of the current streams according to the improved QoS performance conditions).

Running the algorithm in an appropriate software simulation has provided early experience of the described procedure with appropriate profile adaptation. It turned out that switching off one media stream could be done with minor effect on the still ongoing media streams. But considering scenarios of possible service enhancements by an additional media stream (e.g. re-activating of the video stream) a reset of the other, ongoing media streams is currently inevitable in order to achieve the appropriate synchronization between all streams.

Once in a while, this causes small (visible and/or audible) disturbances of the ongoing service. The current revision of the procedure still leave possibilities in order to speed up the process required for the complete handover. By “in advance” determination of the access capabilities of candidate AP’s by the network or by the UE (which indeed would increase the general processing load of the considered system) the system would become more robust in the case of fast changing access conditions. Too extensive latencies that cause service breakdown and re-establishment reduction would be avoided by the proposed speedup of the procedure.

6. SUMMARY

This paper has described the concept of total conversation and its benefit when applied in telecommunication networks especially for disabled persons. Depending on the type of disability (in hearing, speech or vision), it explained the basic requirements that have to be considered in case of the implementation of total conversation service mobility in fixed/mobile converged networks like NGN. It explained how this service could be generally adopted according to the given context like effective bandwidth, network latency and network delay by applying the presented procedure. The framework of this procedure implies also the possibility to suspend the service on one device (e.g. mobile terminal) and to pick it up on another one (e.g. fixed terminal) by taking into account the different device capabilities. Even if the given paper is inspired by the NGN architecture as defined by the ITU, the proposed service mobility procedure is not restricted to NGN. It can be also applied to any IP- related network architecture that includes convergence of fixed, wireless and mobile access technology. Future work can be done by the reduction of further latencies as indicated above. The described procedure has focused on the service continuation aspects in case of changing QoS requirements in the access network. Even if access and device capabilities are most crucial for the performance of a total conversation service, the current work will be continued in order to consider also the capabilities of core networks as well as transit networks and their security related aspects.

REFERENCES

- [1] ITU-T World Telecommunications Standardization Assembly 2008: Resolution 70 - Telecommunication/ICT accessibility for persons with disabilities; October 2008
- [2] ITU Plenipotentiary Conference 2010: PP 10 Resolution 175 - Telecommunication/Information and Communication Technology accessibility for persons with disabilities, including age-related disabilities; October,2010
- [3] ITU-T F.703: Multimedia conversational services; November 2000
- [4] ITU-T Y.2091 Terms and definitions for Next Generation Networks; March 2011
- [5] ITU-T Y.2201 Requirements and capabilities for ITU-T NGN; September 2009
- [6] ETSI ES 282 001 Telecommunications and Internet converged Services and Protocols for Advanced Networking (TISPAN); NGN Functional Architecture; September 2009
- [7] ITU-T H.510 Mobility for H.323 multimedia systems and services; March 2002
- [8] K. El-Khatib, Y. Zhen, E. Zhang, N. Hadibi and G. v. Bochmann: Personal and service mobility in ubiquitous computing environments; Wireless Communications & Mobile Computing, 2004; 4
- [9] Zhijun Lei and Nicolas D. Georganas: Context-based Media Adaptation in Pervasive Computing; University of Ottawa 2004
- [10] R.L. Aguiar et al: Scalable QoS Aware Mobility for Future Mobile Operators; IEEE Communications Magazine, Vol.44, June 2006
- [11] Leo Lehmann: Implementation of Multimedia Service Mobility in Fix/ Mobile Converged Networks, International Conference on Networking and Services (ICNS 2007); June 2007
- [12] ITU-T Y.2804: Generic framework of mobility management for next generation networks; February 2008
- [13] ITU-T Y.2111: Resource and admission control functions in next generation networks; November 2008
- [14] ITU-T Y.2018: Mobility management and control framework and architecture within the NGN transport stratum; September 2009
- [15] 3GPP TS 26.114 IMS Multimedia Telephony; Media handling and interaction; (Release 10), March 2011
- [16] IETF RFC 5194 Framework for real-time text over IP using the Session Initiation Protocol (SIP); June 2000.
- [17] ITU-T H. Sup1: Application profile - Sign language and lip-reading real-time conversation using low bit rate video communication; Mai 1999
- [18] ITU-T Y.2001 General overview of NGN; December 2004
- [19] 3GPP TS 23.228 IP Multimedia Subsystem (IMS); (Release 10); March 2011

LABQOS: A PLATFORM FOR NETWORK TEST ENVIRONMENTS

Luis Zabala, Armando Ferro, Cristina Perfecto, Eva Ibarrola, Jose Luis Jodra

University of the Basque Country (UPV/EHU)

ABSTRACT

This paper proposes the deployment of a network software platform for experimentation called LabQoS that will allow the scientific community to establish scientific experiments relating to measure the performance of applications and services on the Internet and other network environments. Previous experiences have already deployed measurement systems, but they don't incorporate the concept of control of the scenario to enable the performing of experiments. This LabQoS capacity makes it an innovative platform unique in its approach. A module called Test Environment Builder has been designed in order to monitor the operating parameters, and to propose adaptation strategies to handle the scenario control variables. We define an experimental data model that identifies the entities to be considered in the management of the platform. Assessment mechanisms are proposed to study the dynamic sensitivity of the control variables with respect to the parameters. LabQoS is based on QoS METER architecture that addresses technological aspects such as user management, deployment tools, data collection, reporting, security, etc.

Keywords— Testing Platform, Measurement Tools, Traffic Monitoring, QoS, PQoS, QoE.

1. INTRODUCTION

The role of the scientific environment in the development of Information and Communications Technologies (ICTs) related applications is twofold: first, it helps in the development of new technologies and products; second, it studies the proper design of applications to meet user expectations. In both cases it is essential to carry out complex experiments involving many actors and there are many variables to be controlled. The deployment of these experimental scenarios is a major challenge for the scientific community that may help to study the impact of new technologies on future society.

In the scientific community, there are many initiatives [1] [2] related to the development of tools to measure network performance for different multimedia applications. At the same time, many models and ways of estimating the impact of the quality perceived by the user for different services [3] are being studied. But, in general, the scope of the experiments to test these tools and models is very limited, because most of them are based on too specific approaches [4]. Researchers tend to propose simplistic scenarios

because they want to study individually and, thus, isolate each of the parameters that they are interested in. Once the experiments have been carried out, even researchers from the same field have difficulties to agree on the validity of results from one and another, since they do not have any platform for sharing experiences, reproducing the experiments or for performing larger scale ones. By the former we want to emphasize the need to experiment in more heterogeneous environments, with more elements, and also to make the experiment available to larger population samples than those performed in a laboratory environment. [5][6][7] show some experiments related to residential broadband Internet access services, about the Quality of web sites, or about QoS Provision Assessment where relatively small populations were used.

The LabQoS service aims to develop a software platform that allows scientists to do experiments on the Internet and other network environments for measuring performance of new network technologies and for analyzing the impact of new services on end users.

This paper presents the software platform proposed to achieve these objectives. The paper is divided into the following parts. Section 2 introduces the test environment called LabQoS. Section 3 describes QoS Meter architecture. Section 4 explains how Test Environment Builder works. Section 5 shows the general architecture of the system. Section 6 presents the experimental data model. Section 7 details the control mechanism of the experiment. Finally, in section 8, some conclusions are drawn and possible future improvements are proposed.

2. LABQOS TEST ENVIRONMENT

LabQoS is a software system that allows the deployment of controlled experimentation scenarios in order to make the testing of different environments easier. It also allows to manage a set of users appropriately representative to draw appropriate scientific conclusions. This requires solving, in addition to many scientific issues, others of a technological nature. The availability of QoS METER decentralized architecture [8] enables the deployment of measurement tools in heterogeneous environments and provides solutions to many of the technological requirements of the system.

This project will focus on developing more scientific features that allow to use the architecture as a laboratory for quality of service (LabQoS) and to carry out experiments aimed at serving the research community.

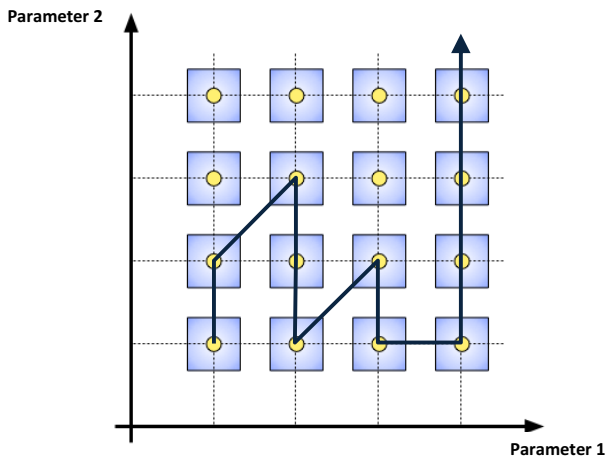


Figure 1. Experimentation map

Figure 1 shows a diagram depicting the working points that an experiment may require. This means that LabQoS test platform should allow experiments to be performed with different combinations of parameters in a stable way. It is complex to achieve it and it requires incorporating into the system: monitoring functions, logic to cross the experiment map, adaptability to make successive approximations to the working points and operating mechanisms to manage other elements of the experimental scenario that allow to control study parameters. It has not been identified any architecture that allows the control of the scenario for experiments measuring Internet services. There are certainly experiences [9] [10] deploying measurement systems as QoSMETER, but none of them incorporates the scenario control concept in order to complete the experiments.

The architecture proposed for LabQoS basically consists on using QoSMETER architecture and incorporating the control intelligence in the experimental scenario of that architecture. Specifically, it will be incorporated within the container module that the users joining an experimental campaign must install on their computers. The container provides some features that are already available such as the deployment of tools, version control, user identification, measure configurations and security systems. In order to add the test scenario control functionality, a new module called Test Environment Builder (TEB) will be designed and it will be installed in the client container (see Figure 2). Each test environment will require the development of an specific measurement tool that will be integrated with the container through an interface already defined. To validate this architecture, it is suggested the integration of a tool for measuring quality of video on the Internet [11]. The integration into this measurement platform requires to implement the necessary interfaces to communicate with the platform.

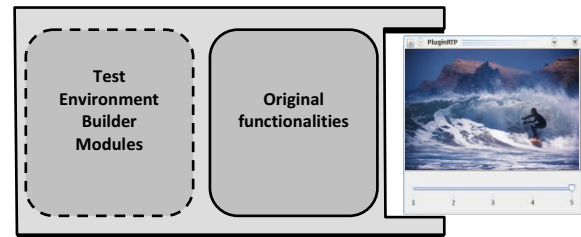


Figure 2. QoS METER architecture container

3. QOSMETER MEASUREMENT PLATFORM

QoSMETER is a generic platform for measuring quality of service. It is designed to host several measurement tools that can cooperate in the analysis of the quality of service in a distributed network. QoSMETER architecture provides a central server which is responsible for the basic functions of the management tools. It also offers a client application which allows to run the desired measurement tools. The client application is called container and it is responsible for providing the appropriate interfaces between the platform and the tools.

3.1. QoSMETER system operation

A QoSMETER typical working case starts when a user downloads a measurement tool container, a software that can contain multiple QoS measurement tools. This container downloads the latest tool configuration, performs the measurement tests in a programmed and controlled environment and finally stores test results in the appropriate storage server. See Figure 3. Once a user has performed several tests correctly, he can request personalized reports. Central server collects users' report requests and schedules their generation, depending on server status and the reports it has previously done. It contacts storage servers in order to achieve the measurement test results and once retrieved, processes them and generates and distributes the corresponding report.

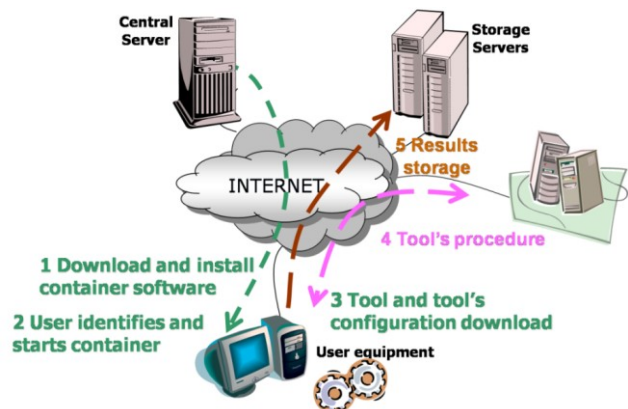


Figure 3. System operation in QoS METER measurement platform

3.2. QoS METER architecture

QoS Meter architecture is divided into four functional parts:

- The Central Server which is in charge of offering platform's general services such as authentication, version control, user profiles, etc.
- The Tool Container which offers the tools to perform measurements through a SOAP interface with the central server.
- The Storage System for results, where data gathered from the measurement campaigns triggered from the different utilities that the system has deployed, are compiled.
- Automated Report Management System is responsible for results related information processing and format as well as for its distribution to final receivers.

3.2.1. Central Server

Central server is the natural root of QoS METER. This is where the system's main logic and functionality resides, as can be seen in Figure 4.

Next central server's main functionalities are described.

- Tool configuration and schedule. User measurement tools need to be configured in order to operate correctly for any environment they are used. A centralized configuration system is proposed so that it can be easily controlled and managed. Users have a web interface to create and manage their tools' configuration, depending on the way they require to perform the measurement. The centralization of the configuration is also useful if users want to use the tools in different machines. This way, configuration is done once and downloaded several times if needed; tools will periodically ask central server for their configuration, querying for it in case a new version is available.
- User administration. QoS METER is going to have lots of final users whom tests are going to be stored and processed by the system. For that reason, it has to correctly identify and maintain some user information to match test results and configuration with the user who has made it. It also should distinguish between user groups, which have different permissions, in order to grant access to different tools and reports.
- System management. Finally, the central server has also the responsibility to make QoS METER administration easier. It has a web administration console through which the system administrator can see all system status and manage tools' configuration and report definition.

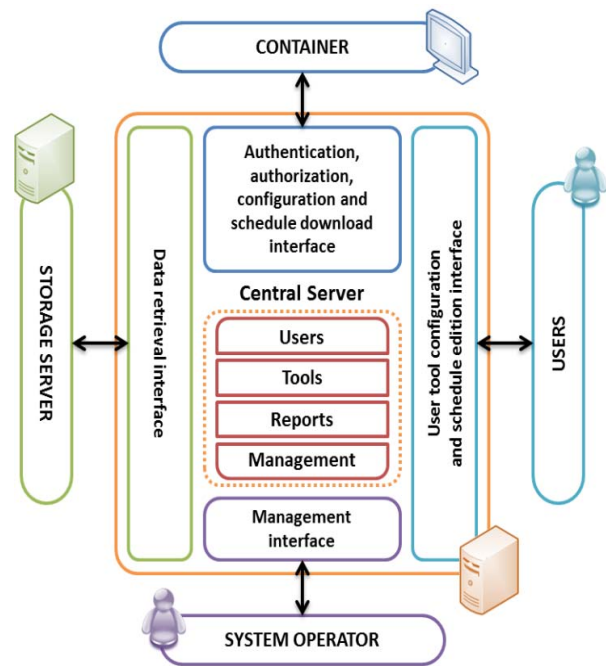


Figure 4. QoS METER Central Server elements

3.2.2. Measurement Tool Container

Measurement tools are the most important part of QoS METER; measurement results are the basis of the system. In order to have the highest number of measurement tests done by users, one of the basic constraints of measurement software is to have an easy to use, intuitive and attractive user interface. For this reason a generic client, called container, has been designed. The container facilitates the development and integration of new measurement tools in an easy to use and configure end-client software.

To acquire the highest number of test results it is necessary to cover a broad range of users, from technicians to others with less computing knowledge. In addition, container software must ensure a controlled environment to run the tests, implementing a standard interface with QoS METER central server, so erroneous values are avoided.

Container software is divided into two different layers: user interface layer and client layer. This way, core functionalities are made independent from the methods to access them. Container's main logic is located at client layer. It is divided in four functional modules: (a) container, (b) security, (c) scheduler and (d) tool manager.

The container module maintains the other modules and is responsible of the instantiation and visibility among them.

The security module is in charge of the user authentication process. It asks the user his username and password and afterwards it asks the central server if they are correct. To

perform this enquire it uses a secure communication link, since the user credentials are exchanged, concretely SOAP above HTTPS. An authentication token is defined, so users do not have to provide their credentials each time they try to do a test. This token is retrieved via secure SOAP after a successful user authentication. It consists in a digitally signed XML file, which contains the relationship between a user and its permissions, in a similar way than an X-509 certificate. All future communications performed by the container, like tool configuration downloading or results storage, will use this token as an authorization method. The security module will transparently include the token in every SOAP message header, so the rest of the modules will not have to deal with this task.

The scheduler's main objective is to execute planned sets of tests and avoid erroneous execution of them. It has a scheduling configuration file where central server lists the order and time periods when the container should carry out the tests. Its most important part is monitoring a set of functions that retrieves local machine info like CPU, memory and network usage, plug-in execution time, resource consumption and so on.

The tool manager module deals with user selected measurement tools that will perform tests in the container. Firstly, it contacts the central server and retrieves the tool list available for this user, offering them through the appropriate user interface. Each tool will have a different configuration, also stored in the central server and ready for download. The tool itself can also be downloaded in case it is not locally accessible in the client machine. Finally, measurement results should be stored in the corresponding store server. These functionalities are performed by different units: (a) plug-in manager, (b) download manager, (c) configuration manager and (d) result manager.

3.2.3. Storage servers

Once measurement values are obtained, they have to be stored in a persistent server, so they are processed afterwards. A storage server responsible either to communicate with clients to recover their measurement results, and to statistically analyze these results has been design. This server is implemented as a web service, using SOAP as the message protocol.

This server has to communicate not only with measurement clients but also with our central server as can be seen in Figure 4. The central server needs to access storage server to ask for processed QoS parameter values in order to calculate the necessary QoS estimation algorithms. Data requests are usually statistical summaries of data grouped by some condition, mainly time periods.

Data representation is one of the most important tasks of this server. A generic language has been defined to describe such measurement values, so that new tools would not need to dedicate much time to develop its storage servers; tools will only need to describe their results in such a way that a generic storage server can understand them.

3.2.4. Report management system

The report management system provides QoS METER platform with the functionalities related to report management, distribution and configuration.

Report generation, one of QoS METER main purposes is to process the obtained measurement parameter values in order to evaluate user Internet access links, distinguishing among requested services (HTTP, e-mail, ftp, etc.). The central server must contact storage servers to acquire measurement results for one or many tools, dealing after with the processing and result representation using the proper algorithm each time. To generalize the report generation process, a report description language is specified. Two different concepts are identified when dealing with this task. The first is the required plain or pre-processed data which is the input of the report and that can come from one or many measurement tools, so multiple parameter QoS evaluation can be done. Each tool's data group is called a data-set. The second is the parameter processing and result representation, called view. A report can have one or many data-sets and views, depending on its contents.

Advanced reports, besides previously cited reports, the system design takes into consideration the possibility of generating extended analyses that require special information processing. For example, system administrators may want to carry out statistical analyses concerning QoS levels in different situations, or to identify root causes of poor QoS levels in a geographical or temporal basis. Since these reports need the utilization of advanced valuation mechanisms or statistical processing, with heavy CPU and memory consumption requirements, the system is provided with a communication interface to external modules. The reason to develop this interface to external processing modules is twofold. Some processing modules have been developed in order to apply different statistical analyses to a set of sampled input QoS parameters. Since these modules are high resource consuming, central server operation could be affected. On the other hand, this interface allows our system to integrate third-party data analysis modules publicly available, as e.g. NetMiner mentioned in [12].

Reports are based on summaries from the information compiled in the storage server for each of the measurement tools. The central server asks the storage server for the data using SOAP for remote method invocation.

A system for report presentation/transformation has also been developed. It is based on templates and the data from the specific service or from the report's local storage if the report has already been requested and is thus kept in the report cache. The system is configurable to allow for report generation according/specific to users' needs.

4. TEST-ENVIRONMENT BUILDER

Test Environment Builder (TEB) is the new system developed to be integrated into QoS METER measurement

platform. TEB adds the capacity to create test environments related to experiments where the quality of service is measured. TEB should provide the following features:

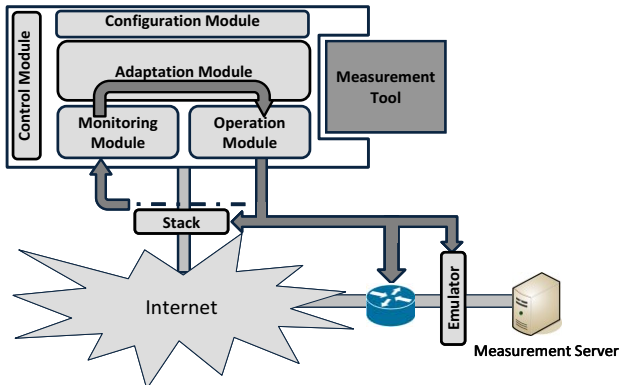


Figure 5. TEB's application scenario

- Configuration of test scenarios. After the LabQoS central service distributes the configuration information for each experiment, the container will distribute that configuration to each of the elements involved in the architecture.
- Monitoring control parameters. In each experiment, the container must control its basic parameters in order to make decisions to adapt and modify working points in the experimental map and proceed to the classification of the measures.
- Adaptation strategies of the experiments. The system must be capable of acting on the scenario in order to maintain control of the experiment parameters and, in this way, the required experimental results will be obtained.
- Management of network elements. The parameters of the experiment must be adjusted operating on specific elements of the scenario that offer this possibility. For that, the container requires an operation module that allows it to interact with network elements.
- Logic for going through the experiments. A multidimensional experimental map will be defined in each experiment. That map indicates the working points you want to go across during the test. Each point on the map refers to a particular combination of experiment parameters. The system must have the appropriate logic to perform the experiments at all points of interest.

Figure 5 shows the layout of the main elements of the architecture in a concrete application scenario based on an emulator [13].

5. SYSTEM MODULES

The architecture of the Test Environment Builder is based on modules. There are the following ones:

- Configuration module. It allows to load the configuration parameters in a flexible way.

- Adaptation module. It is responsible for keeping a stable behavior in the experimental scenario, taking into account the most interesting parameters.
- Operation module. It is responsible for connecting to the elements which are controlled by this module in order to handle network parameters.
- Monitoring module. It allows to read the value of the parameters which are interesting in the test scenario.
- Control module. It coordinates the activities of the different modules and the test execution.

At the same time, it is possible to distinguish other lower level modules in each one of those main modules as Figure 6 shows.

Before describing in more detail those five modules, it is necessary to remind that, as Figure 2 showed, the container of QoS METER is the element which will be modified in order to implement the design of LabQoS service. Therefore, the new functionalities provided by those modules will be added to the container.

5.1. Configuration module

The configuration module is responsible for all tasks related to configuring the system, i.e., it is responsible for obtaining the configuration data, validating those data and verifying that all are correct, initializing the data model defined below with the configuration obtained and, at last, informing about the completion of the configuration to the control module.

5.2. Adaptation module

The adaptation module is responsible for maintaining the scenario in a stable way, against changes introduced on the Internet or the network. For this reason, it is obvious that this module contains the most of the intelligence of the system and, therefore, it is the most complex.

The adjustment is basically to measure the parameters of quality of service and compare the values obtained with those ones which are desired for the test scenario. Based on this comparison, the system will act in such a way that you can get a test scenario as close as possible to the desired one.

This set of operations belongs to the "adaptive cycle" and they should be run periodically. The purpose of those operations is to keep the scenario as stable as possible.

5.3. Operation module

This module is responsible for simulating or emulating all network parameters that the scenario needs. For this, it offers an interface that can dynamically load different operators. Using such operators, the system can interact with different elements that can modify network parameters.

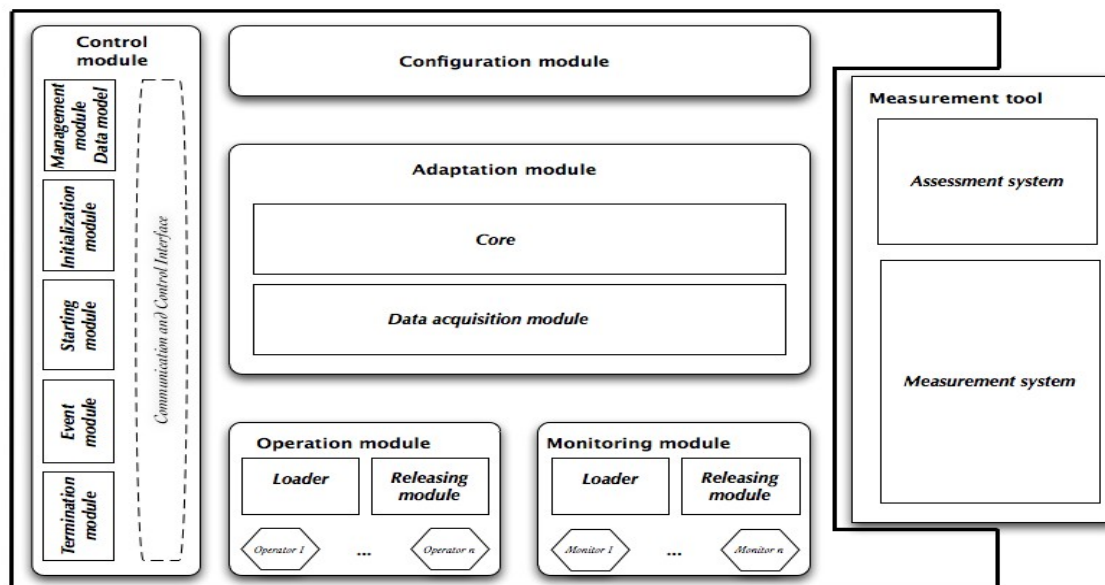


Figure 6. System modules in LabQoS

5.4. Monitoring module

This module provides the capacity of performing QoS parameter measurements that are needed to carry out the adaptation of the scenario. The monitoring module, as well as the operating module, just provides an interface which can dynamically load monitors, thus giving a greater flexibility to the system.

5.5. Control module

The control module is responsible for keeping track of the system with the aim of giving it all what it needs for the different measurement tests. It is responsible for system initialization and termination, management of the communication between modules, definition of interfaces needed to connect different modules, event control...

6. DATA MODEL

LabQoS' data model is called IBTE (Information for Building a Test Environment). It defines the format for storing the information related to the test environment which has been built. Every module of the system has access to it and knows how to interpret it. The model is based on a set of entities that are related to each other in order to take an effective control on the test scenario:

- "Test" entity: It represents all the network situations to be achieved during the performance of a test. An

experimentation map, for example the one showed in Figure 1, tries to bring together all the working points of the test. A network situation is defined with parameters of quality of service.

- "Snap" entity: This entity represents each one of network situations that compose the experiment. So, it is figured as a working point in the experimentation map and it is required to store the values of the parameters which define the working point.
- "Parameter" entity: It represents a parameter of quality of service. A set of parameters define an experiment. The "parameter" entity has to contain enough values to define the distribution of the parameter in the experimentation map.
- "Control Variable" entity: It is the item to be modified, in order to achieve a variation of a specific QoS parameter. There is a relation between the parameters that will be changed and the variables which control the variation. It is possible that different control variables modify the same parameter.
- "Policy" entity: This entity allows us to define different forms of action on the control variables, such a way that the wanted value is achieved for a specific parameter. Thus the system gains in flexibility.
- "Box" entity: It allows us to store all the measures collected, such a way that repeating snaps unnecessarily is avoided. It also controls the snap inside the test.

Figure 7 shows the five entities that make up the model, as well as the relations between them.

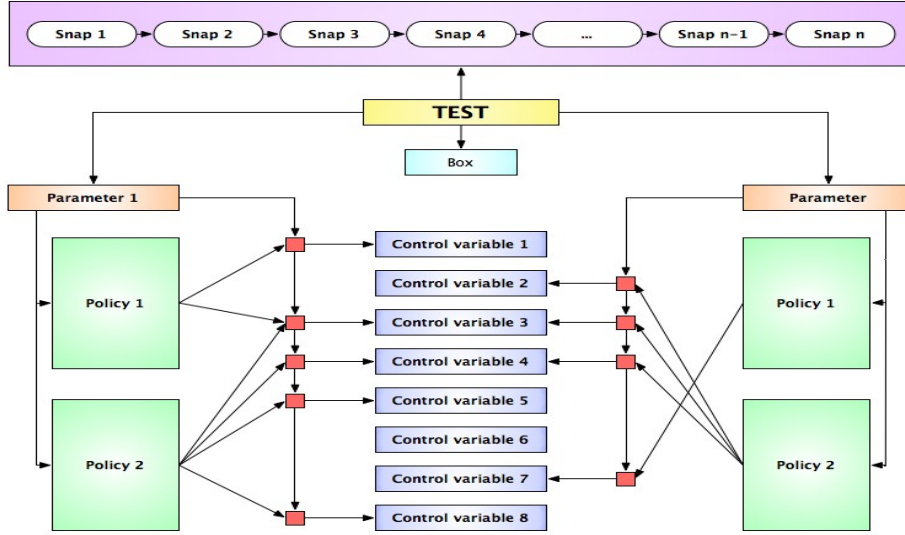


Figure 7. Data model

7. TEST CONTROL MECHANISM

It is important to keep the test working at appropriate coordinates all the time for each case. For this, a control mechanism is established to allow to adjust the test parameters on the basis of the information monitored. In case of identifying deviations in the value achieved with respect to the desired one, a policy will act to control parameters in each cycle of adaptation. The adaptation strategy is based on the analysis of the parameters' gradient ∇ , control variables' sensitivity δ and the necessary correction factor to decide the form and intensity of adaptation.

The "control variable" allows to identify the relationship between a parameter and the different elements of the scenario which can control and act on the parameter. It is defined the concept of sensitivity δ that relates numerically to the variation of a parameter with respect to the variation of a control variable. The mathematical expression that will be used to calculate the sensitivities of all the parameters is as follows.

$$\delta = \frac{\Delta \text{ControlVariable}}{\Delta \text{Parameter}} \quad (1)$$

This definition of sensitivity provides a degree of intelligence to the system, because it allows us to relate the changes in a parameter to the ones in a particular control variable. The gradient ∇ of a parameter set the difference between the measured value of the parameter and the desired one.

A numerical value called corrector ε is also defined. It represents the theoretical variation experienced by the parameter, due to the modifications in other control variables in the current adaptation cycle. For that reason, the relationships between the parameters and control

variables can be identified in the model data and it is possible to know which variables are affected and in the adaptation strategy for each parameter.

After calculating the sensitivity, the system can predict how it will be affected by the variable in that parameter. The mathematical expression used for this is deduced from the definition of sensitivity (1), finding the increase of the parameter. The increase in the control variable is determined by the modification that is to be carried out. This cycle is repeated for all control variables, so the final correction factor is the sum of all parameter increments calculated for each of the variables.

$$\text{Correction} = \sum_i \Delta \text{Parameter}_i \quad (2)$$

Having calculated the gradient and the correction factor, a parameter adaptation strategy begins, using the right policy. The first step is to study whether there is an available policy related to the control variable, in order to act on any parameter. Once confirmed that the policy applies, you have to check all the control variables of the policy and act on those which are available. It is also necessary to inspect if there are conflicts in the strategy. The value to be put in the control variable is calculated using an expression that is also deduced from the definition of sensitivity (1), finding the increase of the control variable. The final increase of the parameter depends on the gradient and the correction factor using the following expression:

$$\Delta \text{var} = \delta \cdot (\nabla \text{par} + \varepsilon) \quad (3)$$

The appropriate adaptation policy uses this calculation in order to send configuration changes in the test scenario through the operation module.

8. CONCLUSIONS

This paper proposes the deployment of a network software platform called LabQoS that will allow the scientific community to establish scientific experiments related to the performance measurement of Internet applications and services. LabQoS capacity to handle the experimental scenario makes it an innovative platform unique in its approach. The logic of the experimental adaptation is one of the most important scientific contributions of this work.

There are many complaints about how difficult it is to check the results of a scientific paper about network technologies. The main reason is the lack of infrastructure availability to replicate the experiment. LabQoS offers a solution to share tests and have a common platform where researchers can test and validate other researchers' scientific experiments.

Finally, it is also remarkable that another consequence of the LabQoS platform design has been the detection of possible improvements for the QoS METER measurement system. Since the Test Environment Builder consists of modules which work inside QoS METER container, new proposals have emerged to enhance the configuration system and the result sending system of QoS METER. For that reason, our next steps on LabQoS platform will be directed towards getting better integration of Test Environment Builder within the new enhanced version of QoS METER.

REFERENCES

- [1] M. Dabrowski, J. Sliwinski, F. Strohmeier, 'The MOME Workstation as a Platform for Automatic Analysis of Measurement Data', Distributed Cooperative Laboratories: Networking, Instrumentation, and Measurements, 409–424.
- [2] D. Morato et al., 'The European traffic observatory measurement infrastructure (etomic): A testbed for universal active and passive measurements', Proc. of Tridentcom 2005, 23–25.
- [3] F. Liberal, A. Ferro, J.L. Jodra, J.O. Fajardo, "Application of General Perception-Based QoS Model to Find Providers Responsibilities. Case Study: User Perceived Web Service Performance," 2005.
- [4] K. Campowsky et al, 'Pan-European Testbed and Experimental Facility Federation–Architecture Refinement and Implementation', International Journal of Communication Networks and Distributed Systems 5 (2010), No.1/2, pp.67-87.
- [5] SSRC, "Survey on residential broadband internet access services", 2007.
- [6] J.v. Iwaarden et al, "Perceptions about the quality of web sites: a survey amongst students at Northeastern University and Erasmus University " Information and Management vol. 41, pp. 947-959, 2004
- [7] E. Babulak, "Quality of service provision assessment for campus network," presented at Mobile Future and Symposium on Trends in Communications, 2003.
- [8] R. Partearroyo et al. 'QoS METER. Generic quality of service measurement infrastructure'. Towards the QoS Internet. Coimbra (Portugal), May 15-19, 2006.
- [9] C. Elliott, 'GENI–Global Environment for Network Innovations', in 33rd IEEE Conference on Local Computer Networks, 2008. LCN 2008, pp. 8–8.
- [10] M. Carbone et al, 'Wireless link emulation in OneLab', in 2nd International Workshop on Real Overlays And Distributed Systems. Warsaw (Poland), 2007.
- [11] D. Basoko, "Sistema de medida de calidad de calidad de flujos multimedia bajo IP". ETSI Bilbao. October 2009.
- [12] M. Baldi et al, 'Data mining techniques for effective and scalable traffic analysis', 9th IFIP/IEEE International Symposium on Integrated Network Management, May 2005, pp. 105-118.
- [13] . Linux Network Emulator.
<http://www.linuxfoundation.org/en/Net:Netem>.

POSTER SESSION

SHOWCASING INNOVATIONS FOR FUTURE NETWORKS AND SERVICES

- P.1 A Trust Computing Mechanism for Cloud Computing
- P.2 The Energy Label A Need To Networks And Devices
- P.3 A distributed mobility management scheme for future networks
- P.4 Toward Global Cybersecurity Collaboration: Cybersecurity Operation Activity Model
- P.5 Context Representation Formalism and Its Integration into Context as a Service in Clouds
- P.6 Supporting technically the Continuity of Medical Care: Status report and perspectives
- P.7 Coexistence of a TETRA System with a Terrestrial DTV System in White Spaces
- P.8 Mobile cloud computing based on service oriented Architecture: embracing network as a service for 3rd party application service providers
- P.9 RBAC for a configurable, heterogeneous Device Cloud for Web Applications

A TRUST COMPUTING MECHANISM FOR CLOUD COMPUTING

Mohamed Firdhous^{1*}, Osman Ghazali², Suhaidi Hassan³

InterNetWorks Research Group, Universiti Utara Malaysia, Sintok, Kedah Darul Aman, Malaysia

Email: mfirdhous@internetworks.my¹, osman@uum.edu.my², suhaidi@uum.edu.my³

ABSTRACT

Cloud computing has been considered as the 5th utility as computing resources including computing power, storage, development platform and applications will be available as services and consumers will pay only for what consumed. This is in contrast to the current practice of outright purchase or leasing of computing resources. When the cloud computing becomes popular, there will be multiple vendor offering different services at different Quality of Services and at different prices. The customers will need a scheme to select the right service provider based on their requirements. A trust management system will match the service providers and the customers based on the requirements and offerings. In this paper, the authors propose a trust formulation and evolution mechanism that can be used to measure the performance of cloud systems. The proposed mechanism formulates trust scores for different service level requirements, hence is suitable for managing multiple service levels against single trust score. Also the proposed mechanism is an adaptive one that takes the dynamics of performance variation along with cloud attributes such as number of virtual servers into computations. Finally the proposed mechanism has been tested under a simulated environment and the results have been presented.

Keywords — Cloud Computing, Trust Formulation, Trust Evolution, Quality of Service

1. INTRODUCTION

Cloud computing has become the new paradigm in networked computing and it has been identified as the 5th utility after electricity, water, gas and telephony [1]. The emergence of cloud computing has helped organizations to change their strategy towards the investment in computing resource from own and operate to pay for what is used. For cloud computing to be accepted by a wider audience, the users need an assurance that we would receive what has been promised. This kind of assurance can be provided by a Service Level Agreement (SLA) signed between the parties. But, the clients require a method to identify the service providers who could meet their requirements. A

reputation management system that quantifies the service levels would be an ideal solution from which users can select a service to suit their budgets. In this paper, the authors propose a mechanism for computing trust metrics that would form the basis for a reputation management system.

This paper is divided into six sections. Section 1 introduces the paper, while Sections 2 provides a brief introduction to cloud computing. Section 3 discusses trust and quality of service in depth and Section 4 introduces trust formulation and evolution mechanisms proposed in this paper. Simulation environment and the results are presented in Section 5. Section 6 concludes the paper along with suggestions for future work.

2. CLOUD COMPUTING

Cloud computing has been identified as the 5th utility in the line of electricity, water, telephony and gas [1]. Cloud computing has been given such a name due to the similarity between these services with respect to the way they have been accessed and paid for. Utilities have been accessed and consumed by consumers without worrying about how the services have been generated and paid only for the actual consumption of the service. With the advent of cloud computing, even the computing services will be accessed by users in a similar fashion and paid only for the services accessed. Prior to the arrival of cloud computing, computing resources were either purchased outright or leased from data center provided at fixed rates, irrespective of usage.

Cloud service providers host their services on the Internet and make them available to the prospective customers. Customers can access these services whenever they would want them and pay only for the services accessed. Service providers host their services on virtualized systems so that the same resource can be sold to multiple customers achieving maximum utilization from the resources. The virtualized systems provide the customers a sense of feeling that the resources are dedicated only for them whereas the actual resources are shared between multiple users [2]. Sharing resources this way increases the productivity of the systems while decreasing the cost of resources per user.

Cloud services are currently marketed under three different categories namely Infrastructure as Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS) [3].

*Mohamed Firdhous is a Senior Lecturer attached to the Faculty of Information Technology, University of Moratuwa, Sri Lanka. He is currently on leave pursuing his PhD at the Universiti Utara Malaysia.

Provision of raw computer infrastructure in terms of virtual computers is known as IaaS in cloud computing terminology. Once a virtual computer has been purchased, users can install the operating system of their choice and applications independent of other systems hosted on the same physical infrastructure. PaaS is the provision of facilities and Application Programming Interfaces (APIs) to support the complete life cycle of building and delivering web applications and services. SaaS is a model of software deployment where the user applications are hosted as a service and made available to users over the Internet [4]. Figure 1 shows the layered architecture of a typical cloud computing system. This figure includes two additional layers namely the physical hardware layer and the virtualized hardware layer in addition to the cloud service layers.

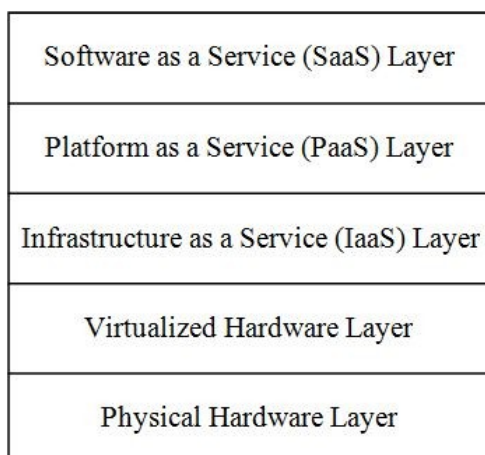


Figure 1. Layered Architecture of Cloud System

From Figure 1, it can be seen that the cloud system is made up of five layers in total. The bottom most layer, the Physical Hardware Layer is usually made up by server class computers in data centers, clusters, grids, storage networks or any other computing systems. This is the workhorse layer which provides the necessary physical resources in terms of processors, memory, bus, storage, networking etc., to carry out the basic computing operations.

Virtualized Hardware Layer running on top of the physical hardware is created by virtualization software. The virtualization software slices the physical hardware into virtual machines in such a manner that each virtual machine will act like an independent computer running its own operating system along with other resources. These virtual computers can be pooled together to act as single resource pools. The capability of pooling the resources together makes the system elastic in the sense the virtual computers can be brought online and assigned to pools on demand. Similarly virtual computers can also be destroyed when demand subsides. This ability to create and destroy virtual computers dynamically is the basis on which IaaS is built upon. VMware, Virtual Machine Monitor (VMM), Xen, Kernel-based Virtual Machine (KVM) are some of the main products in this market.

IaaS Layer provides the clients with the facility of computing infrastructure similar to raw computing hardware [4]. Clients can install the operating system of their choice and any application development platform as if they own their own hardware. Clients are relieved from managing the physical resources such as physical computers, power and the networking but they have full control over the operating system, storage, and applications. Clients also have the flexibility of purchasing different virtual hardware components from different vendors and combine them together to form their own systems. There are several commercial IaaS providers specializing in different types of IaaS services from who customers can purchase the service they wish. Amazon provides two types of IaaS services, namely Amazon Elastic Computing Cloud (EC2) that provides flexible computing capacity and Amazon Simple Storage Service (S3) that provides flexible storage services over the Internet. IBM Smart Business Test Cloud provides a complete test environment comprising operating systems, middleware, storage, network, images and data. This reduces both the cost and time of software development drastically. The Nirvanix Storage Delivery Network and Oxygen Cloud are flexible cloud storage service that can be accessed over the internet. Interactive Intelligence provides a comprehensive set of on-demand services for cloud-based communications applications under the name of Communication as a Service (CaaS).

The Platform as a Service (PaaS) Layer extends the IaaS by abstracting it by providing an operating system and development tools creating an environment that supports the complete software development life cycle. The PaaS Layer eliminates the hassle associated with managing virtual computing instances and provides a uniform programming platform to the end user. Google's App Engine, Amazon Elastic Beanstalk and Force.Com platform are typical PaaS offerings in the market. Google App Engine supports Python and Java programming languages along with other tools for developing and hosting web applications. The App Engine sandboxes the application to provide a secure environment for applications. The sandboxed environment isolates the application and makes it independent of the underlying hardware, operating system and physical location of the web server. App Engine also provides a distributed data storage with query and transaction processing. The Elastic Beanstalk is the PaaS service provided by the Amazon to deploy and manage any Java application in the Amazon Web Service (AWS) cloud. The Elastic Beanstalk helps any Java based web application to be loaded to the AWS as a standard Java Web Application Archive and be deployed as cloud based application. Force.com platform is a slightly different from App Engine and Elastic Beanstalk. Force.com only allows developers to create add-on application that can be integrated to the main salesforce.com application and hosted on the salesforce.com's infrastructure. These application add-ons are to be built using a proprietary Java-like programming language called Apex. The user interfaces need to be developed using Visualforce another proprietary software.

SaaS is the top most layer in the cloud services stack. Applications that were usually installed and run on individual computers are made available over the Internet as services under the SaaS. This relieves the customers from purchasing, installing, running and managing software applications. There are several commercial SaaS providers in the market and the new offerings are everyday. Google Apps, Customer Relationship Management (CRM) solution by Salesforce.com, IBM LotusLive and SAP CRM are some of the prominent SaaS offerings in the market.

Table 1 provides a summary of commercial cloud service providers along with the names and types of services offered.

Table 1. Summary of Commercial Cloud Services

Service provider	Name of service	Type of service
Amazon	Elastic Compute Cloud (EC2)	IaaS
	Amazon Simple Storage Service (s3)	IaaS
	Amazon Elastic Beanstalk	PaaS
Nirvanix	Nirvanix Storage Delivery Network	IaaS
Google	Google App Engine (GAE)	PaaS
Microsoft	Windows Azure Platform	IaaS
Rackspace	Rackspace Cloud Servers	IaaS
SalesForce	Force.com	SaaS
	Force.com Platform	PaaS
HP	HP Software-as-a-Service (Opware)	SaaS
GoGrid	GoGrid	IaaS
ElasticHosts Ltd	ElasticStack	IaaS
Flexiant Ltd	FlexiScale	IaaS
Oracle	Sun Cloud	IaaS
IBM	Blue Cloud	IaaS

3. TRUST AND QUALITY OF SERVICE

The trust and reputation have their origin in the social sciences that study the nature and behavior of human societies [5]. Trust has been studied by researchers in diverse fields such as psychology, sociology, and economics [6]. Trust management systems play an important role in distributed systems such as peer to peer systems, grid computing, cluster computing and sensor networks [7-11]. Trust management systems help nodes to select the right peer to interact with [12].

Trust basically represents a node's competence, benevolence, integrity or predictability and any mathematical model defined to represent trust must be

capable of representing all these aspects [13]. Several authors have attempted to model trust [14-17]. All these models discussed lack theoretical formulation of trust and stopped at proposing some ideas only. For Services offered in commercial would become successful only when they deliver the promised Quality of Service (QoS) [18]. A mechanism is necessary for clients to select the right service provider who could meet their requirements. A trust system built based on the QoS of different service providers will be useful in matching the capability and requirements of both service provider and clients. In this paper, the authors propose a trust mechanism based on QoS that can be used by clients to select the service providers.

QoS has been studied extensively by several researchers and reported in literature based on various QoS metrics such as response time, throughput and network utilization [18]. Xiong and Perros derive a model for computing QoS of cloud computing based on the required percentile response time [18]. They have used the M/M/1 queuing model for the analysis. Though this analysis sheds a certain amount of light into the performance of cloud computing system, the queuing model used in the analysis does not represent the real cloud environment. The cloud system is based on the virtualization of the hardware and the capability of spawning virtual machines dynamically to meet the customer requirements. Hence the model needs to be changed to M/M/n where n represents the maximum number of virtual machines that can be spawned by a physical computer in order to represent the real environment. In this paper, an analysis will be carried out based on the M/M/n queuing model for three types of customers, namely;

1. Customers who require a guaranteed level of service
2. Customers who require an average level of service
3. Customers who require basic level of service with no guarantees

Type I customers who require a guaranteed level of service would be willing to pay a comparatively large fee for the guarantee. This type of service is required for mission critical services. Type II customers who require an average level of service would pay a lower fee and would be happy when they receive the service with slight variations. This is suitable for essential but non critical services. Type III services are for non essential non real time services. The customers should be charged the lowest fee for this type of service.

4. BUILDING OF TRUST

Trust formation, evolution and propagation are central issues in trust management [13]. In this paper the authors propose a model for trust formation and evolution based the Quality of Service of cloud nodes. Figure 2 shows the proposed system that is used to form, evolve and manage the trust of computing nodes in a cloud system. The trust formulation unit computes the initial trust values based on the type of service and level of service. Service monitor

monitors the performance of the service provider and informs the trust evolution unit if the service was carried out satisfactorily or not. Trust evolution unit keeps track of the current trust values for different service types and evolves the them based on the feedback received from the service monitor.

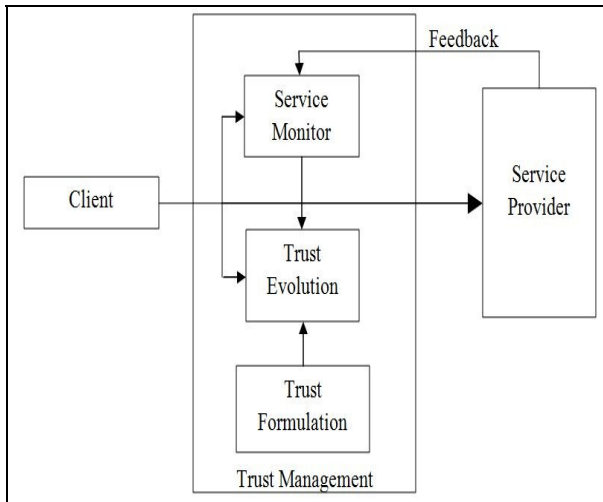


Figure 2. Trust Management System

4.1. Trust Formulation

Figure 3 shows the queuing model for the purpose of formulating trust in the cloud system. The Erlang C queuing model denoted by M/M/n in Kendall notation is used as it is the most suitable model to represent the practical cloud environment.

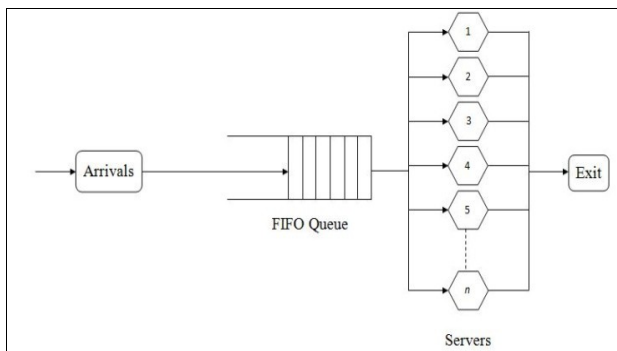


Figure 3. Queuing Model used for Formulating Trust

A FIFO queue with infinite waiting slots is assumed for simplicity. Infinite waiting slots ensure that every customer arriving at the queue be served even after a long waiting time. Every client entering the system is treated equally with no priority and treated according to the First In First Served (FIFS) discipline. Any client leaving the queue and reentering the system due to any reason is treated as a new arrival and added to the queue at the end. The arrival of requests is assumed to be Poisson distributed with a mean of λ and service time is assumed to be exponentially distributed with a mean of μ .

The initial trust values are formulated by computing the probability that the system would meet customers required response time. For example, if a customer requires the response time to be no more than τ_r and the system can meet this requirement with the lowest probability of σ . Then the initial trust score for that class of request is determined to be σ . This initial trust score will be modified according to the feedback received from customers based on the actual performance of the service provider. Let τ , $f(t)$ and $F(t)$ represent the response time, probability distribution function and cumulative distribution function respectively.

If τ_r is the required response time of the customer, the response time should satisfy eq. (1).

$$F(t)|_{t=\tau_r} = \int_0^{\tau_r} f(t) dt \geq \sigma \tag{1}$$

where σ – the required probability

The steady state probabilities can be derived as [19];

$$p_0 = \left[\sum_{i=0}^{n-1} \frac{(n\rho)^i}{i!} + \frac{(n\rho)^n}{n!} \frac{1}{(1-\rho)} \right]^{-1}$$

and

$$p_i = \begin{cases} p_0 \left(\frac{\lambda}{\mu}\right)^i \frac{1}{i!} & \text{for } i < n \\ p_0 \left(\frac{\lambda}{\mu}\right)^i \frac{1}{n! n^{i-n}} & \text{for } i \geq n \end{cases}$$

$$\text{where } \rho = \frac{\lambda}{(n\mu)}$$

The probability distribution $f(t)$ of response time τ is given by;

$$f_\tau(t) = \alpha e^{-\alpha t} \tag{2}$$

$$\text{where } \alpha = \left[\frac{\rho}{\lambda(1-\rho)^2} \cdot p_n + \frac{1}{\mu} \right]^{-1}$$

From Eqn. (1) and (2);

$$F(t)|_{t=\tau_r} = \int_0^{\tau_r} \alpha e^{-\alpha t} dt$$

$$F(t)|_{t=\tau_r} = (1 - e^{-\alpha \tau_r}) \tag{3}$$

Equation (3) can be used compute the initial trust score in terms QoS requirement, given the mean arrival rate, service time and the number of virtual servers.

4.2. Trust Evolution

The trust evolution module updates the trust value based on the feedback received from users. The feedback received from the users can be of two types, namely positive response where the actual response time is less than the required response time or otherwise indicating an inferior

performance than required. The positive response would be used to improve the trust score while the negative response would reduce the trust score based on the algorithm shown in Figure 4.

```

required response time =  $\tau_r$ 
actual response time =  $\tau_a$ 
compute normalization parameter ( $\delta$ ) =  $\frac{|\tau_r - \tau_a|}{\tau_r}$ 
if ( $\tau_a \leq \tau_r$ )
    update all trust score where ( $\tau_r \geq \tau_a$ )
     $T_{n+1} = T_n + \delta * T_n$  :  $T_0 = a$  and  $n = 1, 2, \dots$ 
else
    update all trust score where ( $\tau_r < \tau_a$ )
     $T_{n+1} = T_n - \delta * T_n$  :  $T_0 = a$  and  $n = 1, 2, \dots$ 
end

where  $a$  – initial trust score computed
    
```

Figure 4. Trust Evolution Algorithm

This algorithm updates multiple trust scores based on the feedback received. The algorithm bases its decision on the assumption, that if the system meets a lower response time, it can meet all the response times higher than that. Also, if the system does not meet a higher response time, it cannot meet the lower response times. Initial trust score (a) for different response times is computed by the trust formulation unit. The normalization factor has been included into the calculation to reflect the performance of the system directly on the improvement (reduce) the trust scores.

5. SIMULATIONS

The performance of the proposed algorithm was tested using simulations. A simulation environment comprising trust formulation trust evolution, service provider and service monitor units has been setup using GNU Octave. The M/M/n queue was simulated using the qnetworks, the Queuing Networks analysis package for GNU Octave [20]. Figure 5 shows the initial trust scores computed with the mean arrival rate of 200 requests per second and a uniform service time of 75 seconds for different number of servers. From the figure, it can be seen that the initial trust scores increases drastically as the number of virtual servers increases.

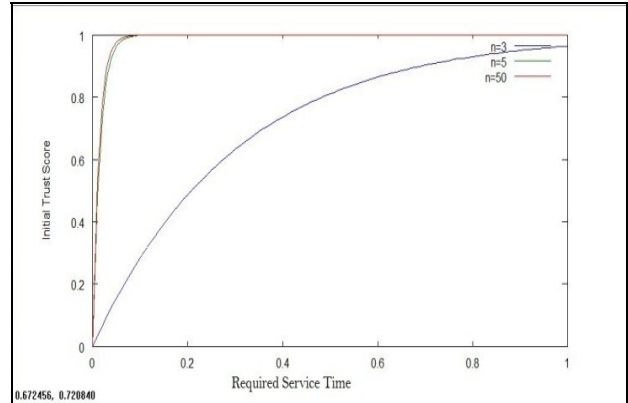


Figure 5. Initial Trust Scores for Different Service Times

Figure 6 shows the changes in trust scores due to continuous positive and negative feedbacks. The continuous negative feedbacks reduce the trust score initially at a higher rate but with time the rate also comes down though a constant time lag was used as the difference. This is due to the reason that the rate of reduction depends both on current trust value and the normalization parameter calculated using the differences in required and real response times. Similarly, during improvement of trust scores, larger trust scores responds fast to improvements compared to the smaller ones. This adaptive nature of rate of improvement (reduction) helps the system to respond to customer requirements fast as the customers who require a higher trust score would also be more sensitive to changes in trust compared to others.

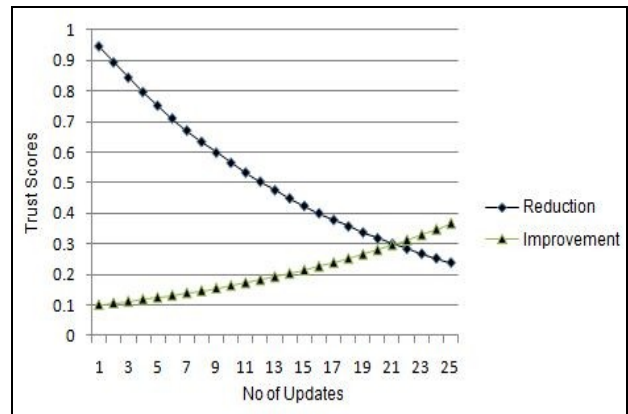


Figure 6. Change in Trust Scores

Figure 7 shows effect of response time on the trust scores. For the purpose of comparison four classes of services characterized by different response time requirements were taken into consideration. The response requirements are namely 0.1, 0.4, 0.7 and 0.9 time units. The time units have been normalized to lie between 0.0 and 1.0 for the convenience of comparison. The 0.1 is the most stringent requirement while 0.9 being the most relaxed.

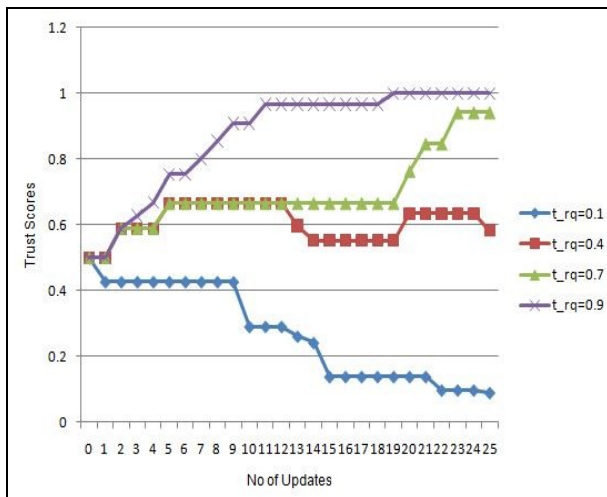


Figure 7. Comparative Change in Trust Scores

From the figure, it can be seen that whenever a more stringent requirement has been met, all the trust values of the relaxed requirements have also been improved. This is due to the reason that, if the system could meet a stringent condition it could easily meet relaxed requirements. This fact should reflect on the trust scores and hence all the respective trust values have been positively updated. Conversely, when a relaxed condition is not met, all the trust scores of the more stringent requirements are reduced. This is due to the reason that the failure to meet a relaxed requirement would necessarily an indication that more stringent performance requirements will not be met.

The figure also shows that the most stringent condition indicated by the requirement of 0.1 continues to decline. This is due to the reason that the negative performance of the system for any requirement lower than this requirement would affect this one. Hence, it is obvious from the results that it is very difficult to meet strict performance requirements unless special attention has been paid to these requirements. On the other hand, the trust score of the most relaxed performance requirement designated by the time response of 0.9 shows continuous improvement. This is due to the collective improvement of all the more stringent requirements. The other two plots show mixed results due to the random effect on their trust scores.

6. CONCLUSIONS

This paper presented a trust formulation and evolution model for cloud computing. Cloud computing has become the new paradigm in computing and accepted as the 5th utility after electricity, water, gas and telephony. For cloud computing to be accepted by different types of users, it needs to provide assurance to clients on service quality depending on the user requirements. Trust system would help users to select service providers based on the quality requirements. In this paper, the authors have proposed a trust system built based on the response time. The trust system provides a trust score between 0 and 1 for different

levels of services and continues to improve these values based on the performance of the system. Hence the proposed system would be more useful for providing differentiated services at different quality levels. The proposed mechanism has been evaluated using a simulation environment setup with Octave the open source Matlab clone. The simulation results show that the proposed system works satisfactorily under constrained simulated environment. The proposed mechanism must be tested rigorously under a more open environment and in the face of adversaries in order to evaluate the ruggedness and resilience of the mechanism. The authors propose to carry out this in a future research.

REFERENCES

- [1] R. Buyya, C.S. Yeo, S. Venugopal, J. Broberg, and I. Brandic, "Cloud computing and emerging IT platforms: vision, hype, and reality for delivering computing as the 5th utility," *Future Generation Computer Systems*, vol. 26, no. 6, pp. 599-616, June 2009.
- [2] M.D de Assuncao, A. di Costanzo, and R. Buyya, "Evaluating the cost-benefit of using cloud computing to extend the capacity of clusters," in *18th ACM international symposium on High performance distributed computing (HPDC '09)*, Munich, Germany, pp. 141-150, 2009.
- [3] C. Vecchiola, S. Pandey, and R. Buyya, "High-performance cloud computing: A view of scientific applications," in *10th International Symposium on Pervasive Systems, Algorithms, and Networks*, Kaohsiung, Taiwan, pp. 4-16, 2009.
- [4] R. Prodan, and S. Ostermann, "A survey and taxonomy of Infrastructure as a Service and web hosting cloud providers," in *10th IEEE/ACM International Conference on Grid Computing*, Banff, AB, Canada, pp. 17-25, 2009.
- [5] H. Yu, Z. Shen, C. Miao, C. Leung, and D. Niyato, "A survey of trust and reputation management systems in wireless communications," *Proceedings of the IEEE*, vol. 98, no. 10, pp. 1755-1772, October 2010.
- [6] Z. Gan, J. He, and Q. Ding, "Trust relationship modelling in e-commerce-based social network," in *International conference on computational intelligence and security*, Beijing, China, pp. 206-210, 2009.
- [7] B. Cai, Z. Li, Y. Cheng, D. Fu, and L. Cheng, "Trust decision making in structured P2P network," in *2009 International Conference on Communication Software and Networks*, Macau, China, pp. 679-683, 2009.
- [8] V. Vijayakumar and R.S.D.W. Banu, "Security for resource selection in grid computing based on trust and reputation responsiveness," *International Journal of Computer Science and Network Security*, vol. 8, no. 11, pp. 107-115, November 2008.
- [9] S. Mishra, D.S. Kushwaha, and A.K. Misra, "A cooperative trust management framework for load balancing in cluster based distributed systems," in *International Conference on*

Recent Trends in Information, Telecommunication and Computing, Kochi, Kerala, India, pp. 121-125, 2010.

[10] Y. Wu et al., "Automatically constructing trusted cluster computing environment," *The Journal of Supercomputing*, vol. 55, no. 1, pp. 51-68, January 2011.

[11] K.K. Tae, and S.S. Hee, "A trust model using fuzzy logic in wireless sensor network," *Journal of World Academy of Science, Engineering and Technology*, vol. 42, no. 13, pp. 63-66, 2008.

[12] M. Firdhous, O. Ghazali, S. Hassan, N.Z. Harun, and A. Abas, "Honey bee based trust management system for cloud computing," in *3rd International Conference on Computing and Informatics (ICOCI 2011)*, Bandung, Indonesia, 2011.

[13] M. Carbone, M. Nielsen, and V. Sassone, "A formal model for trust in dynamic networks," in *First International Conference on Software Engineering and Formal Methods (SEFM'03)*, Brisbane, Australia, pp. 54-61, 2003.

[14] K.M. Khan, and Q. Malluhi, "Establishing trust in cloud computing," *IT Professional*, vol. 12, no. 5, pp. 20-27, 2010.

[15] Z. Song, J. Molina, and C. Strong, "Trusted anonymous execution: a model to raise trust in cloud," in *9th International Conference on Grid and Cooperative Computing (GCC)*, Nanjing, China, pp. 133 – 138, 2010.

[16] H. Sato, A. Kanai, and S. Tanimoto, "A cloud trust model in a security aware cloud," in *10th IEEE/IPSJ International Symposium on Applications and the Internet (SAINT)*, Seoul, South Korea, pp. 121 – 124, 2010.

[17] T.F. Wang, B.S. Ye, Y.W. Li, and Y. Yang, "Family gene based cloud trust model," in *International Conference on Educational and Network Technology (ICENT)*, Qinhuangdao, China, pp. 540 – 544, 2010.

[18] K. Xiong, and H. Perros, "Service performance and analysis in cloud computing," in *World Conference on Services - I*, Los Angeles, CA, USA, pp. 693 – 700, 2009.

[19] N.J. Gunther, *Analyzing computer system performance with Perl:PDQ*. Berlin Heidelberg, Germany: Springer-Verlag, 2005.

[20] M. Marzolla, "The qnetworks toolbox: a software package for queueing networks analysis," in *17th International Conference on Analytical and Stochastic Modeling Techniques and Applications (ASMTA 2010)*, Cardiff, UK, pp. 102-116, 2010.

THE ENERGY LABEL A NEED TO NETWORKS AND DEVICES

*Virgilio Puglia
Italtel, Italy*

ABSTRACT

The energy consumption of Information and Communication Technologies (ICTs) is today relevant also compared with the other industries. The evolution of ICT will determine enormous improvements in our daily lives, but will also increase energy need. So energy consumption becomes one of the key aspects in the evolution. Today every vendor manifests a trend to improve energy saving, but the lack of agreed indexes, terms, definitions and procedures makes it difficult for the customer to realize a relevant comparison. This paper presents a holistic approach in order to identify energy key indexes in all the Life-Cycle Assessment (LCA) of devices and networks. This study starts with the analysis of what was already implemented in other sectors and it proposes a devices energy label and a network energetic classification. A case study on the proposed method application is described. The results of the study demonstrates the need to adopt regulatory energetic indexes in order to ensure competition and energy saving.

Keywords—Energy Cost, Life-Cycle Assessment, Data Centers, Embodied energy, Energetic Classification.

1. INTRODUCTION

The climatic change and global warming are relevant and complicated problems that involve the entire world. This was demonstrated over a decade ago by the United Nations Framework Convention on Climate Change (UNFCCC) [1] and later by the Kyoto Protocols [2]. The Information and Communication Technologies (ICTs) can help to address this problem [3] and ITU is one of the most important stakeholders in this aspect as discussed during the 16th edition of COP conference [4].

ITU in the guideline: “ICTs for e-Environment” [5] shows a complete overview of the impact that ICTs have on the environment and climate change as well as their role in helping to mitigate and adapt to these changes. While in the report “NGNs and Energy efficiency” [6] is shown the evolution of the networks and the beneficial effect that will allow to reduce global CO₂ emissions by 15 per cent by 2020.

The energy consumption of the telecommunication sector is relevant also if compared with other industries. This aspect is known since 2007 when the Australian Computer Society demonstrated that the ICT, used by Australian businesses, generated 1.52% of the total national carbon dioxide emissions [8]. This value was close to that of the civil aviation and the metal production Industries.

The energy consumption, in the USA and worldwide, has been estimated respectively in 9.4 % and 5.3 % of the total electricity [5] produced, also if this data is quite approximated, it gives a consumption reference amount.

This paper won't discuss the enormous benefit to environment derived by ICT use [7], but the analysis will focus on optimizing the consumption for the Carrier Telecommunication Network itself. This consumption can be classified in the following four areas: Cooling, Telecommunication, ICT, Offices.

For example, in 2010 the percentage of energy consumption per area of the incumbent Italian operator [10] was the following: Cooling 25%, Telecommunication 58,63%, ICT 7,27%, Offices 9,09%;

This paper presents a new approach aimed to give a holistic systemic vision of network electrical consumption and describes key factors that will contribute to develop representative indexes of Network Environment compatibility. These indexes will allow an immediate perception of the relationship between the network technology and the environment impact. This will facilitate the electrical power consumption improvement, developing an environment sensibility that will overtake today's approach to subsystem vision.

The remainder of this paper is organized as follows. Section 2 summarizes current situation in other ICT areas (Data Centers, Residential and Office environments) not covered by this study but relevant in energy consumption. Also it introduces basic and necessary energetic concepts derived by other sectors. Section 3 describes Telecommunication Life-Cycle Assessment. Section 4 explains the need of international standards. Section 5 proposes the Holistic approach for Carriers with energetic indexes and explains the need of energetic class definitions. Section 6 presents a case study. Finally, Section 7 contains the conclusions and final remarks.

2. BACKGROUND

Internet's electricity use in the world can be classified in four areas [5]: Data Centers (includes cooling) 12,95%, PCs & Monitors 67,7%, Modems/routers/etc. 19,23%, Phone networks 0,12%.

Even if this paper is focalized on the Carrier Networks and Carrier devices, it will briefly describe also the two environment (Residential and Office equipments – Data Centers) that determine the highest need of electricity and its elevate environment impact. Also to better explain the Life-Cycle Assessment (LCA) it will introduce the Embodied Energy concept which is at the base of the relation device-energy.

2.1. Residential and office equipments

The electrical consumption of residential and office equipments, although higher in quantity, determines a relatively lower impact on the environment because during winter it constitutes an *Internal heat gains* [11] of the building. The devices are present in the working places or at home and this source leads to increases in internal temperatures and reduces the consumption due to the heat produced by the heating devices. In the energetic balance, the electrical consumption determines a reduction of fuel used by heating devices (recoverable system losses). This gain, according to Standard UNI/TS 11300-1 [12] and EN ISO 13790 [13], is calculated with the following equation:

$$[1] \quad Q_{H,nd} = Q_{H,ht} - \eta_{H,gn} \times Q_{gn};$$

Where:

$Q_{H,nd}$ is the energy needs for space heating;

$Q_{H,ht}$ is the total heat transfer by transmission and ventilation of the building;

Q_{gn} are the total solar and internal heat gains of the building;

$\eta_{H,gn}$ is the gain utilization factor.

More details about the heat gain from PCs and other office equipments are documented in [11]. The trend is to reduce the power need of this devices for example by voluntary program like 80 PLUS [9] or the US Energy Star qualification started in the US [14] and then adopted also by the European Union [15]. In this area the road is already drawn.

2.2. Data Centers

The Data Centers significantly contributes to power consumption [5] but for this area is already a consolidated practice to consider it as a complete system. So it's easier to define global indexes to evaluate all the buildings, facilities and rooms which contain enterprise servers, server communication equipments, cooling equipments and power

equipments. See for example the PUE index [16] or the new proposed index DPPE [17]. Google is one of the leader companies in optimizing datacenter that makes public Benchmark Data [18]. Other organizations like The European Union implement study and promote voluntary program to improve Datacenter, with a code of conduct [19]. Also in this area the road is already drawn.

2.3. Embodied energy

The Embodied energy is the total energy expended to make any product, bring it to the market, and dispose of it. This includes all activities which contribute, both directly and indirectly, to the construction process. This concept is used in more sectors like civil Construction [20], [21] and constitutes a good parameter to evaluate the environment impact of a manufacture. The optimization of the Embodied Energy must be the target of all the study. But these concepts are today difficult to implement in a world globalized where there is a high mobility of tangible between countries with different laws and rules. That makes it difficult to trace.

3. DEVICES LIFE-CYCLE ASSESSMENT (LCA)

The Life-Cycle Assessment (LCA) of Network devices can be divided into three parts corresponding to the three life phases of the product: manufacturing with Rise Telecommunication Energy (RTE), use with Dispose Energy (DE) and disposal at the end of life with the Final Energy (FE).

Each part depends directly by the Vendor that influences it with your device design, manufacture process and modality of transport. The Carriers perceive it by financial aspects. Specifically the RTE influence the Capital expenditures (CAPEX), the DE and FE influence the operational expenditure (OPEX), see Figure 1.

3.1. Rise Telecommunication Energy

The RTE is the total energy expended to make the device, and bring it to the end user. This is a portion of the Embody Energy.

A correct evaluation of RTE is complicated because the vendors try not to make public these sensible data and in some countries there is no transparency. Some authors like Tailor, in other sectors, use the device prices as a measure [20]. But this information can misrepresent the RTE because the devices prices are usually influenced by the business opportunity. Usually the final price to the carrier is determined by a market that impose a final price lower than the General Price List (GPL). The GPL is also doped by the market and cannot be considered as a reference.

3.2. Dispose Energy

The DE is the energy used by the device to perform its work. The real energy consumption depends on the devices used and your equipment. More interfaces and boards are equipped in a device and more is the required energy. This energy depends also if the interfaces are used. A punctual energy evaluation can be very complicated, and is quite not relevant with our target: have a macro index to give an immediate evaluation and push the Vendor to produce devices with low consumption.

3.3. Final Energy

The FE is the energy that will be wasted at the device's end of life. This energy can be high in case it is necessary to implement a process of raw materials reuse not provided during the project phase. The project must take into account the dismantling and recovery for reuse and recycling. In the European Community the directive WEEE [22] together with the directive RoHS [23] defines the recycling and recovery targets for all types of electrical goods.

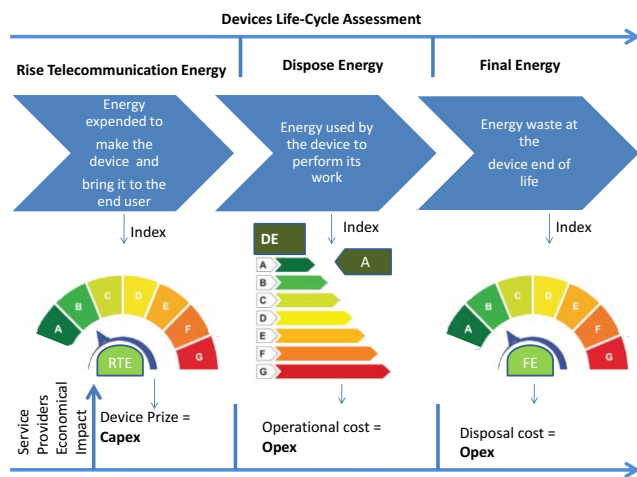


Figure 1. Devices Life-Cycle Assessment (LCA).

4. STANDARD NEED

The lack of agreed indexes, terms, definitions and procedures makes it difficult to analyze the following aspects: evaluate energy consumption (RTE, DE and FE), compare minimum energy performance requirements by networks, understand and compare data gathered on measured energy use of networks, have an immediate feeling of network efficiency.

For this reason, a set of International Standards is needed for assessment and calculation, rating and labeling, and standards for best practice and improvement of energy performance in networks. Such standards would enable meaningful comparisons of actual energy use, and the potential for energy saving.

The organization that should provide this, in our opinion, is ITU both for their attention to the climatic change and for

their authority. In the last four decades this organization defined many of the Telecommunication standards.

5. HOLISTIC APPROACH FOR CARRIER

When a system is constituted by many subsystems, individual optimization of each subsystem usually does not correspond to the system optimization. Vice versa an optimized system can often have some subsystems not optimized. The Carrier networks are a composite system constituted by many subsystems (like DWDM Backbone, IP Backbone, ATM backbone, STP network, etc). In the majority of carriers, all this subsystems are considered and optimized stand alone. Usually there are different departments in the carrier company that work on the different areas without sufficient interaction. Often the target of every department is to optimize our expenditure.

This mode of operating has an impact also on the energy consumption. For this reason, the achievement of optimization requires a holistic approach.

In the last 20 years the telecommunication evolved faster than other sectors (household appliances and build) and this doesn't allow to have enough time to metabolize a holistic vision's approach. In order to speed up the slow evolution of thought, it is important to build upon the experiences in other sectors.

The proposed approach tries to implement in the Telecommunication sector the evolution of thought already realized in the household appliances and house sector.

At the beginning of the '90, the vendors of refrigerators defined a scale of energetic performance that allowed the consumers to compare different models and vendors. The target of these indexes was to optimize the single appliance. This successful initiative encouraged the definition of similar scales also for others household appliances. A similar evolution starts for building in the '90, multiple subsystems with your stand alone indexes: the heating with your performance, the building subsystem with your transmittance, the other cooling with other performances. The start of the 2000s saw the beginning of a new thought that in a few years unified the single subsystems (build, house appliance, cold devices, illumination, house appliances) in a unique system (holistic approach) with the definition of a unique energetic index. See for examples [11], [12], [13].

5.1. Network Energy Indexes

The Telecommunication Networks are realized by multiple devices. These devices can be energetically characterized by the three parameters already presented: RTE, DE and FE. The Indexes associated to the network in each of the three parts that constitute the LCA is shown in the following formulas, where TOT is the total number of Sub-networks present in the network:

$$[2] RTE(Net) = \sum_{i=1}^{TOT} RTE(Sub.net i);$$

$$[3] DE(Net) = \sum_{i=1}^{TOT} DE(Sub.net i);$$

$$[4] FE(Net) = \sum_{i=1}^{TOT} FE(Sub.net i).$$

This study proposes to estimate these three indexes by taking in consideration three factors: the devices maximum energy consumption or energy cost, the maximum capability to transmit information and the distance covered by the information. These three factors must be evaluated in defined standard conditions that do not depend upon temporary condition of work and allow an independent evaluation use.

The total network index will be evaluated with the following formula:

$$[5] E(TOT) = RTE(Net) + DE(net) + FE(net).$$

All these indexes will have as unit of measurement Wh²/Gbit.

5.2. RTE evaluation method

It is proposed to estimate RTE with the following formula:

$$[6] RTE(device) = \frac{\sum_{i=1}^n M(i)*W(i)}{(\sum_{i=1}^{TOT} (Gbit_i*3600*R_i*\log_{10}(Km_i)))}$$

Where “M” is the Specific energy cost of material “i”, “W” is the quantity expressed by the weight of the material “i” and “n” is the total number of these materials. An example of M evaluation, even if derived from the building sector, is available in [24]. We hope that an international organization for standardization like ITU will implement reference documents, with energy cost/coefficients, of materials.

For those devices whose materials or quantities cannot be determined, it is proposed to disadvantage the vendor by evaluating the energy as the devices weight multiplied by the worst energetically material standardized.

This method will also make it more convenient for the vendor to provide the necessary data.

The denominator of this formula is explained in the paragraph 5.5.

The RTE of a sub network is derived by the sum of RTE per device.

$$[7] RTE(Sub.net) = \sum_{i=1}^n RTE(device i).$$

5.3. DE evaluation method

Today in the devices datasheet we can find only the electrical parameters necessary to make possible wiring and cooling dimensioning. These data don't allow to compare consumption between different vendors and it is complicated to understand the electrical advantage proposed by different devices/technology. It is proposed to estimate DE with the following formula:

$$[8] DE(Device) = \frac{Energy_{max}}{\sum_{i=1}^{TOT} (Gbit_i*3600*R_i*\log_{10}(Km_i))}$$

Energy_{max} is the max energy consumption for each device expressed in KWh.

The denominator of this formula is explained in the paragraph 5.5.

The DE of subnetwork is derived by the sum of DE per device:

$$[9] DE(Sub.net) = \sum_{i=1}^n DE(device)_i.$$

5.4. FE evaluation method

The FE depends to the facility to separate raw materials that can be reused in a new process.

If a material requires an expensive process to be separated from the device, it will not be removed and so it will be discarded.

It is proposed to evaluate the recycled raw material that can be separated from the devices in 1 hour by only mechanical man work. All the material that cannot be separated will be wasted, so this Energy will be lost in the global energy balance. The following formula indicates this concept:

$$[10] FE(device) = \frac{\sum_{i=1}^n M_i*Waste_i}{\sum_{i=1}^{TOT} (Gbit_i*3600*R_i*\log_{10}(Km_i))}$$

Where “M” is a convectional energy cost of material that will be wasted and “Waste” is the quantity, expressed by the weight, of the material discarded. About the conventional energy cost see [24] and what already explained in 5.2 paragraphs.

The denominator of this formula it is explained in the paragraph 5.5.

The FE of a sub network is derived by the sum of FE per device.

$$[11] FE(Sub.net) = \sum_{i=1}^n FE(device i).$$

5.5. Formulas contextualization

All the previous formulas must be contextualized with the technological typology in analysis.

For example in a DWDM Backbone

- “n” in formula [7], [9], [11] is the total DWDM devices (OLA-OADM-TS-OXC) constituting the network;
- “TOT” is the total interfaces equipped;
- “Km” is the span length, when it is major of 10 Km or 10 Km when minor;
- “Gbit” is the capacity of the interfaces “i” expressed in Gbit per second.

- “R” takes in consideration the Performance Transmission of the devices. This value is equal to 1, when the BER (Bit Error rate) is equal to a defined standard value (It is proposed for example 10^{-15}), and it is less than 1 when BER is lower.

For the IP Backbone it is possible to use the same formulas with the following contextualization.

The “Km” is the distance considered in the following two situations:

1. between two adjacent IP devices (router-switch-hub) not necessarily at the same OSI layer when the Backbone IP isn’t interconnected via a transport network;
2. Between IP devices and Transport devices (DWDM-SDH-ATM Switch) when the BB IP is interconnected via a transport network.

The distance will be computed only when it is more than 10 Km; this value represents the operating distance with 1000BASE-LX/LH interfaces and SMF (ITU-T G.652) [25], [26]. If the distance is less than 10 Km, the Km parameter will be assumed to 10. For the evaluation of the stand alone devices indexes it is proposed to consider a conventional distance (ex. 0,5 Km for SX, 10 Km for LX, 80 Km for ZX, etc) that needs to be defined by an international organism. For the Evaluation of networks (so devices in a real context), we will use the real distances. The “R” factor is necessary to make DE index independent from operating point. The interfaces of IP devices have usually relevant quantity of packets discarded when they work at 100% of the Bandwidth load. The percentage of packets discarded decreases with the Bandwidth Load reduction but it is strongly influenced by the packet size. To be independent from IP packet size we consider standard Ethernet interfaces used at 90% of Nominal Bandwidth. This load value is a good compromise to guarantee not to have packets discarded.

The number 3600 in the denominator is inserted to converts the Gbit/s in Gbit/h.

5.5.1. Information transmission trend with the distance

About the formula used to evaluate distances effect on the information transmission it is necessary to highlight the following aspects.

There are various physical impairments that affect signal and limit information transmission distances. Attenuation is the most important for all technologies. But there are others phenomena more complex, technology dependent. For example in DWDM technology there are impairments like: Scattering, Bending Losses, Polarization Mode Dispersion (PMD), Four Wave Mixing (FWM), etc. [27]. Resuming all this phenomena with a formula is very hard and not relevant to the scope of this study. In this paper is used the decadic logarithm of distance to approximate the information transmission trend with the distance. This approximation in physics dimensional analysis provides a nondimensionalization. The logarithm remove unit of length from the equations.

5.6. Energetic Class

For each typology of devices we propose to define a unique labeling, compliant with the European directive 92/75/EEC [28], that permits an intuitive classification into seven classes, similar to the devices tag shown in Figure 2.

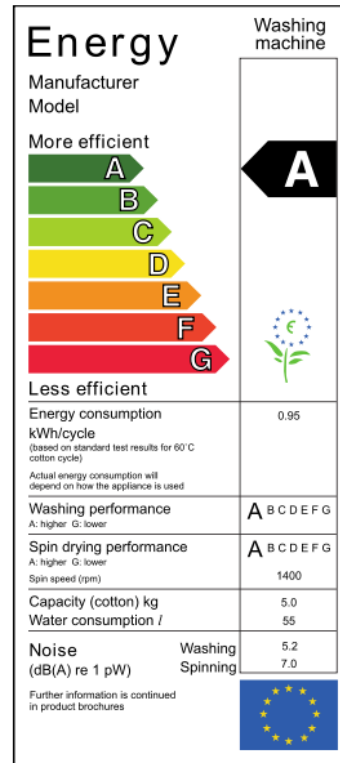


Figure 2. Example of labeling (European directive).

The three indexes RTE(device), DE(devices) and FE(device) could determine three different labels that indicate the energy efficiency of a single device.

5.7. RTE-DE-FE Scale

In this paper it isn’t indicated a specific value for the energetic scale. We think this indication must derive from a more appropriate evaluation also of the future technological evolution and it requires more study. The target of the energetic scale must be an improvement of the devices’ energy consumption. At the beginning, we propose:

1. that all the devices should be considered in the F class (worst class);
2. to define the limit between class F and G by a benchmark. We can identify the RTE, DE and FE values of the best devices, in term of energy, available on the market. These values minus 1 Wh²/Gbit will be considered the threshold between class F and G. This will stimulate the Vendors to improve and exceed the F class.

The other thresholds will be more complex to determine because we have to take in consideration the energy evolution in the next 30 years.

6. CASE STUDY

To explain the concepts described in the previous paragraphs, we show an application in a simple network constituted by two identical POPs (Firenze and Arezzo), each one is realized with a Cisco 7600 used as a BRAS (broadband remote access server). This Network use a Backbone DWDM realized by two Cisco ONS15454. The architecture is show in the following picture Figure 3.

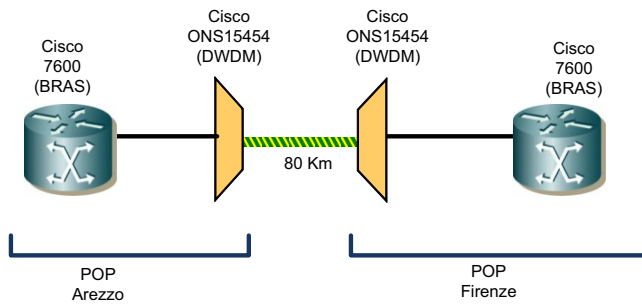


Figure 3. Case study Network.

For all the involved devices there is no material indication in the data sheet. In this case study in the RTE and FE formula will use the energy value of Aluminum virgin extruded-anodized (227 MJ/kg [24] equivalent to 63.05KWh/Kg) as “M” (conventional energy cost of material).

6.1. BRAS energy index

The following table shows the data of the BRAS (see [29], [30], [31], [32]).

In the first and second column of Table 1, we can find the device’s (BRAS) kit list (the card’s part number and its quantity), in the third the total throughput of the component, in the fourth the R factor (as before defined), in the fifth, the weight of the component and in the sixth the energy consumption.

Table 1. BRAS data.

Product Number	Q. ty	Cap.ty Gbs	R	Kg	W
7609S-SUP720BXL-R	1			0.0	0
CISCO7609-S	1			67.5	0
7600-SIP-400	2			14.5	370
SPA-1XOC12-ATM	2	1.244	0.95	0.9	41.0
SFP-OC12-IR1	2				
7600-ES+2TG3CXL	2				
XFP-10GLR-OC192SR	4	40	0.9	7.3	594
WS-SUP720-3BXL	2			10.4	656
PWR-6000-DC	2			32.0	2120
	T O T	41.2		132.6	3781

With M= 63.05KWh/Kg, “Km” for ATM interfaces = 15 Km and “Km” for 10Gbps interfaces = 10 Km.

$$[12] RTE(7600-BRAS)= 62.1114 Wh^2/Gbit;$$

$$[13] DE(7600-BRAS)=0.0281 Wh^2/Gbit.$$

To evaluate the FE it is assumed to recuperate only the chassis (67.46 Kg) so the weight of the wasted material is 65.10 kg. M=63.05KWh/Kg and the result is:

$$[14] FE(device)= 30.4936 Wh^2/Gbit.$$

6.2. DWDM energy index

The following table shows the data of the Cisco ONS 15454 (see [34], [35], [36], [37]). In the first and second column of Table 2, we can find the device’s (ONS15454) kit list (the card’s part number and its quantity), in the third the total throughput of the component, in the fourth the R factor (as before defined), in the fifth the weight of the component and in the sixth the energy consumption.

Table 2. Cisco ONS15454 data.

Product Number	Q. ty	Cap.ty Gbs	R	Kg	W
15454E-CC-FTA=	1				
15454E-SA-ETSI=	1				
15454E-BLANK-FMEC=	10				
15454-FBR-STRG=	1				
15454E-BLANK=	5				
15454E-PWRCBL-010=	2			26	115
15454E-AIR-RAMP=	2			49.9	230
15454E-TCC2P-K9=	2			1.4	52
15454E-CTP-MIC48V=	1			0.2	0.38
15454E-AP-MIC48V=	1			0.2	0.13
15216-DCU-SA=	2			4.6	0
15216-DCU-1150=	2			7	0
15216-DCU-450=	2			7	0
15454-OSCM=	2	0.3	1	3.4	54
15454-OPT-AMP-17C=	2			3.8	46
15216-MD-40-ODD=	2			9.4	70
15454-10E-L1-C=	4	40	1	5.2	200
ONS-XC-10G-S1=	4				
15454-40-SMR1-C=	2			6.4	120
	T O T	40.3		124.5	887.51

Where:

$$M=63.05KWh/Kg \text{ and } Km =80, BER= 10^{-15}.$$

$$[15] RTE(ONS 15454)=28.4307 Wh^2/Gbit;$$

$$[16] DE(ONS 15454)= 0.0032 Wh^2/Gbit.$$

To evaluate the FE, it is assumed to recuperate only the chassis (26 Kg) so the weight of the wasted material is 98.50 kg, M=63.05 KWh/Kg. The result is:

$$[17] FE(ONS 15454)= 22.4933 Wh^2/Gbit.$$

6.3. Network energy index

The Network Energy index will be evaluated by using the formula [2], [3], [4]. In this case study, for simplicity, it is assumed that the real distance is equal to conventional distances.

- [18] RTE(Net)= 181.0843 Wh²/Gbit;
- [19] DE(Net)= 0.0626 Wh²/Gbit;
- [20] FE(Net)= 105.9740 Wh²/Gbit.

By the formula [5] the result is:

- [21] E(TOT)= 287.1209 Wh²/Gbit.

6.4. Energy index result

The result of RTE and FE is altered by the lack of information in the data sheet; nevertheless this demonstrates the validity of the method that will incentive the vendor to be more careful. The formula [5] even if correct theoretically, at the moment is inappropriate because the selected approximation to FE determines a parameter not congruent with the others. The sum risks to compromise the global index.

7. CONCLUSIONS

In this paper it is presented a new approach that aims to provide simple and useful Devices Energetic Classification both to devices and networks. This will push the Vendor to realize devices with low consumption and help Carriers and Providers to have cost savings. With this proposed method, it will be possible to compare energy needs for different devices and solutions also from different vendors. This paper shows the necessity to create for the Telecommunication sector a group of agreed energetically indexes, terms, definitions and procedures to favor the global CO₂ emissions reduction and cover the gap already covered in other technological sectors. We hope this study will be a stimulus for international organization for standardization, like ITU, to provide: the Telecommunication energetic scale, the reference documents with energy cost/coefficients of materials and common energetic indexes, terms, definitions and procedures. The gap to fill in is so big that it will be necessary further research in the above mentioned subject area.

REFERENCES

- [1] <http://unfccc.int/2860.php> ;
- [2] http://unfccc.int/kyoto_protocol/items/2830.php;
- [3] Communication from the commission, Action Plan for Energy Efficiency: Realising the Potential http://ec.europa.eu/energy/action_plan_energy_efficiency/doc/com_2006_0545_en.pdf ;
- [4] <http://www.itu.int/themes/climate/events/cop16/index.html>;
- [5] <http://www.itu.int/ITU-D/cyb/app/docs/itu-icts-for-e-environment.pdf>;
- [6] NGNs and Energy Efficiency, ITU-T Technology watch Report 7, August 2008;
- [7] Towards a High-Bandwidth, Low-Carbon Future http://www.climaterisk.com.au/wp-content/uploads/2007/CR_Telstra_ClimateReport.pdf ;
- [8] [Audit of Carbon Emissions resulting from ICT, Australian Computer Society, August 2007](#);
- [9] <http://www.plugloadsolutions.com/80PlusPowerSupplies.aspx> ;
- [10] EC Data Centre Codes of Conduct - Stakeholder meeting, London 10 November 2010, IT energy efficiency & telcos, Flavio Cucchiatti http://re.jrc.ec.europa.eu/energyefficiency/pdf/CoC/DC_Stakeholder%20Meeting.%20London%2010%20November%202010/Telecom_London%20Nov%202010.pdf
- [11] “Environmental design CIBSE guide A”, CIBSE, 2006, ISBN-10: 1-903287-66-9, ISBN-13: 978-1-903287-66-8;
- [12] UNI/TS 11300-1, “Energy performance of buildings, Part 1: Evaluation of energy need for space heating and cooling”, may 2008;
- [13] EN ISO 13790:2008, “Energy performance of buildings -- Calculation of energy use for space heating and cooling;
- [14] Energy Star Office Equipment Product Specification, Attachment A, www.energystar.gov;
- [15] 2003/269/EC, *Official Journal of the European Union*, L99/47, L99/48, 17.04.2003;
- [16] The Green Grid, 2007, “The Green Grid Data Center Power Efficiency Metrics: PUE and DCiE,” Technical Committee White Paper;
- [17] http://re.jrc.ec.europa.eu/energyefficiency/pdf/CoC/DC_Stakeholder%20Meeting.%20London%2010%20November%202010/GIPC%20London%20Nov%202010.pdf ;
- [18] <http://www.google.com/corporate/datacenter/efficiency-measurements.html> ;
- [19] http://re.jrc.ec.europa.eu/energyefficiency/html/standby_initiative_data_centers.htm;
- [20] “Hybrid Life-Cycle Inventory for Road Construction and Use”, Graham J. Treloar, Peter E. D. Love, and Robert H. Crawford; JOURNAL OF CONSTRUCTION ENGINEERING AND MANAGEMENT ASCE, DOI: 10.1061/(ASCE)0733-9364(2004)130:1(43);
- [21] “The Environmental Impacts of Residential Development: Case Studies of 12 Estates in Sydney”, Bill Randolph, Darren Holloway, July 2006 (Updated March 2007);
- [22] European Community Directive 2002/96/EC on waste electrical and electronic equipment (WEEE). <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2003:037:0024:0038:EN:PDF> ;

- [23] European Community Directive 2002/95/EC on Restriction of Hazardous Substances Directive (RoHS);
- [24] “The energy embodied in building materials updated New Zealand coefficients and their significance”, George Baird, Andrew Alcorn, Phil Haslam, *IPENZ Transactions*, Vol. 24, No. 1/CE, 1997;
- [25] IEEE 802.3z standard;
- [26] http://www.cisco.com/en/US/prod/collateral/modules/ps5455/ps6577/product_data_sheet0900aecd8033f885.html;
- [27] A. Gumaste, T. Antony, “DWDM Network Design and Engineering Solutions”, Cisco Press, 2002.
- [28] Council Directive 92/75/EEC of 22 September 1992 on the indication by labelling and standard product information of the consumption of energy and other resources by household appliances;
- [29] http://www.cisco.com/en/US/partner/docs/routers/7600/Hardware/Chassis_Installation/7600_Series_Router_Installation_Guide/cis_76xx.html;
- [30] http://www.cisco.com/en/US/partner/docs/routers/7600/Hardware/Module_and_Line_Card_Installation_Guides/ES40_Line_Card_Installation_Guide/es40_chap2.html#wp1174038;
- [31] http://www.cisco.com/en/US/partner/prod/collateral/routers/ps368/data_sheet_c78-49152.html;
- [32] http://www.cisco.com/en/US/partner/prod/collateral/routers/ps368/product_data_sheet0900aecd8027c9e6_ps708_Products_Data_Sheet.html;
- [33] http://www.cisco.com/en/US/partner/prod/collateral/modules/ps6267/product_data_sheet0900aecd804dc62d.html;
- [34] http://www.cisco.com/en/US/partner/docs/optical/15000r8_0/15454/sonet/procedure/guide/r80procd.html;
- [35] http://www.cisco.com/en/US/partner/prod/collateral/optical/ps5724/ps2006/product_data_sheet0900aecd800e4d24.html;
- [36] http://www.cisco.com/en/US/partner/prod/collateral/modules/ps2831/ps6085/product_data_sheet0900aecd80351fd6_ps2006_Products_Data_Sheet.html;
- [37] http://www.cisco.com/en/US/partner/prod/collateral/optical/ps5724/ps2006/ps5320/product_data_sheet0900aecd803fc3e8.html;

A DISTRIBUTED MOBILITY MANAGEMENT SCHEME FOR FUTURE NETWORKS

Ved P. Kafle, Yasunaga Kobari and Masugi Inoue

National Institute of Information and Communications Technology, Tokyo, Japan
{kafle,kobari,inoue}@nict.go.jp

ABSTRACT

Unlike Mobile IP protocols, which specify centralized mobility management schemes, the future network is envisioned to embrace distributed approaches to mobility so that it can avoid a single point of failure and triangular routing problems as well as optimize the handover process. This report presents a distributed mobility management scheme of HIMALIS (Heterogeneity Inclusion and Mobility Adaptation through Locator ID Separation) architecture where mobility signaling takes place through the control network composed of end hosts and dedicated functional nodes in the network. Moreover, the signaling functions are network layer independent; therefore, the proposed scheme can be applied to a future networking environment where multiple protocols do coexist in the network layer. It discusses the architectural components, mobility protocol, and an analysis of performance results obtained from an emulation system.

Keywords— Future network, new generation network, ID/locator split, distributed mobility management

1. INTRODUCTION

To address the mobility problem of the Internet, the Internet Engineering Task Force (IETF) has developed Mobile IPv4 [1] and Mobile IPv6 [2] protocols that use a centralized mobility anchor point called the Home Agent (HA). The mobile hosts (MHs) detect their own mobility and initiate signaling to update their new location information in the home agent. These protocols involve longer signaling path between MHs and HAs, thus have longer handover delay. To complement these protocols with local optimization, Fast Handover for Mobile IP [3] and Hierarchical Mobile IP [4] have been developed. Similarly there is also a network-based mobility management protocols called Proxy Mobile IPv6 [5], which does not require MHs to possess any mobility control functions. The access routers called Mobile Access Gateways (MAG) detect host mobility and carry out location update signaling with the Local Mobility Anchor (LMA).

These mobility management protocols require data packets to be tunneled through a fixed node such as the HA and the LMA. This requirement makes the protocols vulnerable to a single point failure as well as sub-optimal routing. These problems are avoided in our newly proposed HIMALIS (Heterogeneity Inclusion and Mobility Adaptation through Locator ID Separation) architecture [6], where both hosts

and dedicated network nodes known as local name server (LNS) and ID registry (IDR) perform mobility-related control functions distributedly. This architecture is based on the ID/locator split concept [7], which is being currently discussed in ITU-T Study Group 13 [8]. The mobility functions are implemented in the identity layer, a new layer inserted between the transport and network layers. These functions can be implemented over any (or heterogeneous) network layer protocols, such as IPv6, IPv4, or new protocols introduced in the future networks. This will promote the coexistence of multiple protocols in the future networks. The preliminary concept of HIMALIS mobility management was published in [9]. This paper presents the realization of the concept by the detail design of mobility functions and performance analysis by implementing them in an emulation system.

The remainder of this paper is organized as follows. Section 2 briefly describes the mobility related components of the HIMALIS architecture. Section 3 discusses the mobility signaling process. Section 4 presents an overview of the emulation system developed for performance evaluation in distributed environments. Section 5 concludes the paper.

2. MOBILITY COMPONENTS

2.1. Naming Method

In HIMALIS architecture [6], a global hostname (i.e. the host's globally unique name) is composed of its local hostname (which is unique only in the local domain) and a globally unique domain name. The global hostname is formed by concatenating the local hostname and domain name using concatenation symbol #. The global hostname is thus in the format similar to mypc#mydomain.org or sen01#adomain.com, where mypc and sen01 are local hostnames and mydomain.org and adomain.com are domain names. A host ID is then generated by concatenating prefix, scope and version fields with the cryptographic hash value of the global hostname and a parameter, i.e., host ID = concatenation (prefix, scope, version, hash(global hostname, parameter)). The host's public key can be used as the parameter.

A hostname and a host ID play the similar role that both identify the host. The differences between them lie in their structures and usages. Hostnames are usually denoted by variable-length, human readable and memorable alphanumeric characters, while IDs are denoted by fixed-length bit strings which are not human-friendly to memorize. Hostnames are used during a communication initialization

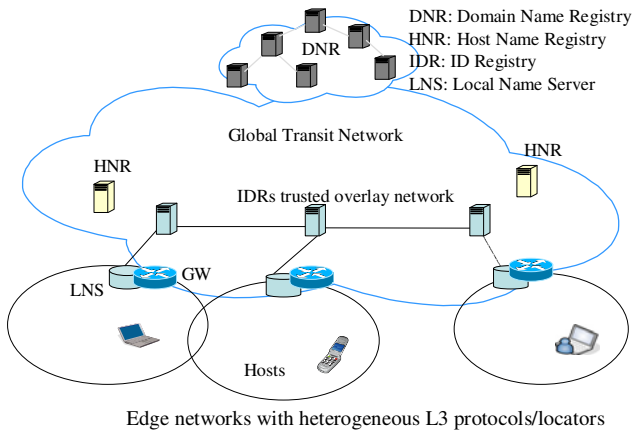


Figure 1. HIMALIS architectural components

process to find locators and authenticate and authorize hosts, while host IDs are used in communication protocols and packet headers to identify sessions or packets.

The bindings among hostnames, IDs, locators and other information (such as security keys) are stored in (and provided through) a resolution system consisting of two registries: Domain Name Registry (DNR) and Host Name Registry (HNR).

The DNR is similar to the conventional DNS (Domain Name System). It stores the bindings between a domain name and the ID/locator of HNR which, in turn, stores information about hosts that have the domain name in their global hostnames. The DNR records are static, i.e. they do not change often, as HNRs are considered to be fixed nodes which do not change their IDs and locators.

The HNR stores the hostnames, IDs, locators and security information such as public keys or shared keys of the hosts

that derive their global hostnames from the domain names managed by the HNR. The HNR records are dynamic as they often need to be updated when hosts change their locators due to mobility.

2.2. Architectural Components

Fig. 1 shows the HIMALIS architecture components, which are the edge (or access) networks, global transit network, gateways (GW), and control nodes such as DNR, HNR, LNS and IDR. The edge network can use heterogeneous layer 3 (L3) local protocols and locators, while the transit network use a global L3 protocol. The GWs connecting edge networks to the transit network possess ID/locator mapping functions to translate the L3 protocol information of the packets flowing through them. The DNR and HNR and LNS are components of the hostname to ID/locator resolution system. The DNR and HNR are located in the global transit network while the LNS is located in the edge network. Each edge network contains at least one LNS. The LNS function can exist in an independent node or be collocated with the GW. In this paper, we consider the latter case. Besides hostname to ID/locator resolution, the LNS is also responsible for hostname registration in the HNR when the host comes into existence in the network for the first time, host authentication and locator assignment when the host attaches to the edge network, and mobility signaling when the host moves from one edge network to another. The mobility signaling takes place through the IDRs overlay in the global transit network. The IDRs form a trusted overlay network, which is used to distribute the mobile host's ID/locator binding updates among LNSs during mobility.

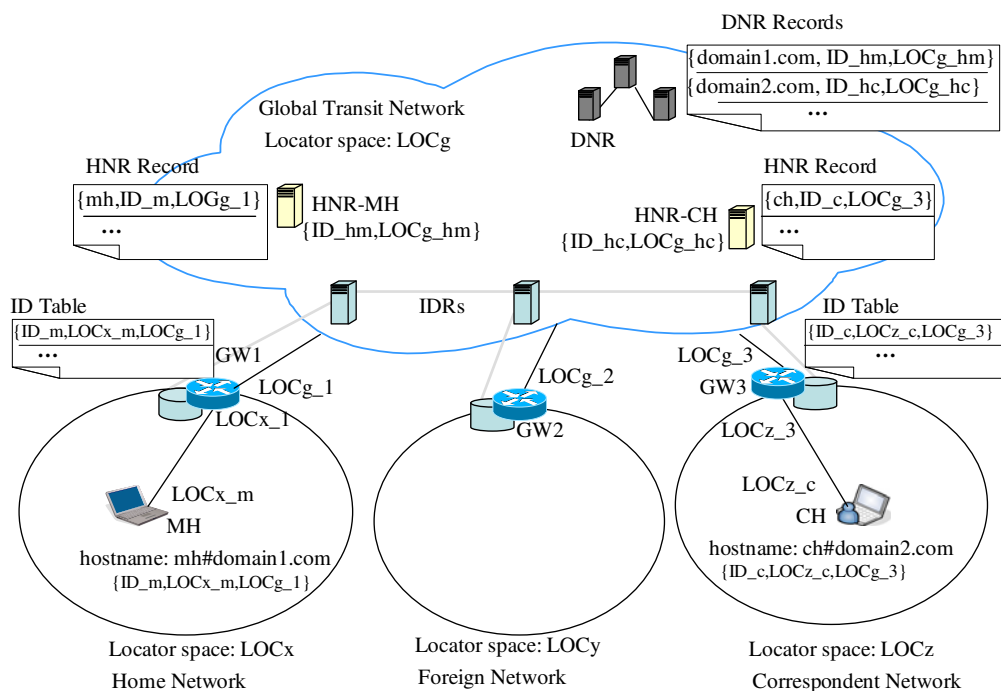


Figure 2. Heterogeneous network layout

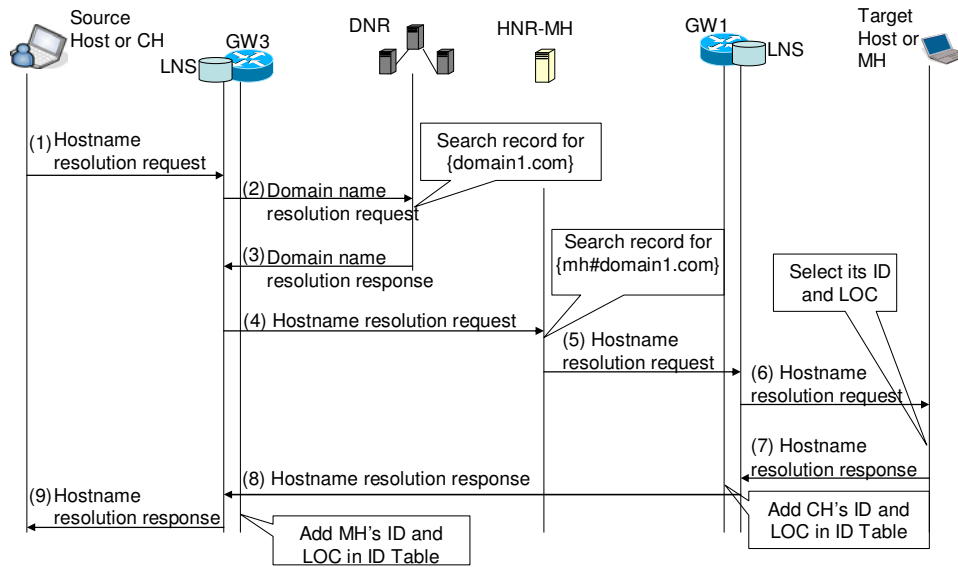


Figure 3. Hostname resolution in heterogeneous networks

2.3. Heterogeneous Network Model

Fig. 2 illustrates the heterogeneous network layout used to describe the HIMALIS mobility management scheme. There are three edge networks connected to the global transit network through GW1, GW2, and GW3. All of these networks use different L3 protocols and locator spaces. LOC_g represents the locator space used in the global transit network, while LOC_x, LOC_y, and LOC_z represent the locator spaces used in the mobile host (MH)'s home network, foreign network and the correspondent host (CH)'s network, respectively. The MH's hostname, ID, and local locator are mh#domain1.com, ID_m, and LOC_{x_m}, respectively. Similarly, ch#domain2.com, ID_c, and LOC_{z_c} are the CH's. Besides local locators, the MH and CH also have global locators LOC_{g_1} and LOC_{g_3}, which are in fact GW1's and GW3's locators derived from the global locator space. The GWs also have local locators from the edge networks' locator spaces. LOC_{x_1}, LOC_{y_2} and LOC_{z_3} are the local locators of GW1, GW2, and GW3, respectively.

The binding among the MH's hostname, ID and global locator is stored in HNR-MH whose ID and locator are ID_{hm} and LOC_{g_hm}. Similarly, the binding among the CH's hostname, ID and global locator is stored in HNR-CH whose ID and locator are ID_{mc} and LOC_{g_hc}. The bindings among the MH's and CH's domain names and HNR-MH's and HNR-CH's IDs and locators are stored in the DNR.

The GW maintains an ID table to store the bindings between IDs and locators of all hosts associated with the edge network. For example, GW1 stores the MH's ID and locator in its ID table after the MH gets attached with the home network. The network attachment process requires the host to present its credential to the LNS, which is the responsible authentication agent of the edge network. After verifying the host's credential from the host's HNR, the

LNS allows the host to configure a locator from the local locator space and informs about the global locator. The LNS keeps a record for hostname, ID and locator bindings of all hosts currently attached to the edge network. It also provides the ID/locator bindings to the GW, which caches them in the ID table. Along with the local host's ID/locator bindings, the LNS also provides the remote hosts' ID/locator bindings to the GW as they are obtained through the hostname resolution process discussed below. The GW's ID tables are updated by the LNS for mobility. Similarly, the host also maintains an ID table to store its peers ID/locator bindings (not shown in the figure).

Note that in HIMALIS architecture, all signaling messages (e.g. host registration, authentication, hostname resolution, mobility, etc.) flow through the LNS. The signaling packets are distinguished from application data packets by the value of "packet type" field in the identity header. The GW checks the value of this field of all packets entering into the edge network from the global transit network, and forwards them either to the LNS or to the destination host

2.4. Hostname Resolution Process

In HIMALIS architecture, a host (say MH) is known only by its global hostname to another host (say CH) before they have established a communication session. To start a session, the CH initiates a hostname resolution process to obtain each others hostname to ID/locator bindings. The hostname resolution process is shown in Fig. 3 and summarized below.

The CH sends a hostname resolution request message to the LNS. The message contains the CH's hostname, ID, local and global locators, as well as the target MH's hostname and intended application type. The host caches the request message in the hostname resolution cache, which is used to temporarily store the content of hostname resolution requests that are under progress. The LNS also stores the message in

its hostname resolution cache and checks its domain name resolution cache to find if the MH's domain name has recently been resolved. The domain name resolution cache stores the target domain name, HNR's ID, locator, and validity (TTL). If no entry for the MH's domain name is found, the LNS sends a domain name resolution request to the DNR to obtain the HNR-MH's ID and locator in the domain name resolution response message and then adds these values to the domain name resolution cache. After resolving the MH's domain name to HNR-MH's ID and locator, the LNS sends the hostname resolution request message to the HNR-MH, which checks its record to find the MH's ID and locator, and then relays the message the MH. The message reaches the MH's GW, i.e. GW1, which forwards it to the LNS. The LNS adds GW3's local locator as the CH's local locator (as would be seen by the MH) to the message content and sends it to the MH. The MH then selects its locator (in case it is multihomed) that would be best suitable for the requested application type and stores the CH's hostname, ID and locator binding in its ID table, and prepares a host name resolution response message containing its hostname, ID and global locator, as well as the CH's hostname, ID and global locator. The message is

then sent to the LNS (GW1), which adds the CH's ID and global locator to its record as well as to the ID table of GW1 as the peer host of the MH. The LNS (GW1) forwards the message to the LNS (GW3). On receiving the message, the LNS (GW3) adds the MH's ID and global locator to its record as well as to the ID table of GW3 as the peer host of the CH. The LNS (GW3) adds GW3's local locator to the message as the MH's local locator before forwarding it to the CH. The CH adds the MH's ID/locator binding to its ID table.

After the CH and MH have each others IDs and locators through the hostname resolution process, they use the IDs in the identity header and locators in the network header of data packets. The IDs present in the identity header do not change as the packets traverse through the networks and are used by the GWs to translate locators (and network layer protocols) in the network header of packets when they pass through the GW. Both the host and GW consult their ID tables to find an appropriate locator related to the given ID.

3. MOBILITY MANAGEMENT SCHEME

Mobility management involves mainly the following two

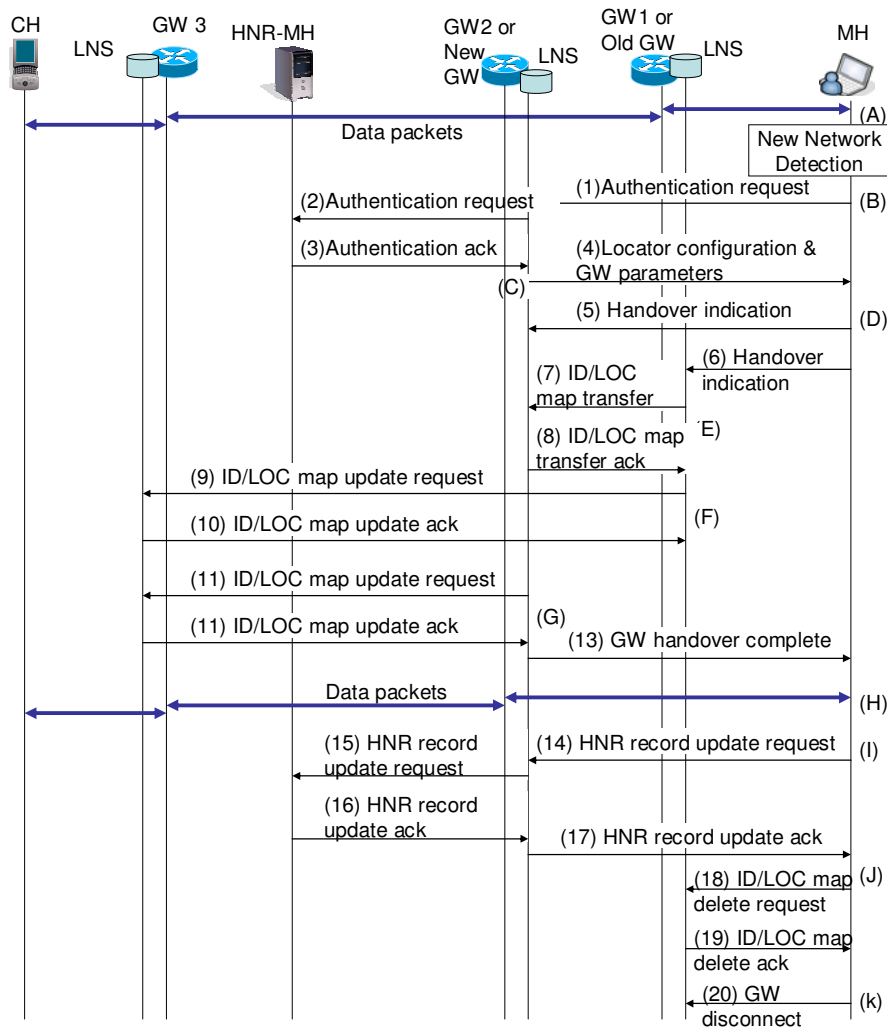


Figure 4. Make-before-break handover signaling sequence

categories of functions: (a) detecting a new network, authenticating to it and getting a new locator from it; and (b) moving ID/locator binding from the MH's old GW to new GW, updating the ID table in the CH's GW, and updating the HNR record.

3.1. Make-Before-Break Handover

The handover process can be of make-before-break type if the MH can start the process to obtain a new locator from the foreign network and perform update about its new locator in the CH's GW while still being connected to the home or old network. This type of handover is also called smooth, soft or proactive handover. It is necessary in this case that the old and new edge networks have some overlapped area from where the MH starts and completes the handover process. The signaling flow sequence for handover is shown in Fig. 4 and described briefly below.

(A) While having data communication with the CH via the home or old network, the MH detects the availability of the foreign or new network.

(B) For attachment with the foreign network, the MH presents its credential to the LNS (of GW2), which authenticates the MH by consulting the MH's HNR. (C) Upon authentication, the LNS sends a local locator (LLoc) (or information required to configure it) and other system parameters such as GW2's ID, LLoc and global locator (GLoc) to the MH.

(D) The MH sends a handover indication message to the old and new LNSs. The message sent to the old LNS contains the MH's hostname, ID, old LLoc, old GLoc as well as new GLoc. The message sent to the new LNS contains the MH's hostname, ID, old GLoc as well as new LLoc and GLoc. The message also includes a flag to indicate if the new LNS has to contact the old-HNR to obtain MH's ID/locator map. The old and new LNS may send acks to the MH for the receipt of handover indication signaling (not shown in the figure.)

(E) After receiving the handover indication message, the old LNS transfers the MH's ID/locator bindings containing to the MH's new GLoc (i.e. to the new GW). The ID/locator bindings also include information about all CHs currently communicating with the MH. Alternatively, the old LNS may transfer the ID/locator mapping cache of mobile host to the new LNS only after receiving an ID/locator map transfer request from the later (useful when the old LNS cannot receive the handover indication message directly from the MH, such as in break-before-make handover discussed in the next subsection.)

(F) The old LNS sends an ID/locator map update request to the CH. The message includes: CH's hostname, ID, GLoc as well as MH's hostname, ID, old GLoc, and new GLoc. The CH's LNS stores the updates in a buffer and sets a flag on for the MH's entries in the record for indicating that the entries are about to be updated by new values stored in the buffer when the ID/locator map update request comes from the new LNS to confirm the updates. The CH's LNS replies the old LNS with an ack message. If no ack is received, the old LNS resends the ID/locator map update request.

(G) The new LNS also sends an ID/locator map update request to the CH's LNS to confirm the update. On receiving this message, the CH's LNS updates its record as well as GW3's ID table and sends an ack to the new LNS. On receiving the ack, the new LNS sends a GW handover complete message to the MH and MH's ID/locator bindings to GW2's ID table. (H) Data communication takes place through the new GW. (I) After receiving the GW handover complete message, the MH sends an HNR record update message through the new LNS. (J) The MH sends an ID/locator map delete request to the old LNS, which deletes the MH's entry from its record and the GW's ID table and sends back an ack to the MH. (K) The old LNS and MH exchange signaling messages for the graceful termination of the connection or link in the old network.

In this way, smooth handover completes and the communication session between the hosts continues through the new GW without any interruption. Although not shown in the figure, all the ID/locator binding update signaling messages are transferred through the trusted network of IDRs in the global transit network. The LNS sends (receives) these messages to (from) the IDR with which it has a pre-established security association.

3.2. Break-Before-Make Handover

In break-before-make handover, the MH gets disconnected from the old network (e.g. due to not having an overlapped coverage area of both networks) before it starts exchanging handover signaling messages with the new network. This type of handover is also called hard or reactive handover.

The sequence of control messages exchanged for a hard handover differs slightly from that of the smooth handover. As soon as the MH detects that it has been disconnected from the old network, it searches for a new network, authenticates itself, and acquires a new locator. It then sends a handover indication message to the new LNS by setting the flag value to request the later to contact the old LNS for transferring the MH's ID/locator maps. The new LNS accordingly requests the old LNS for the MH's ID/locator map entries. The old LNS transfers the ID/locator map to the new one. The old LNS may also command the old GW to buffer the MH's packet until the MH's new locator is known through the new LNS. After having the MH updated its HNR record, the new LNS sends an ID/locator map delete request to the old LNS, which deletes the entries of the MH from its record as well as from the GW's ID table. In this way hard handover completes and the communication session continues through the new GW.

4. PERFORMANCE ANALYSIS

4.1. Implementation Overview

In order to verify the feasibility of the HIMALIS architecture and protocols, we have implemented an emulation system in Linux user space. As shown in Fig. 5, the system has three components: module, platform, and controller. The modules implement architectural functional components such as the host, DNR, HNR, and GW (which

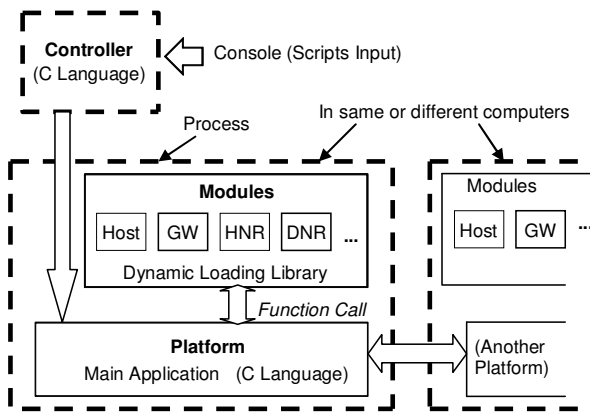


Figure 5. Implementation layout

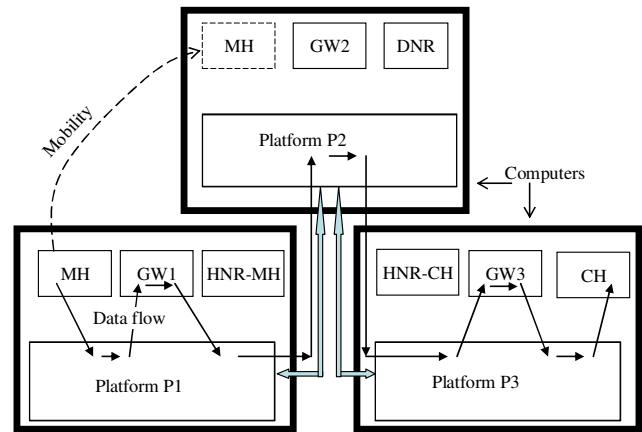


Figure 6. Experimental setup

also includes LNS and IDR functions). The modules are created on the platform by issuing commands from the controller. One or more platforms can be created in the same physical machine or in different machines. The platforms are linked together by TCP sockets, irrespective of if they are located in the same machine or in different machines. Modules communicate to each other through the platforms. For experiments, we have developed an echo application which is similar to the existing ping application. In order to make two hosts communicate to each other, the echo application in sender mode is attached to the sender host and the echo application in responder mode is attached to the other host. The hosts create real data packets that include headers of the transport, identity, and network layers. These packets are sent via the GWs which translate the network layer protocols when they connect two networks that speak dissimilar protocols.

4.2. Experimental Setup

To measure mobility performance, we configured the emulation system as shown in Fig. 6. Three platforms (P1, P2, and P3) are created in three different computers. The MH’s home network exists in P1, while the CH’s network is in P3. That is the MH is connected to GW1 and the CH is connected to GW3. The MH and CH initiate an echo application session between them. The path taken by data packets from the MH to CH is shown by arrows. With this setup (say Setup 1), while communicating with the CH, the MH moves to GW2 located in P2 and executes the mobility signaling functions as discussed in the previous section. We measured handover delays (i.e. time elapsed from the instance the MH detects the new network to the instance it

receives the GW handover complete message from the new GW) in both soft and hard handover cases and the results were as shown in Table 1. It shows that hard handover delays were shorter than soft handover delays for the reason that the MH took longer time to complete the handover signaling in the later case because it was involved in data communication while executing the handover signaling functions. While in hard handover the MH stopped its application data communication and dedicatedly executed the handover procedure. Although the handover delay was larger in smooth handover, no packet was lost. While in hard handover, the packets sent from the CH during the handover time got lost.

We repeated the experiment by slightly changing the configurations. In Setup 2, ten CHs were created in P3, all of which were communicating with the same MH. In this setup, the MH’s handover process completed only when the ID/locator map entries of all CHs stored in GW3 was updated by separate ID/locator update messages sent from GW2. In Setup 3, ten CHs were created in different platforms (three in P1, three in P2, and four in P3) to check how the handover delays would vary if the CHs were located in different networks. The table shows that handover delays of Setup 2 are similar to those of Setup 1. However, delays in Setup 3 are larger than those in the other setups, due the reason that the new GW with which the MH attaches after handover takes longer time to send ID/locator binding update messages to GWs located in different networks and receive the acknowledgements from them. The handover process completes only when the new GW received the ID/locator binding update acknowledgements from all correspondent GWs.

5. SUMMARY AND FUTURE WORK

We presented the distributed mobility management scheme of HIMALIS architecture. Separating signaling and data plane functions in the network and assigning them to the LNSs and GWs, respectively, help make the mobility scheme distributed and secured. The LNSs distribute ID/locator binding updates promptly through the trusted network of IDRs to GWs so that the MH becomes able to receive (send) packets from its new location in a short time.

Table 1. Handover delays in different setups

	Handover Type (S:smooth, H:hard) / Setup#					
	S/1	S/2	S/3	H/1	H/2	H/3
Handover delay (ms)	209	211	253	129	130	211

In future studies, the scheme will be implemented in real protocol stack and tested in a large-scale testbed network such as JGN-X, a large-scale testbed network spread over Japan. Based on the performance evaluation results, the scheme will be optimized by revising the protocol functions. The mature content of this proposal will be then brought to ITU-T for standardization.

REFERENCES

- [1] C. Perkins, "IP mobility support for IPv4," FC 3344, IETF, Aug 2002.
- [2] D. Johnson, et al. "Mobility support in IPv6," RFC 3775, IETF, June 2004.
- [3] R. Koodli, "Mobile IPv6 fast handovers," RFC 5268, IETF, June 2008.
- [4] H. Soliman, "Hierarchical Mobile IPv6 (HMIPv6) mobility management," RFC 5380, IETF, Oct 2008.
- [5] S. Gundavelli, et al., "Proxy Mobile IPv6," RFC 5213, IETF, Aug 2008.
- [6] V.P. Kafle and M. Inoue, "HIMALIS: Heterogeneity inclusion and mobility adaptation through locator ID separation in new generation network," IEICE Trans. Commun. Vol. E93-B, No. 3, March 2010.
- [7] ITU-T Recommendation Y.2015: General requirements for ID/locator separation in NGN, Jan 2009.
- [8] ITU-T Study Group 13, Future networks including mobile and NGN, <<http://www.itu.int/ITU-T/go/sg13>>
- [9] V.P. Kafle and M. Inoue, "Mobility management in HIMALIS architecture," Proceedings of IEEE CCNC 2010.

TOWARD GLOBAL CYBERSECURITY COLLABORATION: CYBERSECURITY OPERATION ACTIVITY MODEL

Takeshi Takahashi* Youki Kadobayashi† Koji Nakao‡

* National Institute of Information and Communications Technology, Tokyo, Japan,
takeshi_takahashi@ieee.org

† Nara Institute of Science and Technology, Nara, Japan, youki-k@is.naist.jp

‡ KDDI Corporation, Tokyo, Japan, ko-nakao@kddi.com

ABSTRACT

The importance of communication and collaboration beyond organizational borders is increasingly recognized with regard to maintaining cybersecurity. Yet organizations still face difficulties communicating and collaborating with external parties. Among these difficulties is the absence of a common vocabulary, as organizations do not always share the same terminology in describing operations, and this consumes unnecessary time and can lead to miscommunication. This paper addresses the problem by introducing a cybersecurity operation activity model that provides the foundation for defining such vocabulary. The model also facilitates understanding and review of cybersecurity operations and their associated activities. This paper demonstrates the model's usability by visualizing the domains of cybersecurity operations and services and concludes that the model has sufficient usability as a foundation for building vocabulary and as a tool for visualizing cybersecurity issues, which will help expedite communication beyond organizational borders.

Keywords— activity model, cybersecurity operation, miscommunication, vocabulary, cybersecurity collaboration

1. INTRODUCTION

As cyber-society develops, cybersecurity has become one of the greatest concerns for organizations. Cyber-threats traverse national borders, and coping with them inevitably requires communication and collaboration beyond organizational borders. Inter-organizational communication and collaboration, however, remain inefficient, while attacks grow increasingly efficient.

The lack of a common vocabulary is one factor causing this difficulty. It renders verbal communication inefficient and time-consuming, even resulting in miscommunication that leads to severe operational flaws. These days, it is quite common to see situations in which organizations use the same vocabulary to describe different issues. This is due to the lack of common terminology, and matters are exacerbated when communicating in non-native languages.

To handle this, this paper introduces a cybersecurity operation activity model that can serve as the foundation for build-

ing a common vocabulary for describing cybersecurity operations. Using the vocabulary enables description of cybersecurity operation processes that differ among organizations, which should serve to streamline and expedite communication across organizational borders.

To build the model, we have held repeated discussions with cybersecurity organizations, including those running cybersecurity operations in the United States, Japan, and South Korea. Since such discussions sometimes address sensitive issues that contain confidential information, we hold the discussion sessions separately. Together with these efforts, we have extracted knowledge and vocabulary from various related works, including cybersecurity operation guidelines [1, 2], cybersecurity information description schemes [3–6], security frameworks [7,8], and ontologies [9–15]. Based on the above, we extracted common cybersecurity operation activities, structured them, and built the activity model. Though some of the above items provide cybersecurity models, they focus on operation of specific entities or specific areas such as incident operation. Different from these, the proposed model describes a series of cybersecurity operation activities beyond entity borders and provides a complete picture of such operations in an incident-centric manner.

To demonstrate the model's usability and applicability, this paper uses the model to describe operations of several cybersecurity services. It also demonstrates the model's usability for visualizing cybersecurity. To build a common vocabulary, this model and vocabulary defined here eventually need to be brought to standard bodies so that many stakeholders can recognize, share, and advance them.

The rest of this paper is organized as follows: Section 2 proposes the activity model for cybersecurity operations, Section 3 demonstrates the model's usability and applicability, and Section 4 concludes the paper.

2. ACTIVITY MODEL

Operation processes consist of operation activities, and are different among organizations, but individual activities should be the same among them. This paper focuses on cybersecurity operation activities, and introduces a cybersecurity operation activity model. This model is designed

from the viewpoint of administrating and maintaining cybersecurity within one organization. To make the model simple and process-independent, in-house communication activities, such as reporting, is omitted from the model. Note that communication between an organization and its outsourcing company are regarded as one type of in-house communication, and is thus omitted from the model.

To structure assorted cybersecurity operation activities, this section introduces four operation stages, i.e., Preventive, Enforcement, Detection, and Responsive stages, and describes cybersecurity operations as the iterations of the stages as shown in Figure 1. The Preventive stage deploys preventive measures against possible cyber-risks and incidents by installing and configuring ICT assets. The Enforcement stage enforces security measures and policies prepared in the Preventive stage. Routine operations during peacetime fall into this stage. The Detection stage detects cybersecurity incidents and risks. Upon detection, it triggers Responsive stage operations. The Responsive stage handles detected incidents and risks and collaborates with external entities when needed. For each of the stages, we defined activities and built a cybersecurity operation activity model, as shown in Figure 2. The activities of each stage are detailed below.

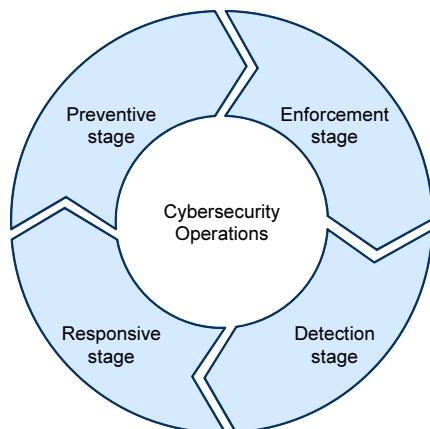


Figure 1. Cybersecurity Operation Stages

2.1. Activities in Preventive Stage

Activities in this stage deploy preventive measures against potential cyber-risks and incidents.

2.1.1. Secure Infrastructure Provisioning

This activity equips and maintains ICT infrastructure with security provisions so that the system may function effectively, efficiently, and securely. This is achieved by running the following sub-activities.

Software and Hardware Development is for developing software and hardware by designing, implementing, testing, and maintaining them. Creating and installing software patches are also included here. For instance, developing

an enterprise resource planning (ERP) software requires review of proper security considerations. Bugs need to be cleared and exception handling should be meticulously implemented.

System Integration is for integrating hardware and software so that they can work effectively, efficiently, and securely. For instance, an organization may integrate a new system, with new hardware and software, with a conventional system. This requires meticulous design, configuration, and thorough testing. Recently, a great deal of vulnerability is created because of poor integration skills rather than security flaws of individual software components.

Network Integration is for integrating network components and building effective, efficient, and secure networks. The system may be integrated with a router, switch, and security appliances such as a firewall and IPS/IDS. This may also involve deploying security zoning, be it either physically or logically. Proper installation and configuration are needed. Note that though this paper describes Secure System Integration and Secure Network Integration separately, they need to be collaborated for maintaining cybersecurity; otherwise they push responsibilities onto each other, which leads to vulnerabilities.

Service Subscription is for subscribing to appropriate external services, e.g., from internet service providers (ISPs), application service providers (ASPs), and cloud service providers (CSPs), so that the system may function effectively, efficiently, and securely. An organization needs to choose proper subscriptions for data center services. For instance, laws applicable to data in a data center may differ depending on the data center's location, thus proper subscription and configuration are needed for maintaining the security level. Proper subscription management for users' subscribing to external services is also needed. For instance, one user may subscribe to a service while another does not, or users may use only domestic cloud services.

Note that the above activities cover a wide range of sub-activities, and thus could be regarded as independent from cybersecurity operations, though they are partly overlapping.

2.1.2. Security Policy Design

This activity identifies security policies, manually or automatically, and implements them in the system. Traffic filtering rules, including blacklists, packet filtering policies, and content access policies, may be established and implemented in the system. The rules may be defined based on manual configuration. They can also be identified automatically by using policy-mining technologies that study users' system usage and identify proper security policies. Many IDS/IPS products already implement this automatic policy identification option.

2.1.3. Measurement Design

This activity establishes measurement schemes, including logging and packet monitoring. For instance, the target and

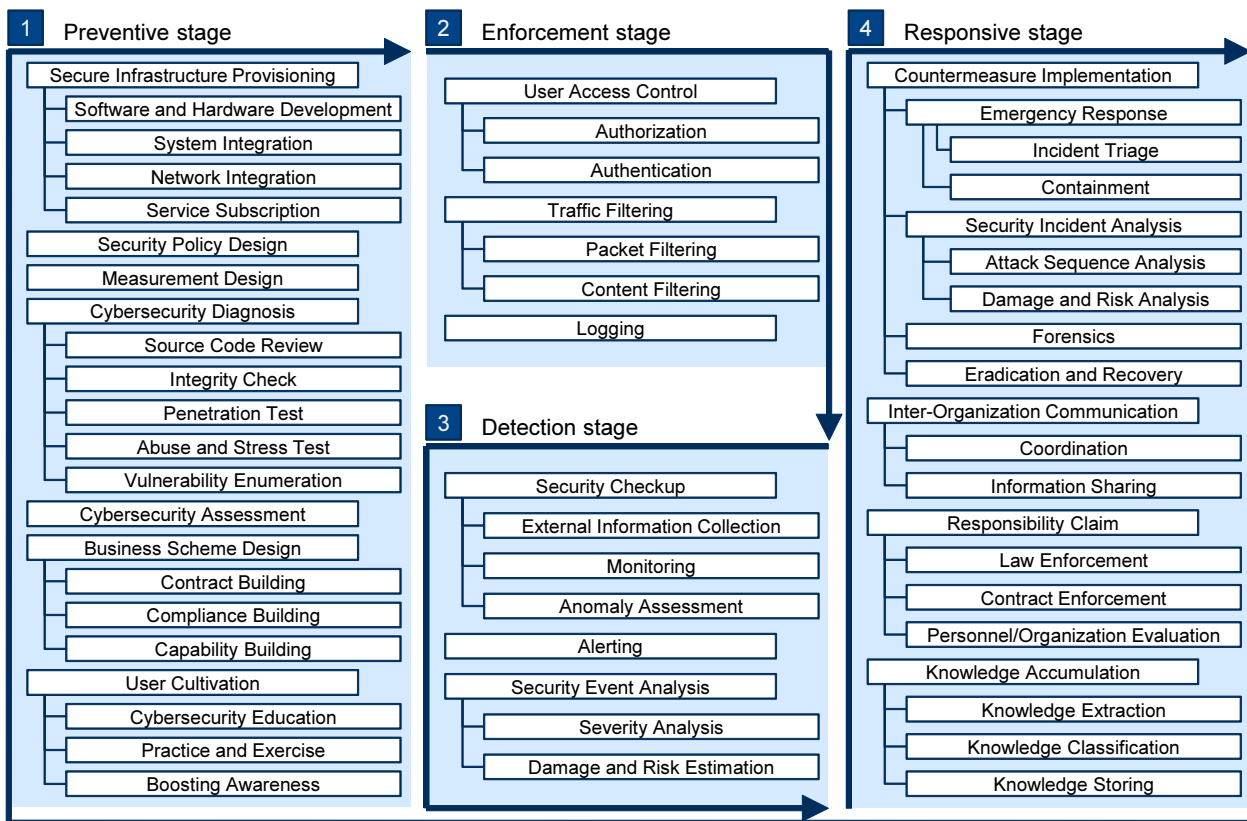


Figure 2. Cybersecurity Operation Activity Model

format of logging are defined. Apart from that, an integrity check list may be built, which is useful for reviewing compliance with the security policy defined by Security Policy Design. Criteria may also be provided for judging anomalies.

2.1.4. Cybersecurity Diagnosis

This activity diagnoses cybersecurity risks within systems by conducting tests and assessments¹. It is usually run throughout the system development lifecycle, thus is often run by Infrastructure Provisioning. It includes the following sub-activities.

Source Code Review is for checking whether software contains vulnerabilities; i.e., white box tests. Based on known vulnerability patterns of programming language, verification is made on the existence of similar such patterns, and diagnosis is made for the existence of source code that has potential risks of, for instance, buffer overflow and vulnerability against injection attacks.

Integrity Check is for checking the integrity of an ICT system with a predefined security policy. For instance, individual users may wish to customize their systems, and this activity comprehensively investigates that customization. Checks

¹It could be argued that some activities, such as penetration tests, could be classified as being in the Detection stage. This is merely a matter of taxonomy and is outside the interests of this paper.

are made on whether the password length is long enough as defined in the security guidelines.

Penetration Test is for attempting to penetrate a system and checks systems for presence of security risks. In doing this, the difficulties of intruding need proving. This activity may use vulnerability scanning tools.

Abuse and Stress Test is for checking whether a system has sufficient resistance against abuses and excessive burdens. To check resistance against Distributed Denial of Service (DDoS), spam, etc., the tester imposes an excessive network burden on a network. The system must verify the strength of its planned resistance against such attacks.

Vulnerability Enumeration is for investigating a system to find vulnerabilities, which can be done by external security service providers as a service. For instance, this activity scans ICT systems to detect vulnerability by utilizing past knowhow and vulnerability information, such as vulnerability notes, and warning information as well as a checklist. This can be done manually or by running a vulnerability scanning tool.

2.1.5. Cybersecurity Assessment

This activity assesses a system's cybersecurity status based on the vulnerability information provided by Cybersecurity Diagnosis and the internal audit report. It may score the cy-

bersecurity status considering the vulnerability level and information confidentiality.

2.1.6. Business Scheme Design

This activity designs business schemes needed to handle security incidents and includes the following sub-activities.

Contract Building is for building contracts that compensate for possible losses caused by cybersecurity incidents. For instance, an organization may create a contract for subcontractors defining monetary compensation for confidential information leakage, which serves as a deterrent.

Compliance Building is for building compliance that regulates the usage of ICT assets. For instance, an organization may create a compliance item defining sensitive information and restrictions on disclosing it to outside parties.

Capability Building is for building the cybersecurity handling capability inside an organization. This activity may include establishing an incident response team or system administration department. Team-building inside these organizations is also included in the activity.

2.1.7. User Cultivation

This activity cultivates users and increases the awareness, knowledge, and skills of users and includes the following sub-activities.

Cybersecurity Education is for educating users on using the system properly. Users are taught, for instance, the functionality of various ICT assets, their appropriate usage, and troubleshooting techniques.

Practice and Exercise is for training users to use the system properly. Users are given exercises so that they can put the knowledge learned from Cybersecurity Education into practice when needed. As an exercise, an email with a virus attached may be sent to users, and those who open the attachment receive a warning and are entered in a cybersecurity review course.

Boosting Awareness is for boosting awareness of cybersecurity. It may involve publicizing via posters, flyers, brochures, and other media to increase awareness of the importance of cybersecurity and the threat of security incidents.

2.2. Activities in Enforcement Stage

Activities in this stage enforces security measures and policies prepared in the preventive stages.

2.2.1. User Access Control

This activity provides appropriate accessibility to ICT assets for authorized users, and includes the following two sub-activities.

Authorization delegates proper access rights to authorized users. For instance, different access rights may be assigned to

personnel depending on their position so that sensitive management information is only accessible to board members.

Authentication confirms the user as authentic; that is, that claims the user made are true. When a user accesses a system with an ID, this activity authenticates the user's access rights.

2.2.2. Traffic Filtering

This activity filters traffic, both incoming and outgoing ones, following predefined policies. This can be done by routers, firewalls, and/or hosts. The following sub-activities are the two types of traffic filtering.

Packet Filtering inspects and discards packets following pre-defined packet filtering rule. For instance, packets may be blocked from sources identified by blacklists. Or, in the case of a high traffic burden, arbitrary packets could be dropped to sustain the system's functionality.

Content Filtering inspects the content of the traffic and discards them following the content security policy. Email filtering and website access blocking are typical examples.

2.2.3. Logging

This activity logs transactions handled and observed by the system. It logs not only system alerts but also data such as user access histories. The logs are used, for instance, to track the activities of an ID or service, etc. They are crucial and indispensable tools for running Detection stage operations.

2.3. Activities in Detection Stage

Activities in this stage detect cybersecurity incidents and risks and include the following sub-activities.

2.3.1. Security Checkup

This activity comprehends a system's current status by collecting relevant information such as system logs, and checking the system's health. It is usually performed routinely to ensure that systems are secured with the latest fixes, only permitted applications are running with proper approved releases, and the system is generally free of viruses and worms. It includes the following three sub-activities.

External Information Collection is for collecting external cybersecurity information. This activity may include collecting such information by, for instance, subscribing to vulnerability and warning information services. Such information includes software patches.

Monitoring is for monitoring information obtained from ICT assets, including alerts, logs, and configuration information. For instance, various logs and alerts sent by IDS are monitored, checks are made on whether security risks arise, and determination is made on whether the risks were caused by an attack.

Anomaly Assessment is for assessing the anomaly and security risk level of ICT assets. It can be based on the information obtained by External Information Collection, the result of Monitoring, or can use pre-defined security assessment rules. For instance, checks are made on whether any vulnerability exists by running Cybersecurity Diagnosis in the Preventive stage.

2.3.2. Alerting

This activity raises alerts if Anomaly Assessment recognizes significant anomaly and security risks needing inspection and investigation. For instance, a system may send alert messages to administrators upon detecting security risks beyond predefined criteria. The organization may then deploy emergency procedures.

2.3.3. Security Event Analysis

This activity analyzes and integrates security event information obtained from multiple security appliances to clarify the big picture of the security incident. This may involve confirming if the alerted events were caused by incident; Alerting could be sometimes taken for non-incident events, i.e. false detection. It usually requires the following two sub-activities. Note, as discussed before, operations in this stage may run Cybersecurity Diagnosis in the Preventive stage.

Severity Analysis is for analyzing the severity of security events. An event may be transient and disappear, or it may cause severe trouble. Regardless of the reason for the event, be it either an attack or a benign incident, if it has the potential to cause severe trouble, the severity needs to be analyzed.

Damage and Risk Estimation is for estimating damage and further risks caused by security events. Quick investigation is made on the entire ICT system. For instance, a situation such as functional failure of routers and password leakage needs to be clearly grasped.

2.4. Activities in Responsive Stage

Activities in this stage handle detected incidents and risks to prevent damages from spreading further and to remedy them.

2.4.1. Countermeasure Implementation

This activity handles detected incidents and risks to impede the damages from spreading further and to remedy them and is achieved by running the following sub-activities within each organization.

Emergency Response is for responding to detected incidents and risks with countermeasures. Note that it sometimes may not be needed depending on the type of incident. For instance, incidents that have ongoing effects on the system need this activity while those whose malicious behavior has already stopped may ignore it. It includes Incident Triage and Containment. **Incident Triage** is for evaluating the emergence of incidents and events and prioritizes needed

operations. For instance, incoming incident reports provided by the activities in the Detection stage are interpreted, prioritized, and associated with ongoing incidents and trends. **Containment** is for taking emergency treatment actions to contain the damage. This may include isolation of affected systems and platforms.

Security Incident Analysis is for analyzing incidents detected in the Detection stage. It includes Attack Sequence Analysis and Damage and Risk Analysis. **Attack Sequence Analysis** is for analyzing information and collects trails and evidence of attacks so that the attack sequence can be clarified. For instance, this activity clarifies which vulnerability the attack exploited, when, and how. It may also analyze the network topological problem that allowed the attacks to occur. **Damage and Risk Analysis** is for analyzing damage and further risks caused by security events. It investigates the entire ICT system and identifies the damage. It also clarifies the risks that may result.

Forensics is for collecting evidence, including the identification, preservation, extraction, and documentation of computer-based evidence. Such information may be hidden from view; thus, special forensic software tools and techniques are required. These can be used to identify passwords, log-ons, and other information that may have been deleted from the computer's memory. They can also be used to identify backdated files and to tie a diskette to the computer that created it. In some cases, interception/traceback can be conducted. The activity should be consistent and compatible with any incident response process.

Eradication and Recovery is for eliminating incident causes and effects and returns systems to a normal operational state. Measures implemented in the Preventive stage usually need to be updated here to eliminate the same risk. This could be seen as a preventive stage activity for future security incidents.

2.4.2. Inter-Organization Communication

This activity communicates with external organization to collaboratively handle incidents and includes the following sub-activities.

Coordination is for notifying and coordinating with appropriate internal and external parties on a need-to-know basis to prevent damage from spreading further. Relevant parties include the civilian, national security, and law enforcement communities.

Information Sharing is for sharing information with appropriate internal and external parties. This can be done based on mutual trust between the parties, or based on trustworthy coordination.

Note that collaborative mitigation operations are enabled by Countermeasure Implementation and Inter-Organization Communication and prevent damage from expanding further and to implement countermeasures.

2.4.3. Responsibility Claim

This activity seeks compensation for damages. For instance, a user/organization may be evaluated based on the degree of responsibility and negligence. It includes the following sub-activities.

Law Enforcement is for taking legal actions, which includes commencing civil suits, running criminal investigation and claiming negligence liability. Note that this activity covers a wide range of sub-activities, and is usually regarded as independent from cybersecurity operations though they are partly overlapping.

Contract Enforcement is for initiating penalties based on contracts including employment contracts. For instance, users violating compliance are subject to a defined penalty.

Personnel/Organization Evaluation is for evaluating the user/organization responsible for the detected incidents. This works as a deterrent against future incident.

2.4.4. Knowledge Accumulation

This activity extracts and accumulates knowledge in order to reuse it and includes the following sub-activities.

Knowledge Extraction is for extracting knowledge including expertise on cybersecurity gleaned from incident response operations. Note that sometimes knowledge is extracted from Cybersecurity Research, which may include attack and threat research, but we see it as an adjacent operation of Cybersecurity Operation.

Knowledge Classification is for classifying knowledge and assigning tags so that the knowledge becomes effectively and efficiently retrievable.

Knowledge Storing is for storing knowledge so that it can be reused in the future.

3. USABILITY AND APPLICABILITY

This section demonstrates the model's usability and applicability in terms of its expressiveness and visualization capability.

3.1. Expressiveness of the model

The model is designed to provide the foundation of common vocabulary describing cybersecurity operation and facilitate communications across organizational borders. To facilitate communication, the model needs to be capable of describing assorted operations. To demonstrate its expressiveness, this section uses the model to describe four major services of cybersecurity service providers. Figure 3 highlights the domains of the services, which are detailed in the following sections. Note that services differ among providers, and the demonstration is based on an average service.

3.1.1. Intrusion Detection Service

This service runs intrusion detection, a passive security activity that sits on the network or on selected hosts, on behalf of a customer organization.

To detect intrusions, it needs to store assorted logs by using IDS-supportive devices (Logging). Based on the log, it runs a Security Checkup; it collects a range of information from external organizations, including attack trends, vulnerability information, and attack signatures (External Information Collection), monitors the organization's system (Monitoring), and runs Anomaly Assessment. When an anomaly is detected, it alerts the customer organization (Alerting) and runs Security Event Analysis including Severity Analysis and Damage and Risk Estimation. Upon intrusion detection, it runs Emergency Response.

In some cases, service providers implement Eradication and Recovery. Sometimes collaborative mitigation could be taken since service providers usually monitor multiple customers. They update preventive operations by, for instance, updating attack signatures. Security policy may also be redesigned here to eliminate the same risk in the future (Security Policy Design).

Apart from that, this service typically scans the system for vulnerabilities (Vulnerability Enumeration) as a preventive measure.

3.1.2. Risk Management Service

This service assists customer organizations with managing security risks. Its main activity is Cybersecurity Diagnosis; it may run a Source Code Review, Integrity Check, Penetration Test, Abuse and Stress Test, and Vulnerability Enumeration. With the diagnosis result, it assesses the system's cybersecurity (Cybersecurity Assessment). Based on the assessment result, it may develop risk management guidance for supporting the customer's risk management program. It may also define the criteria of judging anomalies (Measurement Design).

3.1.3. Incident Handling Service

This service handles incidents occurring in a customer organization. Its major domain resides in the Responsive stage, while some service providers may also, for instance, review security policies (Security Policy Design) in the Preventive stage.

Upon detecting incidents, it runs Incident Triage if needed, then may implement Containment, during which it also pays attention to the system's sustainability so that the customer organization can continue to use the system to run its business. In parallel, the service runs Security Incident Analysis, including Attack Sequence Analysis and Damage and Risk Analysis, to understand the details of the incidents and/or risks. Eradication and Recovery are taken based on the Security Incident Analysis results. The service also in parallel runs Forensics, which collects evidence of attacks. Note that

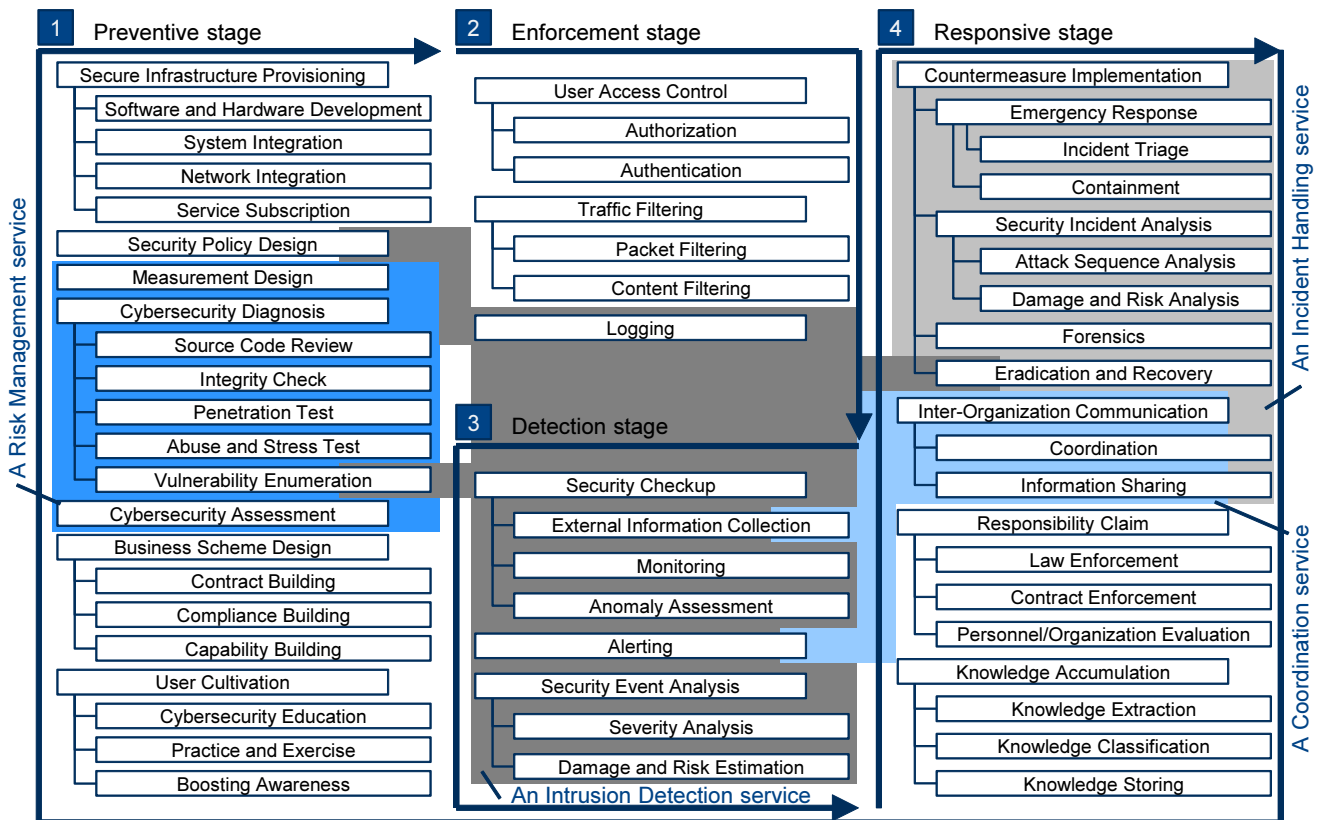


Figure 3. Major Cybersecurity Services

the above Containment procedure should be taken carefully to prevent needed evidence from disappearing.

The service may collaborate with external organizations when needed and may ask a CERT to run Coordination. It then shares information on the attack (Information Sharing), though it needs to protect the customer organization's information confidentiality.

3.1.4. Coordination Service

This service, typically provided by a coordination center such as a CERT, facilitates cooperation among multiple organizations to maintain cybersecurity. Its main service domain resides in the Responsive stage, while it offers several supporting activities in the Detection stage.

The service provider runs Coordination to coordinate various organizations in order to maintain cybersecurity when needed. It receives various reports on cybersecurity incidents and risks from different organizations, and shares that information among them (Information Sharing). It may sometimes lead collaborative mitigation.

On the other hand, user organizations communicate with the service provider to update cybersecurity information (External Information Collection). The provider may sometimes provide alerts directly to user organizations (Alerting).

3.2. Visualization

The proposed model is useful not only for defining vocabulary but also for visualizing security-related issues.

3.2.1. Visualization of Security Service Domains

Various cybersecurity services exist, yet service differences between companies are not so clear. There may be situations in which two different companies provide services with the same name, yet of a completely different nature. A number of SOC service providers are emerging, offering very different features, yet users are unaware of this. Consequently, users subscribing to an SOC service with minimum functionality may depend excessively on the service and overlook implementing other necessary cybersecurity measures.

This model provides a means of visualizing the domain of security services as shown in Section 3.1 and Figure 3. Users can gain a clear understanding of service contents, and clearly recognize what a service does and what they need to do for themselves.

3.2.2. Security Measure Implementation Status Review

Many companies are investing in implementing cybersecurity measures. Yet these investments are sometimes neither effective nor efficient since they tend to be unbalanced. Some

activities receive heavy investment while others do not. To be effective, investment needs to cover all aspects of cybersecurity.

The model provides a means for visualizing the current status of security measure implementation within an organization. Based on the visualization, the organization may decide on which activities to further invest in.

4. CONCLUSION AND FUTURE WORK

The proposed cybersecurity operation activity model provides sufficient expression for describing major cybersecurity operations and is also useful for visualizing cybersecurity. We believe that the model provides a foundation for defining vocabulary. Documenting and structuring cybersecurity activities with the model also contribute to the development of cybersecurity for developing countries.

Nevertheless, to expedite and facilitate communication and collaboration beyond organizational borders, the vocabulary needs to be commonly acknowledged and used among organizations. Building a global standard helps achieve this goal. It helps the model and vocabulary to be recognized and used more by stakeholders. During the standardization process, the model itself may be developed further to accommodate their interests. Eventually, we hope to see the resulting model and vocabulary expedite communication beyond the borders of organizations, countries, and even languages through global standards.

REFERENCES

- [1] NIST Special Publications (800 Series). <http://csrc.nist.gov/publications/PubsSPs.html>, Sept. 2011.
- [2] N. Brownlee and E. Guttman, "Expectations for Computer Security Incident Response," RFC 2350, Internet Engineering Task Force, June 1998.
- [3] The MITRE Corporation, "Making Security Measurable," <http://msm.mitre.org/>, Sept. 2011.
- [4] R. Danyliw et al., "The Incident Object Description Exchange Format," RFC 5070, IETF, Dec. 2007.
- [5] R. Martin, "Making Security Measurable and Manageable," *CrossTalk, the Journal of Defense Software Engineering*, Sept. 2009.
- [6] ICASI, "The Common Vulnerability Reporting Framework (CVRF) v1.0," <http://www.icasi.org/cvrf>, Sept. 2011.
- [7] NIST, "Federal Information Security Management Act (FISMA) Implementation Project," <http://csrc.nist.gov/groups/SMA/fisma/>, Sept. 2011.
- [8] A. Rutkowski et al., "CYBEX – The Cybersecurity Information Exchange Framework (X.1500)," *ACM SIGCOMM Computer Communication Review*, Oct. 2010.
- [9] J. Wang and G. Minzhe, "Security data mining in an ontology for vulnerability management," in *IJCBS*, 2009.
- [10] S. Parkin, et al., "An information security ontology incorporating human-behavioural implications," in *SIN*, 2009.
- [11] B. Tsoumas and D. Gritzalis, "Towards an ontology-based security management," in *AINA*, 2006.
- [12] S. Fenz and A. Ekelhart, "Formalizing information security knowledge," in *ASIACCS*, 2009.
- [13] C. Blanco, et al., "A systematic review and comparison of security ontologies," in *ARES*, 2008.
- [14] G. Denker et al., "Security in the semantic web using owl," in *Information Security Technical Report*, 2005.
- [15] T. Takahashi et al., "Ontological approach toward cybersecurity in cloud computing," in *SIN*, 2010.

CONTEXT REPRESENTATION FORMALISM AND ITS INTEGRATION INTO CONTEXT AS A SERVICE IN CLOUDS

Boris Moltchanov

Telecom Italia

ABSTRACT

Context Management technology itself is not novel and ICT companies are already trying to find a technically feasible solution and appealing marketing usage of the context-awareness. However, after many years of technology scouting and academic scrutiny within this still innovating area, the usage of the personalized and context-aware services is still limited due to totally new business models that shall be put in place. The context information available in the real world from many potential context sources shall be handled as a near real-time, efficiently processed by many devices and be interoperable among different actors dealing with the context. Therefore among a comprehensive context management framework and its efficient representation the context information shall be exposed in an way easy to use and consume. Even more better is if the context information and data are embedded within service clouds using frontier technology of cloud computing embedding not only the tools for a reach, flexible and scalable service creation but also integrating context knowledge and an efficient real-time data management. A solution integrating the context information in the clouds with its efficient context representation and publish-subscribe web service based interface is described in this paper.

Keywords—context, cloud, service, context-awareness, interface, representation formalism

1. INTRODUCTION

The context information acquisition and handling requires efficient and simple interoperable representation formalism, especially for exposing this information to the human readable interfaces such as internet browsers and Machine-to-Machine (M2M) frameworks. This task of selection for representation reference within heterogeneous environments such as telecommunication networks consisting of the mobile network equipment and mobile devices shall be pondered by consideration of many peculiarities including resource- and energy-constrained mobile devices with embedded context sensors, limited and “expensive” connectivity bandwidth and limited data transmitting capacity of the mobile network operators within interested

Thanks to the European Commission for funding the research program and projects

geographic areas, where the context information could be acquired and used. Moreover, the reference choice for the representation formalism should be preferably standardized or de-facto adopted solution allowing a wide employment on many devices and equipment and large usage among large reference market. Telecom Italia, working in the context awareness field for many years, has performed a careful selection of the context representation based on the above mentioned principles, and this work describes this representation as well as its evolution and usage in the telecommunication services. The context information is under embedment within service clouds by means of the web service compliant interfaces implementing publish/subscribe interactions.

This work, once started as based solely on the mobile network operator’s requirements then has been proven and extended within various European research projects aimed to create systems and platforms using or handling the context information acquired from the operator’s customers, physical environment sensors, Internet and from network equipment. Therefore, this work presents the context-awareness and its integration into clouds not as a “vertical” “narrow” solution dedicated only to Telecom Italia’s preferences, rather as a careful academic research work that permitted to create a comprehensive formalism, very simple and interoperable on many energy and resource-constrained devices and at the same time using sufficiently flexible format allowing its extension for additional functionalities and easy integration of comprehensive security features. This created formalism is based on world recognized and widely adopted standards. Currently this formalism with its respective communication interface is under embedding into the XaaS (Everything-as-a-Service) platforms within two ongoing European research projects.

First section describes the requirements listed for creation of the context representation formalism, then the selected form is described with its usage model and its evolution is presented as add-ons to the initially selected formalism, such as a simple query language based on the same reference representation. Some examples of the interface usage in a number of demo and prototypes are shown as proof of concept. The paper wraps up with conclusions and mentioning of the current work in progress on the context management interface integration into clouds.

2. CONTEXT MANAGEMENT FRAMEWORK

Although telecom operators are always handled a lot of customers with their personal devices connected to the fixed line phone or mobile phone networks, the only services provided to the customers by the telecommunication networks where the voice-calls and the text messaging on a mobile, later evolved into reach communication services. Currently many services are offered or supported by a Mobile Network Operator (MNO) are leveraging on a better communication quality control, a wider available bandwidth per customer pondering the Internet navigation and the content distribution, and on a better knowledge about the customers available from the network itself and customers' devices. This last aspect related to the customers' habits, their location, preferences, etc. is brought under the umbrella of the context-awareness. Although the context knowledge and exposal is a very strong driver opening new dimensions within the operator's ecosystem, at the same time this is also a great amount of data to be treated in a real-time or near real-time mode, and it opens plenty of privacy related issues along huge business opportunities. Telecom Italia has started research and development in this field many years ago and is continuing to learn its knowledge and experience evolving the context-awareness and the context management framework within a number of research projects involving many industrial and academic research partners under a severe academic scrutiny. The project list is including but is not limited to C-CAST [8], MobiLife [9], MUSIC [10], OPUCE [11], PERSIST [12], SPICE [13] and currently ongoing 4CaaS [14] and FL-WARE [15] projects. All these projects are contributed to the current results with a severe long academic research to select a solution satisfactory for context management and feasible and applicable to telecommunication network operator environment. Moreover, most even if not all, of the mentioned projects have benefitted from the chosen solution in terms of innovation services prototyping and trialing. The main requirements for the context management from the beginning have been the following ones:

- To involve and use the context information available from the owned operated network and its subscribed customers, from the customers themselves (in both modes "opt-in" and "opt-out") and from the 3rd parties in Internet;
- Run on wide range of heterogeneous mobile devices with acceptable performance, interworking and impacting as little as possible on the battery life of the customer devices and on customer's User Experience;
- Be real- or near real-time for a dynamic service configuration and execution;
- Be distributed and flexible in order to perform context related operations within dedicated context information domains (source, transport, distribution, consumer, etc.);
- Be extendable and scalable in order to extend the system by enriching the context information and to scale for acceptable overall performance;

- Be interoperable and interworking over the large operator's ecosystem including the networks, supporting Information Technology (IT) nodes, storages and myriad of customers' devices;
- Be exposable over a SOA or Web Service interface through Service Delivery Platform or Cloud Solution;
- Be secure and compliant with privacy regulations.

A generalized Context Management Model adopted by Telecom Italia is shown in the following 0

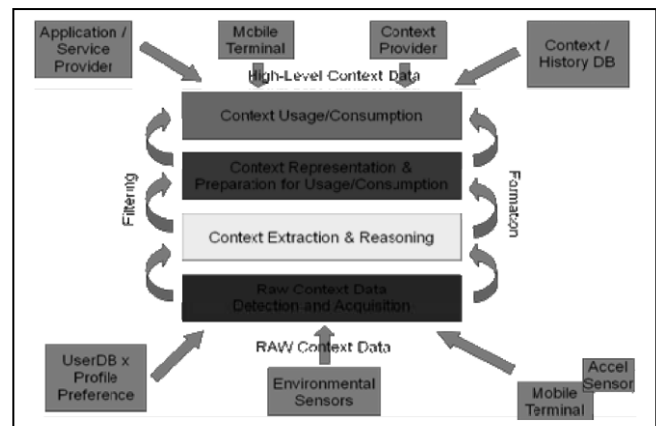


Figure 1. Context Management Model adopted by Telecom Italia

While a simplified context management framework [1] filled with the context represented within transmitted packets is shown in the below 0

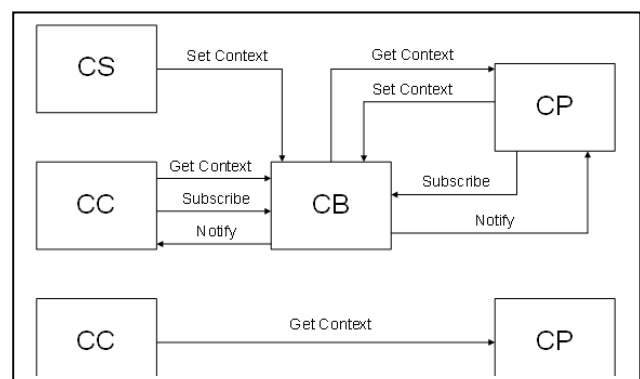


Figure 2. Context Management and Methods used by Telecom Italia

Proposed context processing model is very simple in this centralized scheme [4] based on a single context broker and including roles of context providers or sources and context consumers. Nevertheless this scheme [2] shown above is under current work of its extension to a federated context brokerage concept based on well-know and standardized (also "de-facto") Internet protocols, as one shown in the 0The current evolution is also consisting of integration of the context data and of the context-awareness, as a dedicated Service Enabler, into the Cloud Services technologies, i.e. Context as a Service (CaaS).

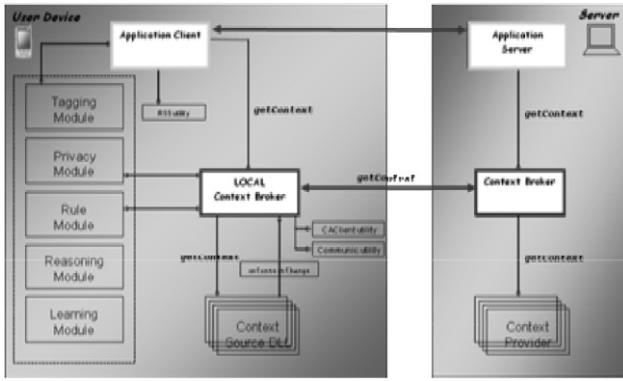


Figure 3. Context Management Model adopted by Telecom Italia

3. CONTEXTML

The ContextML [3] is a context data representation scheme or language chosen mainly with respect to the above listed initial requirements and focused especially onto the portability among the devices within the MNO's ecosystem. This formalism is extendible, able to represent any available within the ecosystem context information and contains a minimum necessary information to transport within its payload using its tag-value schema. Therefore it is simple with minimum overheads and overloads, easy to extend and to process and based on the open standards. Indeed, the ContextML is based on the XML technology adapted to the context-awareness needs.

ContextML includes the following elements:

- *Entity* – a source or owner of the context information available within the context management system. This element must be always present within a ContextML document;
- *Scope* – a context data consisting of a tag name and a value published within the ContextML document, which may contain more than one context scopes regarding the same entity;
- *Time-stamp* – a date and time of the context creation or acquisition that is very important for a real- or near real-time context peculiarity;
- *Validity* or *expiration time* – a context expiration date and time or a validity time till which the context information may be considered valid, that is very important especially for a near real-time concept.

The entity/scope association is shown in the 0

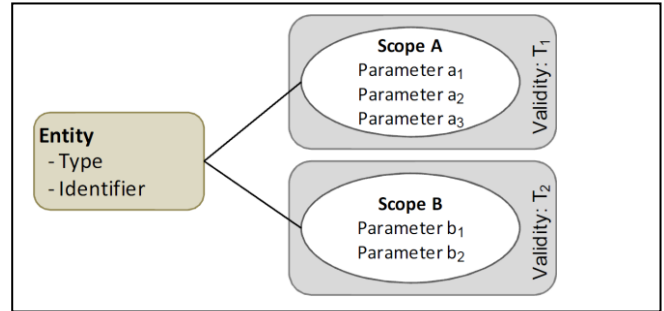


Figure 4. Entity/Scope association in ContextML

A very small snippet of the ContextML is given in the following 0

```

<contextML>
  <ctxEls>
    <ctxEl>
      <contextProvider id="LP" v="1.1.0" />
      <entity type="username" id="Max" />
      <scope>civilAddress</scope>
      <timestamp>2007-02-27T12:20:11+01:00</timestamp>
      <expires>2009-02-27T13:20:11+01:00</expires>
      <dataPart>
        <parS n="civilAddress">
          <par n="room">1037</par>
          <par n="corridor">North</par>
          <par n="floor">2</par>
          <par n="building">B</par>
          <par n="street">Via G. Reiss Romoli
            274</par>
          <par n="postalCode">10148</par>
          <par n="city">Torino</par>
          <par n="subdivision">TO</par>
          <par n="country">Italy</par>
        </parS>
      </dataPart>
    </ctxEl>
  </ctxEls>
</contextML>

```

Figure 5. ContextML document snippet

The ContextML 0 is easy to integrate within a Web Service or any other SOA technologies like REST used by Telecom Italia, is simple to include security elements and is ready to embed into a security suite. The REST communication, very similar to the HTTP, consist of two only communication request-respoce (GET and response) to retrieve any information or to acknowledge (ACK/NACK) a malfunctioning within the response. Therefore no additional overheads provided neither to the communication network nor to its components, such as network equipment, switches and nodes, nor to the communicating parts such as server and client. Moreover, chosen ContextML over REST communication protocol interoperates over a vast range of server platforms and mobile devices due to its XML legacy with a durable and successfully proven best-practice employment and vastest usage experience. One very important ContextML property is its ability to be

automatically processed in M2M operations during the context distribution and its readiness for secondary context acquisition mechanisms such as context aggregation, extraction, reasoning and prediction. There ContextML integrated within clouds is easily available and simply to process for both the cloud supporting infrastructure and for the context-aware service implementation and execution.

Additionally, ContextML has its intrinsic semantic characteristics inherited from the XML including the hierarchically nestled information nodes and the parameter/value data structures. Nevertheless, a more sophisticated context semantic and an application or service domain specific ontology are missing and required to be integrated into a system benefitting from a better intelligence and an autonomous automatic context processing. Therefore the context management model and the employed context representation are continuously evolving towards a higher intelligence, a social artifacts embedding, a better security and privacy management and an autonomous computing and self-QoS provisioning accordingly to the service configurations and SLAs, while always remaining respecting the initially assigned requirements.

4. CONTEXT OPEARTION MESSAGES

The ContextML [3] representation formalism described in the previous section is a format for the payloads in the communications between different components composing the Context Management Framework consisting of the context producers (Context Providers), context using or context aware applications and services (Context Consumers) and centralized context handlers (Context Brokers) mentioned in the first section. ContextML is employed as packets payload within the communication transport protocol between aforementioned components such as REST or XMPP. The messages in this communications are the following ones:

- Context Provider Advertisement is a type of message from a Context Provider to a Context Broker announcing the Internet address of the provider and the context information (context entities and context scopes) it could provide to the system. Thanks to this message a Context Broker knows how and where to retrieve the context information (scopes) regarding a certain entity. There is no alternative of this message for detaching a Context Provider, instead the Context Provider shall periodically send its advertisement repetitively as a keep-alive message, otherwise the Context Broker will clean its record regarding this Context Provider in 5 minutes of inactivity. This is implemented as a simple mechanism of cleaning the Context Brokers from unused or non working Context Providers;
- Context information retrieval and Context Provider lookup requests are the REST GET messages from a Context Consumer to a Context Broker requiring in a synchronous mode the information regarding a Context

Provider (Internet address) being provided with required context scope or context information and with entity's ID and required context scopes. Context information retrieval message could be also used directly to a Context Provider from a Context Consumer being provided with the Context Provider's Internet address;

- Context publishing is a type of message sent from a Context Provider to a Context Broker or to a Context Consumer, or from a Context Broker to a Context Consumer. This message contains the context information (scopes) regarding an entity being before requested or subscribed by a Context Consumer or by a Context Broker;
- Acknowledgement is a type of message sent from a Context Provider or from a Context Broker to a Context Broker or to a Context Consumer respectively on a context information request or on a subscription request indicating a correct context subscription (when no yet context information is available) or a "malformed request" response to a Context Provider or to a Context Broker;
- Context subscription is a type of message sent from a Context Consumer to a Context Broker requiring certain context information (scopes) regarding an entity.

The methods invocation examples are not given here for the sake of the text simplicity and its useless complexity due to the reason that they are solely HTTP-like REST GET and POST messages or their XMPP alternatives. However, in case of a necessity the communication examples could be demonstrated in a demo showing a real communication between existing components – part of the running in production or a test-bed set-ups, taking into account first the privacy measures of a non disclosure of real entities and of their sensitive context information handled by Telecom Italia.

5. CONTEXTQL (CQL)

Additionally to the unconditional context retrieval by the getting context queries the Context Management Platform integrates an event- or context-based publish/subscribe interface [6], [7]. It allows to formulate the event conditions in a specific schema, the Context Query Language (ContextQL or CQL), which is similar to the well known Structured Query Language (SQL). In addition, a callback URL is provided for required conditional context retrieval when the conditions would become true.

An example of a query condition is given in 0

```

<contextQL>
  <ctxQuery>
    <action type="SUBSCRIBE"/>
    <entity>username|boris</entity>
    <scope>civilAddress</scope>
    <validity>180</validity>
    <conds>
      <cond type="ONVALUE">
        <constraint      param="civilAddress.city"      op="EQ"
value="Turin"/>
      </cond>
    </conds>
  </ctxQuery>
</contextQL>

```

Figure 6. ContextCQL example

The operand of the context query messages could assume the following values: matching the context conditions (ON-VALUE), being more or less value (MORE or LESS), on arrival of the context information (ON-AVAILABILITY) and on a contest information update regarding certain entity (ON-UPDATE).

In order to allow to mix the abovementioned conditions thus creating more complex subscription requests and more comprehensive therefore precise entity selections or context retrieval, the CQL supports the following operators: equals (EQ), unequals (NEQ), starts with (STW), contains (CONT), ends with (ENW). Each subscription is bound to a specified validity time (in seconds) but can be easily renewed. All subscriptions are acknowledged by Context Brokers and all subscribed messages are acknowledged by Context Consumers.

6. CONTEXTML AND CQL INTEGRATION INTO CLOUDS

Currently ongoing research and integrated projects with Telecom Italia's context-awareness and context management platform aim to context-awareness integration and context information access during the cloud service creation and execution. The context information access via ContextML and CQL through REST interface is created based on the Enabler concept in Open Mobile Alliance (OMA) [16]. Moreover the same ideas are supported by Telecom Italia within OMA as the new Next Generation Services Interface (NGSI) Enabler specification. The mechanism of the interface is based on the publish/subscribe paradigm supported by both the interfaces ContextML and CQL. The consumer of the context information shall subscribe through this enabler to the context information and entity it desires involve, then the enabler will publish the requested information based on the programmed conditions. While a context provider shall publish its context information to the context enabler that could be subscribed by context consumers: context-aware services and applications. However the context interface allowing to context providers to provision their context into

the clouds is not yet exposed through the cloud context enabler and will be supported as future work. That exposed interface would allow to register and to use the context enabler by an end-user or developers in the same way as the context consumption. Therefore, for the moment, only the context enabler for context consumption is exposed and can be used by the applications and services running in the clouds.

7. USAGE AND APPLICATION EXAMPLES

Research projects listed in the first section of this paper have provided number of prototypes and service demos, some of which still could be found in the projects' respective sites. Telecom Italia on its side, as an industrial partner, largely exploited the context-awareness developed during the last years and implemented in various services in production and services prototypes, beginning many years ago from simple context-aware content share services such as Mosaic shown in 0This service allowed to the Telecom Italia's customers to publish their User Generated Content (UGC) (pictures, audio and video) taken by their mobile phones using the context information about the place, time, and other conditions related to the multimedia content. At the same time mobile customers were able to vote the content placed by other customers. And finally, during this event, only the most voted and related to certain (local) place content has been shown on the big screens installed in different places of Venice city.

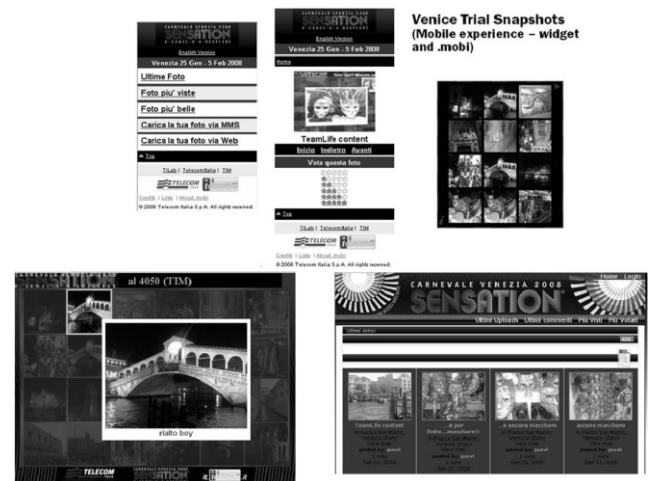


Figure 7. Mosaic platform and Carnevale di Venezia screenshot

Another service is Graffiti or eTourism based on context-aware recommendation concept when mobile customers visiting certain places published their content and feedback regarding visited places. And other customers visiting later the same places were able to both see the content and recommendations left there by other people as well as leave their own content and recommendations for further visitors.. This concept of Points Of Interest recommendations is demonstrated in the 0

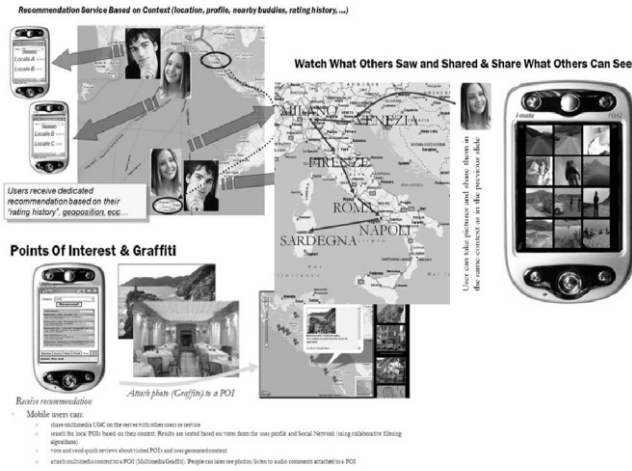


Figure 8. Graffiti service and iTourism solution screenshots

Then many other services, involving context information and context-awareness have been designed and created for local Public Administrations such as Bicistaffetta: shown in the 0The bicycle tour or race moving around Italy across Italian cities and villages has been provided with the localised content left there by other customers and Public Administration. Also here the selection of the content regarding certain place and its local information such as hotels, place to eat, etc. is done based on the customer location for the content tagged in a particular way enabling its context-aware selection.



Figure 9. Bicistaffetta and Valle d'Aosta PA screenshots

Currently Telecom Italia is working on implementation and running of the services oriented to an extensive social-networks usage within its context management platform such as for e-books social writing, editing and commenting as shown in 0 The customers can collaboratively write the books, edit already written content and comments during the writing, editing or reading processes in a way that everyone can contribute to the content or express her/his opinion.



Figure 10. eBook usage scenario of Telecom Italia

Telecom Italia's check-in into places as Social Places concept demonstrated in the 0This is also a mix of the context-aware content selection service interworking with social networking concepts, e.i. a customer can be check-into a certain service or location and receive recommendations, opinions and content for this service or place left by a parent or a friend, or a friend-of-friend.

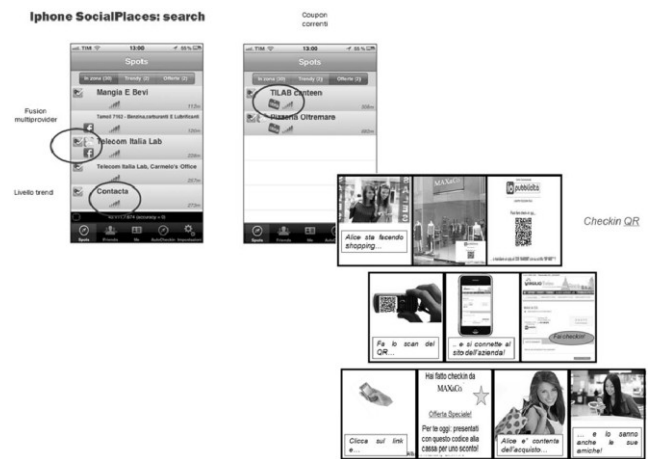


Figure 11. Check-in into places service screenshots

Another very interesting and appealing context-awareness usage in a daily life is context-aware augmented reality. A snap-shot of such kind of service by Telecom Italia is shown below:



Figure 12. SeeAR augmented reality service screenshots

Here a customer observing her/his surroundings over mobile phone's camera receives a lot of useful information regarding the place and objects such as distance to monuments, opening hours, tickets costs and description of museums and other highlights regarding the place as well as the content and opinions, comments or rating left by other customers. A mobile customer may even see which other social related customers (friends, friends of friends, parents, etc.) are close to her/him and eventually contact them or invite to an event.

There are a lot of such services and systems based on or using the context-awareness within innovated telecommunication services of Telecom Italia and listing of all of them as well as giving examples of all the context-aware services is out-of-scope of this paper. Table 1 reassumes the context-aware service classes and employment of a context and content information there.

Telecom Italia does not have yet context-aware or personalized services in the clouds run in production therefore here is no yet example of its usage. However within 4CaaSt [14] EU project there is already created a demo of taxi service employing the context information for both actors the taxi company and for customers of taxi service to efficiently assign and serve the customer by taxi and to provide useful information about the customer to the taxi as well as to taxi customer about her/his trip. Thus the context management integration into clouds has been already started and currently is in an intensive exploitation.

Table 1. Context-aware service classes and content/context usage

Service Class	U G C	Context	Professional Content	Real World Info
Recommendation	X	X (location, comments, opinion, ranking, social)	X (opening hours, days, promotions)	
Entertainment	X	X (location, user preferences, history scores/opinion, social)		
Augmented Reality	X	X (location, ranking, weather conditions, comments, traffic conditions, social)	X (opening hours, days, promotions, description, pictures, video)	X
Public Administration		X (location, weather conditions, traffic intensity)	X (advertisement, video, pictures)	
Events		X (location, any other context relevant for event or customers)	X (advertisement, any other information relevant for event)	

8. CONCLUSIONS AND ONGOING WORK

The context model employed by Telecom Italia during many years initially involved mainly the entities under direct control of TIM (Telecom Italia Mobile) operator, therefore TIM customers, and aimed for usage as internal Telecom Italia provided services and applications. However, this solution has been disseminated and its functionality has been included within Open Mobile Alliance (OMA) standardization as Next Generation Service Interface (NGSI) Enabler [5]. And currently Telecom Italia is extending its context-awareness domain as

context information acquisition from external entities. Therefore a context federation concept involving many federated Context Brokers, integration of the context management and context access by proving interfaces into the cloud technology as a context (service) enabler (Context-as-a-Service CaaS) and a more precise and secure privacy control are strongly required. These three areas of an efficient context federation model, CaaS and a security context information exchange and privacy control remaining interoperable and totally compliant with regulation requirements through the standard and “*de-facto*” solutions used and law-regulations enforcements respectively are three main topics, where Telecom Italia is intensively working within its internal innovation research and development and actively involving the academy research partner within European research projects and initiatives. One of the current technologies under study for further applications within the Telecom Italia’s Context Management Framework are the open-source protocols OpenID and OAuth employed already for a long time by many world-wide ICT companies with large customer base and a number of available context-aware or context providing services. OMA standardization contributes to definition and building of the context enabler for a further wide usage. While peer-to-peer communications protocols are expected significantly contribute to future federated solution of Telecom Italia.

Although Telecom Italia uses a lot of standards formal and de-facto, on another hand Telecom Italia significantly contributes to OMA and W3C working on context awareness and context enabler as far as this area is totally new and not yet standardized.

REFERENCES

- [1] B. Moltchanov, M. Knappmeyer, C.A. Licciardi, “Context-Aware Content Sharing and Casting”, 12th ICIN, Bordeaux, France, October 2008
- [2] M. Zafar, N. Baker, B. Moltchanov, J.M. Goncalves, S.L. Kiani and M. Knappmeyer, “Context Management Architecture for Future Internet Services”, ICT MobileSummit 2009, Santander, Spain, June 2009
- [3] M. Knappmeyer, S.L. Kiani, C. Fra', B. Moltchanov, N. Baker, “ContextML: A light-weight context representation and context management schema”, Wireless Pervasive Computing (ISWPC), 2010 5th IEEE International Symposium, Modena, Italy, 5-7 May 2010
- [4] M. Valla, C. Fra', W.L. Goix, M. Marchetti, E. Paschetta, A. Salmeri, “Architettura e moduli della Context Awareness Platform”, Telecom Italia Lab - Technical Report, DPC2006.01818, December 2006
- [5] M. Bauer, N. Ito, E. Kovacs, A. Schülke, C. Criminisi, L.W. Goix, M. Valla, “The Context API in the OMA Next Generation Service Interface”, In Proc. of the 14th International Conference on Intelligence in Next Generation Networks (ICIN 2010), Berlin, Germany, 11-14 Oct. 2010
- [6] R. Reichle, M. Wagner, M.U. Khan, K. Geihs, M. Valla, C. Frà, N. Paspallis, G.A. Papadopoulos, “A Context Query Language for Pervasive Computing Environments”, 5th IEEE Workshop on Context Modeling and Reasoning (CoMoRea) in conjunction with the 6th IEEE International Conference on Pervasive Computing and Communication (PerCom'08), Hong Kong, 17–21 March 2008
- [7] C. Frà, M. Valla and N. Paspallis, “High Level Context Query Processing: An Experience Report”, 8th IEEE Workshop on Context Modeling and Reasoning (CoMoRea) in conjunction with the 9th IEEE International Conference on Pervasive Computing and Communication (PerCom'11), Seattle, WA (USA), 21–25 March 2011
- [8] Context Casting (C-CAST), FP7 ICT Research Project, Web Site: <http://www.ict-ccast.eu>
- [9] MobiLife, FP6 IST Integrated Project, Web Site: <http://www.ist-mobilife.org>.
- [10] self-adapting applications for Mobile USers In ubiquitous Computing environment (MUSIC), FP6 Integrated Project, Web Site: <http://ist-music.berlios.de>
- [11] Open Platform for User-centric service Creation (OPUCE), FP6 IST Research Project, Web Site: <http://www.opuce.tid.es>
- [12] Personal Self-Improving Smart Spases (PERSIST), FP7 ICT Research Project, Web Site: <http://www.ict-persist.eu>
- [13] Service Platform for Innovative Communication Environment (SPICE), FP6 IST Project, Web Site: <http://www.ist-spice.org>
- [14] 4CaaS, FP7 ICT Integrated Project, WebSite: <http://4caast.morfeo-project.org>
- [15] Future of Internet WARE (FI-WARE), Public Private Partnership (PPP) Progeram Future of Internet core platform project, <http://fi-ware.morfeo-project.org>.
- [16] Open Mobile Alliance (OMA) reference architecture and enablers specification, Web Site: <http://www.openmobilealliance.org>.

ContextML XML schema – available on: <http://contextml.tilab.com>

SUPPORTING TECHNICALLY THE CONTINUITY OF MEDICAL CARE: STATUS REPORT AND PERSPECTIVES

B. Spyropoulos, M. Botsivaly, A. Tzavaras

Biomedical Technology Laboratory, Medical Instrumentation Technology Department, Technological Educational Institute (TEI) of Athens, Athens, Greece

ABSTRACT

The purpose of this paper is to present the status of the R&D efforts of our Laboratory concerning the development and the improvement of hardware and software means, appropriately designed to ensure Continuity of Medical Care among Primary Health-care Agencies, Hospitals and Home Care, according to existing or emerging National, European and International regulations and standards. Our R&D is presently focused on the development of an integrated prototype system, including first, improved equipment facilitating the Continuity of Care, second, software supporting Medical Decision Making during emergency-care delivery, as well as, real and virtual audio-visual monitoring of chronic and terminally ill patients at home-alike conditions, third, a Continuity of Care Record (CCR), complying with the major ANSI E2369-05 CCR, ISO 13606-1:2008 and prEN 13940 Standards, and finally, linking the CCR to appropriate semantically annotated Web-Services, providing for enhanced technical interoperability and medical clarity and simplicity.

Keywords— Continuity of Care Record (CCR), Home Care Equipment, Clinical Decision Software, Semantics.

1. INTRODUCTION

In the near future, due to the explosion of health-related expenditure, a dramatic reduction of traditional hospitalization-length is anticipated. The employment of cost-effective care-schemata for the elderly and the disabled, such as extended and systematic Home-care, seems to be inevitable. The objective of our Research and Development activity is the development and/or the improvement of hardware and software means, appropriately designed to ensure Continuity of Medical Care among Primary Health-care Agencies, Hospitals and Home Care, according to existing or emerging National, International and European regulations and standards. More specific, the objectives and components of the project were and partially are:

- The systematic recording and sorting of International Standards, Codifications, Directives etc. relevant to Continuity of Care.
- The development of a Continuity of Care Record-CCR.

- The development of new and the improvement of existing Biomedical Equipment supporting the Continuity of Medical Care.
- The development of software supporting Medical Decision Making during emergency-care delivery, and observing chronic and terminally ill patients, at home-alike conditions.
- The development of software-means for virtual audio-visual remote supervision of patients, combined with their vital signs monitoring.
- The design and partial implementation of semantically annotated Web-Services ensuring the Continuity of Care.
- The internal evaluation of the deliverables to ensure functional compatibility of the developed components and the compliance of the system with a quality system, leading to accreditation of the service package to be delivered after its completion and functional testing.

The anticipated final outcome is the development of a prototype system, hardware and software, ensuring all aspects of continuity of Care, among any provider, complying with the major ANSI E2369-05 CCR, ISO 13606-1:2008 and prEN 13940 Standards, and finally, providing for enhanced interoperability, due to the adoption of HL7-CDA, combined with the employment of semantics. In the next paragraphs, the present the state of the art in the field of Information Technology supporting the Continuity of Medical Care, based upon appropriate Industrial Property Documents, and the development status of our system, will be reported and documented in details.

2. THE STATE OF THE ART AS DISCLOSED IN RELEVANT PATENT DOCUMENTS

In order to establish the state of the art, an extensive search concerning patented software and equipment has been performed, by employing the European Patent Office search-engine esp@cenet, for continuity of medical care and other terms, related to our project. The following patent documents, beyond the broad scientific literature of the field, and numerous other patents that have been retrieved and evaluated, represent adequately, precisely and legally binding, the “state of the art” for the field of Information Technology supporting the Continuity of Medical Care.

Document WO 2011027006 (A1): "...relates to a system for storing and managing the complete medical records of patients, which essentially comprises a system that allows the common user to access all of his/her medical information, as well as an operation that can be used to monitor and manage the continuity of assisted care system, with the following order of action: a periodic review of all the protocols or templates activated during a given time period, the entry into the monitoring system of the periodicity recommended by the doctor, and the automatic review of compliance by the patient" [2]. This is the most recent patent application retrieved (publication date: 10-03-2011), however, no decision support software or monitoring hardware is mentioned to be included in the proposed invention.

Document WO 2011022178 (A2): "...permits the documentation of medical information on a cell phone, enabling improved communication between health care provider and patient. Electronic medical records will be transferable to the cell phone and be stored. The owner of the cell phone will be able to maintain a longitudinal record of health information and be able to share this information with different health care providers" [3]. This is the second most recent patent application (publication date: 24-02-2011), however, it is limited to the management of some documentation of medical information on a cell phone.

Document US 2010280350 (A1): "The present invention is tele-diagnostics and patient triage method and system for the practices of traditional Chinese medicine (TCM) that acquire patient health condition info through 4 diagnostics process, provide pathogenesis analysis (disease cause, progression, projection), and treatment and prescription options remotely. This method and system established medical business modules with standard syntax through the implementation of medical decision support tools such as relational database, lookup tables, calculators, decision trees, manifestation reference charts, case comparison and statistics, and visualization in disease, pattern and syndrome identification, differentiation and determination" [4]. This patent application (publication date: 04-11-2010), comes closer than the previous ones to our objectives, however, it is limited to the very narrow field of a tele-diagnostics and patient triage method and system for the practices of traditional Chinese medicine. Further, no treatment support and no semantically annotated continuity of care systems are even mentioned.

Document CN 101764857 (A): "...discloses a realization method for mobile handover of a next generation of IP full wireless sensor network. The wireless sensor network comprises a gateway node, a fixed sensor node, a mobile sensor node and a correlative node of the mobile sensor node" [5]. This interesting Chinese patent application (publication date: 30-06-2010), however, it is limited to a realization method for a next generation of IP full wireless sensor network (IPv6).

Document US 2009254361 (A1): "A consumer is provided with continuity of medical care by being prompted to transfer data to a recipient, upon completion of the consumer's interaction with a brokerage system. The data obtained during the interactions is transferred to one or

more recipients selected by the consumer" [6]. This patent application (publication date: 08-10-2009), comes close to our objectives, however, what is claimed is limited to a computer-implemented method of providing a consumer of services with continuity of medical care, comprising prompting the consumer to transfer data to a recipient, upon completion of the consumer's interaction with a brokerage system; and transferring data obtained during the consumer's one or more interactions with the brokerage system to one or more recipients selected by the consumer. The development of new and the improvement of existing Biomedical Equipment supporting the Continuity of Medical Care, the development of software supporting Medical Decision Making during emergency-care delivery or monitoring of chronic and terminally ill patients, at home-alike conditions, and finally, the provision of enhanced interoperability, due to the adoption of HL7-CDA, combined with the employment of semantics, that constitute important features of our proposals, are not at all pointed out in the said patent.

Document US 2009216558 (A1): "An automated system is described for presenting a patient with an online interactive personal health record (PHR) capable of delivering individualized alerts based on comparison of evidence-based standards of care to information related to the patient's actual medical care. A health care organization collects and processes medical care information, including clinical data relating to a patient in order to generate and deliver customized clinical alerts and personalized wellness alerts directly to the patient via the PHR. The PHR also solicits the patient's input for tracking of alert follow-up actions and allows the health care organization to track alert outcomes. Further embodiments include implementing a plurality of modules for providing real-time processing and delivery of clinical alerts and personalized wellness alerts to the patient via the PHR and to a health care provider via one or more health care provider applications, including disease management applications" [7]. This interesting patent application (publication date: 27-08-2009) describes and claims an automated online interactive personal health record (PHR), not a continuity of care system, delivering individualized alerts based on comparison of evidence-based standards of care to information related to the patient's actual medical care, however, does not claim neither the development of equipment and software supporting care, nor semantics employment in their system.

Document US 2009171692 (A1): "An online health care consumer portal for accessing one or more health-related services by a health care consumer. The consumer portal includes an authentication module for identifying the health care consumer upon receiving an online identification token, and a database for maintaining health care information comprising a plurality of health records. The consumer portal also includes a rules engine module for applying a set of rules to the one or more health records corresponding to the consumer to determine an impairment profile of the health care consumer based on the health records corresponding to the health care consumer and an online user interface for providing access to the health-related services. The online user interface is configurable

from a default configuration to a second configuration upon identification by the authentication module, the second configuration adapted to facilitate use of the online user interface according to the impairment profile” [8]. This patent application (publication date: 02-07-2009) is focused on an online health care consumer portal for accessing one or more health-related services offered by a health care consumer.

Document KR 20090001730 (A): “A continuity consultation document generation system and a method, and a recording medium for recording a program therefore are provided to present a continuity consultation document to a user, by rendering the continuity consultation document that is automatically generated up to an entry level” [9]. This patent application (publication date: 09-01-2009) covers only the creation of electronic documents, related to continuity of care.

Document US 2008215372 (A1): “The present invention is directed to a method and device for ensuring patient continuity of care. One aspect of the present method includes providing to a patient a hand-held portable device that has at least a portion of the patient's medical record stored thereon. The patient can then carry the hand-held portable device on his person for access of the information thereon whenever necessary or desired” [10]. This patent application (publication date: 04-09-2008), is limited, according to its main claim, to a method for ensuring patient continuity of care, the method comprising the step of providing to a patient, a hand-held portable device having at least a first portion of said patient's medical record stored electronically therein, such that said patient can carry said at least a portion of said patient's medical record on said patient's person for access by said patient or another authorized individual when access is desired, and so far does not match at all with our proposal.

Document US 2006116911 (A1): “The present invention generally relates to the field of medical services and more specifically to the area of chronic disease management. It comprises a new method for providing a disease management service which utilizes nurse practitioners to engage in regularly scheduled virtual evaluation and management "office visits" with patients, using off-the-shelf videophones for real-time video and audio communications”[11]. This patent application (publication date: 01-06-2006) is focused mainly on achieving reimbursement of “virtual office visits" by public and private insurers.

Document US 2004210458 (A1): “Methods and platforms for enhancing collaboration and communication between a patient and his healthcare team are described. A personal health record is created for a patient and maintained by a service provider. The health record is updated with self-monitored or remote device readings. These readings are sent, in a secure format that insures patient privacy, to the service provider and inserted into a health record via a computer connected to the Internet or via a telephone line without the use of a computer, i.e., by directly connecting an intermediate device to a phone outlet. Other health and wellness data may be written to the health record via a computer or via conventional means” [12]. This patent

application (publication date: 21-10-2004) is one of the pioneer documents in the field, it is focused mainly on the creation of a personal health record for a patient and its maintenance by a service provider, and does not refer to most of the important issues of our proposal.

Document US 2005027569 (A1): “Systems and methods for documenting an encounter and communicating about same are described. The systems and methods of this invention generally comprise an electronic records system for creating and maintaining information in electronic records; a point-of-encounter system in communication with the electronic records system, wherein the point-of-encounter system allows a scribe to document the encounter into a predetermined electronic record; and a library of event-specific templates usable for documenting the encounter. The systems of this invention allow new dynamically-generated templates to be created and added to the library of event-specific templates as needed” [13]. This patent application (publication date: 03-02-2003) is focused on methods for documenting an encounter and communicating about same, and it is rather outdated.

The extensive search and presentation of the context of the state of the art in the field of continuity of Medical Care, concerning our R&D activity presentation, proves that the objectives of the ongoing project, have not yet been met on Industrial Property level, nor accomplished in any commercially available product. Therefore, the ongoing research project remains important for the field.

3. THE METHODOLOGY AND THE STRUCTURE OF OUR PROJECT

Ensuring the Continuity of Medical Care as a patient moves between different points of care is a complicated medical, technical and managerial task. The project is focused presently on the epidemiologically most significant Cardio-respiratory Diseases, and has partially completed the development of several prototype devices and appropriate application software, compliant to all relevant Medical Protocols, Encodings, Guidelines, Technical Standards and Regulations. The detailed presentation of the components of the system and their development status is following.

3.1. Development of a Continuity of Care Record.

There are almost 50 different standards related to Electronic Health Records [14], and this fact means that we are not yet able to solve all interoperability problems that reflect the historical, social and cultural diversity in the progress of e-Health Research, Innovation and Standardization.

ASTM, an American National Standards Institute (ANSI) standard development organization, has as first approved 2005 the E2369-05, Standard Specification for Continuity of Care Record (CCR). The CCR is intended to assure at least a minimum standard of health information transportability, when a patient is discharged, referred or transferred, fostering thus and improving continuity in care [15], [16].

The International Organization for Standardization (ISO) has also recently approved the ISO 13606-1:2008 Standard Electronic health record communication (Part 1: Reference model), which is a new standard for the communication and semantic interoperability of electronic health record extracts [17].

Finally, the European answer to these efforts [18] was the publication of the preliminary European Standard for Continuity of Care (prEN 13940 CONTsys: Systems of Concepts to Support Continuity of Care prEN 14463 [ClaML: A syntax to represent the content of medical classification systems]).

Based on the combination of these standards and on their evolution, a prototype system has been developed [19]-[23], which will allow for the creation of a Continuity of Care Record (CCR). This record is designed to be employed during the post-discharge period and to cover the need to organize and make transportable a set of basic patient information, consisting of the most relevant and recently facts about a patient's condition, together with a care plan, that is, recommendations for future care. The structure of this CCR record will assist the Continuity of Medical Care, in both the administrative and the medical aspects. This module allows for the CCR to be prepared, transmitted, and viewed in any browser, in an HL7 CDA (Clinical Document Architecture)-compliant document format. Furthermore, the CCR is designed to be technology and vendor neutral, allowing thus in different EHR systems to both import and export all relevant data to and from the CCR document.

3.2. Development and improvement of equipment supporting Continuity of Medical Care.

Ensuring the Continuity of Medical Care and especially in the case of cardio-respiratory diseases requires the employment of low-cost equipment that will allow for, first, the collection and recording of electrical and non-electrical biosignals, related to the bodily functions, under any circumstances (e.g. in home care, in rehabilitation centers etc.), and second, measuring of parameters associated with Supported Ventilation, for example, for patients suffering Chronic Obstructive Pulmonary Disease (COPD).

We are being developing various versions of integrated prototype systems based on low-cost, commercially available components, and Notebook PC. The presently developed system includes an Electrocardiography (ECG) acquisition module, equipped with an RF-link between amplifier and PC, a finger pulse Oximetry probe for typical plethysmography based Oxygen Saturation (SpO₂) measurements and the estimation of Heart Rate (HR) and Respiration Rate (RR), and finally, a Respiratory, Carotid and Pulmonary Sounds acquisition module, based on a microphone array [24]-[26].

A device for recording the distribution of the Thoracic Impedance, based on software-driven multiplexed electrodes-array, under high-frequency voltage is under advanced development status.

A sampling-device from the breathing circuit of gas-cylinder O₂, C-PAP, compressor etc supported patient, employing adaptable Flow, Pressure, PO₂, PCO₂ and other sensors has been designed and is being assembled and tested.

3.3. Development of software supporting Medical Decision Making

Ensuring the Continuity of Medical Care often requires supporting Medical Decision Making in the following areas: First, emergency health care delivery outside the hospital boundaries (home-care, hotels, cruisers, airports etc.), second, monitoring of chronic patients or patients discharged from a Hospital, after a major surgical or medical intervention, and finally, treatment of terminally ill patients.

Following software tools have been developed until now: First, software for the detection of shockable ventricular fibrillation (VF) and malignant ventricular tachycardia (VT) based on data acquired by the previously described module. Second, software for Mechanical Ventilation Optimization (e.g. for patients with COPD). This software implements neuro-fuzzy logic and genetic algorithms on data acquired by the gas sampling system described previously. In view of the fact that patient needs are not static, the system will finally produce an advice on the percentage change of Oxygen supply to the patient. Third, software for the acquisition, processing, storage and transmission of various in vitro Diagnostic procedures tested at the point of care. Finally, software for the application of the appropriate Medical Guidelines for the treatment of an emergency patient, according to the patient's condition.

3.4. Design of Semantically annotated Web Services Ensuring the Continuity of Care.

Interoperability of health care information systems has become one of the most crucial and challenging aspects in the healthcare domain. Clinical terminologies and vocabularies, such as SNOMED, ICD-9, ICD-10, and LOINC are already in use for several years and they provide a well established description of the medical domain knowledge. Furthermore, the HL7-CDA (Health Level 7 Clinical Document Architecture) is already in use by several countries, providing for a common representation of clinical documents, enabling the clinical document exchange and facilitating document management. Nevertheless, it is unrealistic to expect that all care providers will agree on adopting a single standard allowing for the interoperability of different health care information systems. The emerging Semantic Web that will employ semantically annotated Web-Services and in which information will have a well-defined machine-interpretable meaning, appears currently to be the most appealing approach towards this direction.

Presently, an application is been developed including: First, software for the creation and the management of a “Care Plan” allowing for any physician to create structured profiles of activities (monitoring, treatment, diagnostic and nursing activities) that should be employed in the post discharge period. These profiles will be defined according to the different diagnoses, taking into consideration each patient’s specific characteristics, needs and demands. Second, software appropriately designed to support Pharmaco-vigilance. Third, software that will consist of prototype ontology based upon the HL7-CDA, and an application that will convert the referral documents into CDA-compliant format and the contents of the CDA-compliant documents into ontology instances. Finally, an appropriately designed semantically annotated Web-service, for the distribution of the documents over IP, providing for adequate security for the transferred data.

3.5. Design of Software for Virtual Audiovisual remote surveillance supporting the Continuity of Care.

Appropriate software has been developed and tested, enabling virtual medical surveillance of sensitive and high-risk populations. In this context, the primary collected Medical data, the properly processed secondary data, and the interactively collected audiovisual patient’s information (fixed and moving images, speech, body sounds, etc.) will be transmitted to the desired surveillance site. The transmission will be achieved through wireless point-to-point links and over IP. The design of software supporting the remote monitoring is based mainly on the “Virtual Server” technique, and fulfills the following requirements: First, transmission of audiovisual data. (e.g. bedside image during Homecare, images from incubators in a Neonatal-ICU etc.), second, transmission of primary and secondary data, which will also appear on the remote screen and finally, transmission of bedside medical images, taken by various equipment (e.g. portable Ultrasound and Doppler, UV/VIS/IR images of the patient), spectral-photometric profile of light reflected from the skin etc. Adequate security for the transferred data is provided.

3.6. Internal evaluation of the project and preparation of a Quality System for future Service Certification

The internal evaluation of the developed components, to ensure their mutual compatibility and their compliance with International Standards, Codes, Guidelines and Regulations, is a rather complicated task. We are presently preparing a systematic recording of this information, related to the Continuity of Medical Care that will be posted on a website, indicating the progress of state of the art regarding this project to facilitate the evaluation procedure, as well as, the preparation of a “Handbook supporting Continuity of Medical Care”.

This manual will comprise of, first, the scientific background relevant to the support of the Continuity of Medical Care, second, the Industrial property rights relevant to the hardware related to the Continuity of Medical Care, third, the International Standards, Codes, medical protocols, guidelines, legislation, etc. and finally an outline of a comprehensive Quality Management System supporting the Continuity of Medical Care (Human Resources, Training, Equipment, Software, Procedures and Guidelines, Certification, Accreditation and Quality Assurance. This handbook, although focused on Cardio-respiratory Diseases, will provide a “reference guide” for the certification of Services, provided from similar applications, related to Continuity of Care in other Medical Disciplines.

Table 1. Overview and short description of the components of the system supporting the Continuity of Care.

Overview of the main components of the system
Software for the formation of a Continuity of Care record, combining the E2369 (CCR), the ISO 13606-1 and the prEN 13940 standards.
An integrated prototype Notebook-based system enabling the monitoring of Biosignals, Thoracic Impedance, Respiratory, Carotid and Pulmonary Sounds, in vitro Diagnostics Point of Care testing, and the post-discharge supported Respiration or Ventilation.
Medical Decision-making support for Cardiovascular and Respiratory Diseases, enabling post-discharge monitoring evaluation for emergency response, and ensuring Mechanical Ventilation settings Optimization.
Design of Semantically annotated Web-Services ensuring the Continuity of Care, including post-discharge Care Plan and Pharmaco-vigilance.
Software for Virtual Audiovisual Remote monitoring, supporting the continuity of care.
The Internal Evaluation Report on the integration success of the hardware components and the software modules of the system.
A Handbook supporting Continuity of Medical Care.

In Figure 1 a schematical outline of the ongoing project is being presented, displaying a description of the main components of the system. The main software components, the developed biomedical hardware and the legal, scientific and industrial property framework and their mutual interactions are clearly indicated.

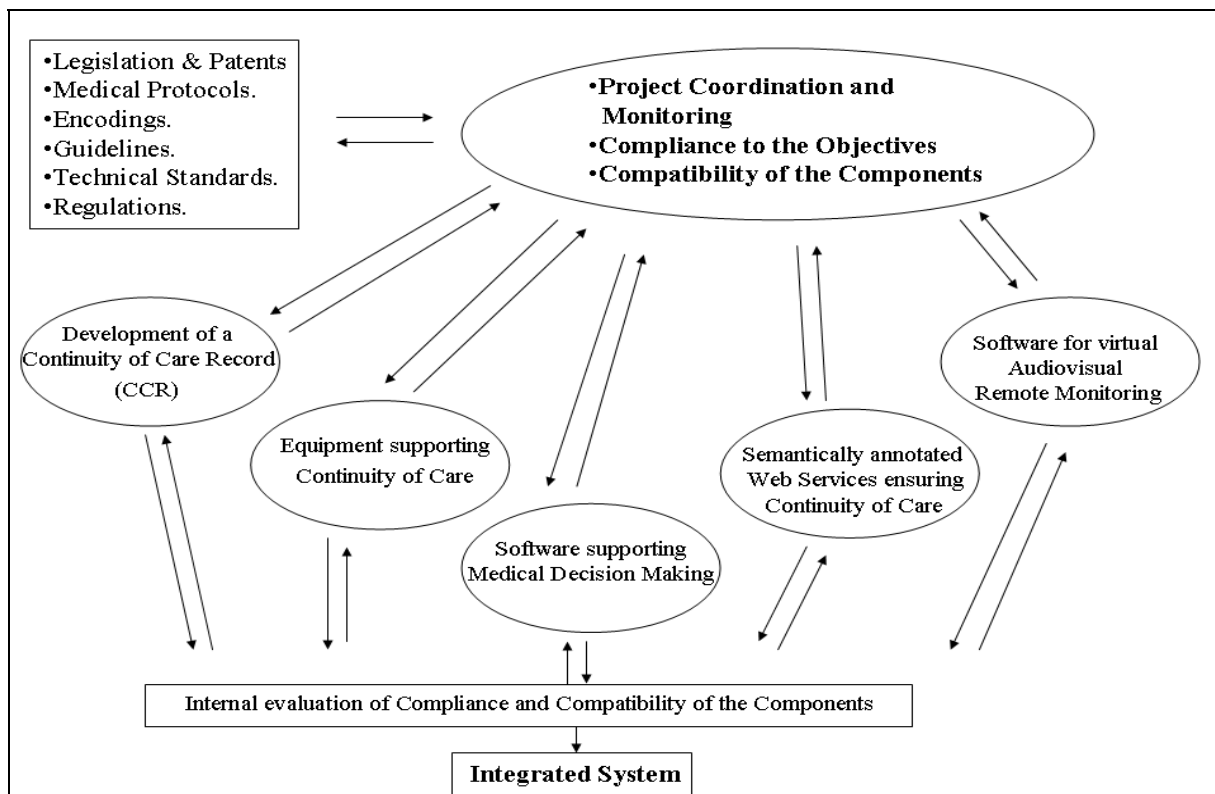


Figure 1. A schematical outline of the ongoing project presenting a description of the main components of the system and their mutual interactions.

4. IMPLEMENTATION REMARKS

Comparing the presented state of the art with the structure and the methodological approach of our ongoing project, it is becoming clear the lack of an integrated system, supporting Continuity of Medical Care. This fact is clearly indicated by the following individual innovative aspects that are being presently under development and testing.

The development of a complete framework to support Continuity of Medical Care could become a prototype for the commercial providers.

The development of Continuity of Care software complying with the requirements of all three major standards ANSI E2369 (CCR), ISO 13606-1 and pre-EN 13940, promotes the improvement of the interoperability between existing systems, and the interim bridging of their lack in several regions.

The employment of recording Respiratory and Pulmonary Sounds by means of a microphone-array, and the recording of the distribution of thoracic Impedance by means of an array of electrodes, multiplexed in pairs to high-frequency voltage, combined with gas sampling and testing from the patient's breathing circuit, may dramatically optimize the lungs-condition monitoring of home-care patients, depended on mechanically supported ventilation.

An example of the present status of development of the combined recording of the distribution of thoracic Impedance by means of an array of electrodes, multiplexed in pairs to high-frequency voltage and of measurements of

Respiratory and Pulmonary Sounds, taken by means of a microphone-array is presented in Figure 2.

The development of Medical Decision Support Software, supporting important Continuity of Care issues, constitutes an innovation, because it allows for automatic assessment of the need for Automatic External Defibrillation, and optimization of patient's supported Respiration, by employing appropriate settings.

The offered possibility of acquisition, processing, storage and transmission of in vitro Diagnostic profiles tested at the Point of Care (PoCT), as well as, the retrieval and employment of Emergency Medical Protocols and Guidelines, supporting the appropriate treatment by the attendant or the first responder, during urgent situations, complete the efficiency and efficacy of the system.

The designed semantically enriched Web-services, assure the continuity of the care of the patient by including, software-tools for the creation and the management of a post-discharge Care-plan, and by providing pharmacovigilance software supporting it.

Finally, the employment of software supporting virtual audiovisual distal supervision of the Continuity of Medical Care, constitutes an innovative application, concerning sensitive, high-risk groups and it enables transmitting audiovisual data, transferring primary (raw) and secondary (processed) data to a remote (distal) terminal, transmission of bedside images and sounds, and adequate safety of the transferred data.

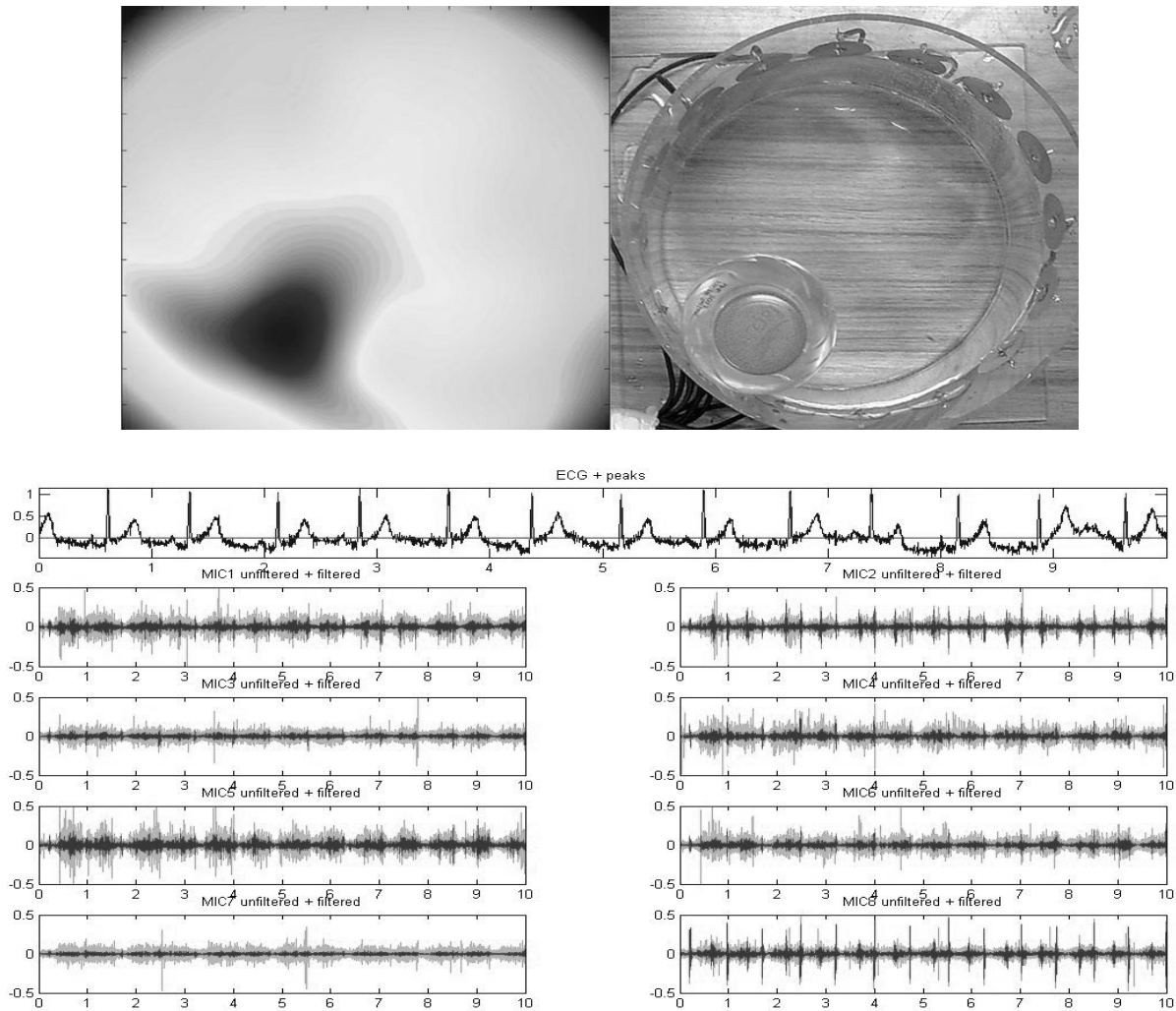


Figure 2. The present status of development of the combined recording of the distribution of thoracic Impedance by means of an array of electrodes, multiplexed in pairs to high-frequency voltage (upper part, image reconstruction taken in a saline solution phantom simulating the patient) and measurements of Respiratory and Pulmonary Sounds by means of a microphone-array (lower part, recorded in vivo on one of the investigators).

5. CONCLUSIONS

The presented, partially already developed, system has not yet been accomplished elsewhere, as reflected in the Industrial Property Documents, thoroughly presented in the introduction that constitutes legally the “state of the art” today. It includes a lot of individual, innovative aspects, as first, the development of a complete framework to support Continuity of Medical Care that lends itself to become a prototype for the commercial providers, second, the development of Continuity of Care software complying to the requirements of the major standards ANSI E2369 (CCR), ISO 13606-1 and pre-EN 13940, promoting the improvement of the interoperability between existing systems, and finally, the interim bridging of their lack in various Health Systems worldwide, especially in less developed countries.

The combined employment of new hardware approaches, allowing for, first, the recording of Respiratory and

Pulmonary Sounds by means of a microphones-array, second, the recording the distribution of thoracic Impedance, by means of an array of electrodes, multiplexed in pairs to high-frequency voltage, and finally the possibility of gas sampling and testing from the patient’s breathing circuit, offer the possibility to an essential improvement, of the lungs-condition monitoring of home-care patients, depending on mechanically supported ventilation.

The development of Medical Decision Support Software, supporting important Continuity of Care issues, constitutes a further innovation, because it allows for, first, an automatic assessment of the need for Automatic External Defibrillation, and second, an optimization of patient’s supported Respiration, by employing appropriate settings, thus, contributing to a reduction of patient mortality, during home-care or similar treatment conditions.

The acquisition, processing, storage and transmission of in vitro Diagnostic profiles tested at the Point of Care (PoCT) and the retrieval and employment of Emergency Medical

Protocols and Guidelines, supporting the appropriate treatment by the attendant or the first responder, during urgent situations, provide further important “medical tools” to the general physician in charge, especially in remote and isolated regions.

The planning of semantically enriched Web-services, assures the continuity of the care of the patient by including first, software-tools for the creation and the management of a post-discharge Care-plan, and second, Pharmacovigilance software supporting continuity of Care. This is achieved by the employment of an HL7-CDA compliant ontology, and appropriately designed Web-services that will deliver the demanded documents, by localizing appropriate instances of this ontology.

The provided software-tools, supporting adequately safe virtual audiovisual distance monitoring of the Continuity of Medical Care, constitute an innovative application, concerning sensitive, high-risk groups and it enables, first, transmitting of audiovisual data, second, transferring primary (raw) and secondary (processed) data to a remote (distal) terminal, and finally the transmission of bedside images and sounds, in various situations, starting from an emergency “telemedicine link” in an accident case to the optical contact of a premature baby to its parents at home, during its stay in a neonatal ICU.

Concluding, it should be kept in mind, that the presented system constitutes, among others, a transient, interim solution, improving the interoperability between various EHR-systems, whenever already present, and partially covering their absence. The creation of structured records, comprising of the most important medical-data of the patient, can be used for post-hospital discharge, improving the conditions of continuity of Health-Care, even in absence of an EHR-system.

Continuity of Care Software and the semantically enriched Web-services, will allow for in the very near future, both, a solution overcoming interoperability barriers, and an alternative approach to the inaccessible at the time Patient Record data. Finally, the Continuity of Care data, accumulated in appropriate Data and Knowledge-Bases, could gradually support Clinical and Epidemiological Research, might shape the Clinical paths followed by evidence based Medicine, and in an “ossified” form, as Clinical Guidelines and Protocols, would in some extend improve, through feed-back, cotemporary Medical Care.

REFERENCES

- [1]. The European Patent Office search-engine esp@cenet at ep.espacenet.com
- [2]. WO 2011027006 (A1).
- [3]. WO 2011022178 (A2).
- [4]. US 2010280350 (A1).
- [5]. CN 101764857 (A).
- [6]. US 2009254361 (A1).
- [7]. US 2009216558 (A1).
- [8]. US 2009171692 (A1).
- [9]. KR 20090001730 (A).
- [10]. US 2008215372 (A1).
- [11]. US 2006116911 (A1).
- [12]. US 2004210458 (A1).
- [13]. US 2005027569 (A1).
- [14]. <http://www.eurorec.org/services/standards/standards.cfm>
- [15]. ASTM, at www.astm.org: E2369-05, Standard Specification for Continuity of Care Record.
- [16]. ASTM, at: www.astm.org: ADJE2369- Adjunct to E 2369 Continuity of Care Record (CCR).
- [17]. ISO13606-1:2008. Health Informatics, EHR communication, Part 1: Reference model at: <http://www.iso.org/>
- [18]. prEN 14463 ClaML: A syntax to represent the content of medical classification systems.
- [19]. B. Spyropoulos, A. Tzavaras, M. Botsivaly, K. Koutsourakis, Ensuring the Continuity of Care of Cardio-respiratory Diseases at Home: Monitoring Equipment and Medical Data Exchange over Semantically annotated Web Services, *Methods of Information in Medicine*, 2010; volume 49; issue 2, pp.156-160.
- [20]. B. Spyropoulos, M. Botsivaly, A. Tzavaras, P. Spyropoulou, Towards Digital Blood-Banking, ITU-T Proceedings of the Kaleidoscope Conference, Mar del Plata, Argentina, 31 Aug - 1 Sep 2009.
- [21]. B. Spyropoulos, Smart Health Care in the city of the future: Patient treatment at home and medical data exchange in the emerging networked society, ITU-T Workshop “ICTs: Building the green city of the future”, United Nations Pavilion, EXPO-2010, 14 May 2010, Shanghai, China.
- [22]. B. Spyropoulos, E. Oikonomi, A. Danelakis, K. Karaboulas, E. Kotsiliti, E. Maridaki, L. Papageorgiou, E. Papalexis, C. Sakellarios, D. Zogogianni and M. Botsivaly, A web-based System supporting the Certification of the Outpatient and Emergency Departments and providing for post-discharge Continuity of medical Care Software, IFMB Proceedings (EMBEC September 14-18 2011, Budapest, Hungary), Springer 2011.
- [23]. B. Spyropoulos, E. Oikonomi, M. Botsivaly, Software supporting the Certification of an IVD-Point-of-Care Testing service according to ISO-15189 and ISO-22870 and its linkage to an ASTM-E2369-05 Continuity of Care Record, Proceedings of the AMIA Annual Symposium 2011, Washington, DC, October 22-26, 2011.
- [24]. A. Tzavaras, P. R. Weller, G. Prinianakis, A. Lahana, P. Afentoulidis, B. Spyropoulos, Locating of the required Key-Variables to be employed in a Ventilation Management Decision Support System, Proceedings of the 33rd Annual International IEEE EMBS Conference, August 30 - September 3, 2011, Boston, MA, USA.
- [25]. G. Angelopoulos, C. Tsigkas, A. Tzavaras, B. Spyropoulos, Digital Multiplexer supported scanning Data Collection Method to be employed in Electrical Impedance Tomography simulation measurements, On-line Proceedings (EPOS) of the 9th Annual Scientific Meeting of the European Society of Cardiac Radiology (ESCR 2010), October 28-30, 2010, Prague, Czech Republic.
- [26]. P. Afentoulidis, A. Tzavaras, B. Spyropoulos, Development status of a low cost system quasi-visualizing pathological alteration of the lungs based on their acoustic response modification, On-line Proceedings (EPOS) of the 10th Annual Scientific Meeting of the European Society of Cardiac Radiology (ESCR 2011), October 27-29, 2011, Amsterdam, NL.

COEXISTENCE OF A TETRA SYSTEM WITH A TERRESTRIAL DTV SYSTEM IN WHITE SPACES

Heejoong Kim, Hideki SUNAHARA, Akira KATO

[†]Graduate School of Media Design, Keio University, 4-1-1, Hiyoshi, Kohoku-Ku, Yokohama, Kanagawa, Japan
heejoong@kmd.keio.ac.jp, suna@wide.ad.jp, kato@wide.ad.jp

ABSTRACT

In this paper, we have investigated the possibility of coexistence of terrestrial truncated radio (TETRA) as a narrow band system with digital television (DTV) in TV white spaces. Based on the system operation mode of a TETRA system, trunked mode operation (TMO) and direct mode operation (DMO), the interoperable power range of fixed and mobile terminals is obtained according to their frequency offsets and power classes.

Keywords—White Spaces, DTV transition, PPDR, TETRA, Interference

1. INTRODUCTION

In general, digital television (DTV) conversion presents both opportunities and challenges for the people who want to develop new wireless services via TV bands. The spectrum in the terrestrial broadcast service bands below 1 GHz is of particular interest because of the favorable coverage costs compared to the higher frequencies that are currently used to provide wireless services, such as mobile internet and wireless fidelity (Wi-Fi). In particular, broadcasters for terrestrial service usually provide their service not for mobile use but for fixed use. In this situation, it is easier to introduce new regional services that have coexisting circumstances, such as time and space.

In the near future, many countries, such as the UK, Korea, and Japan, are planning to switch analog television off and reallocate the television spectrum for DTV service. After that, spatially unused frequency resources within the DTV spectrum, called white spaces or the interleaved spectrum, will be generated and it could be reused for industry promotion and public interest with spectrum sharing techniques [1]. With these techniques, lots of wireless services will be open for universal purposes like a wireless internet services and local information services in some specific areas [2]. Developers in the US are planning to use white spaces as unlicensed bands for commercial, personal, and public safety services [3], [4]. Similarly, in Japan, they conducted research of white spaces in terms of service development and they focused on the revitalization of the local community, including use for public safety services through white spaces [5].

Regarding to public safety service, in WRC-2003, some spectrums were assigned not for global utilization but regional utilization because of spectrum scarcity [6]. However the public safety communication service should be connected universally in order to cope with global disaster. In this situation, white spaces could be a good solution as global services for public safety.

So far most of researches about utilizing white spaces have focused on cognitive radio (CR) [7], [8]. However, CR has some problems to be solved, such as hidden node, transmit-power control, etc [9]. Moreover it will take time to be commercialized and verified in the system operation [10]. On the other hand, a conventional TETRA system could instantly use white spaces with a little system modification, when amending the rules for them. In order to boost white space utilization, it is one of the most efficient methods to allow existing systems to use white spaces. To this end, coexistence analysis between DTV and existing systems should be preceded. However, there have been few researches to review the use of white spaces by existing systems. Therefore, we have investigated the possibility of coexistence of a TETRA system with a terrestrial DTV system in white spaces.

The rest of this paper is organized as follows. In Section II, a TETRA system for public protection and disaster relief (PPDR) service is described. The used interference analysis model including scenarios is given in Section III. Interference assessment results according to the TETRA operation modes, trunked mode operation (TMO) and direct mode operation (DMO), are presented in Section IV. Finally conclusions of our work are given in Section V.

2. TETRA SYSTEM

The main purpose of the PPDR service is to protect against and relieve risky situations like a natural disasters or accidents. In this regard, ITU-R classifies the systems in two ways, public protection (PP) and disaster relief (DR) [11]. Regarding this, the regional frequency distribution was decided at the World Radiocommunication Conference 2003 (WRC-03) [12]. For PPDR technology, study groups of ITU-T are working on technical standardization for emergency communications to develop effective and comprehensive standards [6].

Table 1. Parameters for White Space Detection [13]

Parameter	US value	UK value
DTT sensing	-114dBm	-120dBm
Wireless microphone sensing	-114dBm	-126dBm
Location accuracy	50meters	100meters
Transmit power - adjacent channels	40mW	2.5mW
Transmit power - non-adjacent channels	100mW	50mW
Out-of-band powers	55dBc	<-46dBm
Time between sensing	1minute	1second

TETRA is a set of standards developed by the European telecommunications standards Institute (ETSI) that describe a common mobile radio communications infrastructure. TETRA was designed for reliable, spectral efficient and safe voice communications and data transmission.

In Korea, the telecommunications technology association (TTA) is working on standards for PPDR and TETRA Release 1 is adopted for the PPDR standard. Although TETRA Release 2 has been issued recently, we have used the TETRA Release 1 standard in analysis [6].

In order to use TV white spaces, TETRA should adopt one of the following three detection methods which determine whether spectrum is free. Related parameters are given in Table 1.

- Sensing: Judge the empty space by using spectrum sensing technology, such as cognitive access.
- Geo-Location Database: Obtain the empty space by using the database by an authorized office
- Beacon Reception: Tune cognitive devices to the beacon channel to get channel information

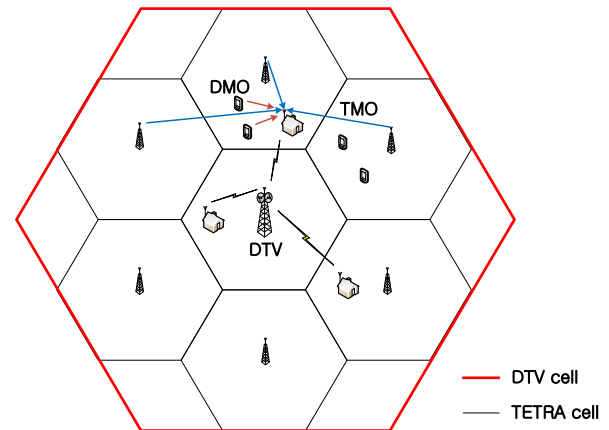
While the sensing and beacon reception methods requires extensive modification of a TETRA system, the geo-location database method allows a TETRA system to used in whitespaces without modification.

The TETRA system is the digital wireless communication system using 4-slot time division multiple access (TDMA) in 25 kHz bandwidth. Each carrier provides four independent physical channels. The $\pi/4$ -DQPSK modulation scheme was chosen to support a gross bit rate of 36 kbps, which means that net data rates up to 28.8 kbps can be offered to some data applications. There are two operating modes in the TETRA standard [14].

- Trunked Mode Operation (TMO): TETRA V+D enables basic voice and data transmission in a circuit switched mode using the network infrastructure.
- Direct Mode Operation (DMO): enables direct mobile-to-mobile communication without the support of the network infrastructure and also mobile-to-repeater communication for a communication range extension.

TMO transmitters are fixed in location and DMO transmitters move freely in space. In order to take these

characteristics into account, we have used the deterministic model and random models, respectively, in calculating interference power from TMO and DMO transmitters.

**Figure 1.** Interference Scenario

3. INTERFERENCE ASSESSMENT

3.1. Scenario

A scenario for analyzing interference is depicted in Fig. 1. A DTV receiver in a cell boundary may be most significantly affected by TETRA systems. There are two kinds of sources of interference with the DTV receiver: TETRA base station (BS) and mobile station (MS). In order not to compromise the reliability and functionality of the DTV system, the TETRA frequency bands for downlink communications are usually planned in the frequency range. However, TETRA DMO transmitter locations are distributed randomly, so the network planning approaches used in cellular systems are not applicable. Thus, it is very important to analyze the interference power of TETRA transmitters in white spaces.

The assumptions made here are:

- TETRA TMO transmitters are centered in hexagonal-shaped cells, and TETRA DMO transmitters are uniformly or randomly distributed over their cell areas (in two dimensions).
- The path loss is proportional to $d^{-\gamma}$, where d is path length and $\gamma = 3.3$.
- TETRA transceivers operate on an adjacent DTV channel with a 25 kHz or 50 kHz offset and the DTV is affected by TETRA transceiver leakage powers with -55 dBc.
- The average power levels from multiple interferers seen by the DTV receiver are additive

3.2. Simulation

As shown in Fig. 1, TETRA BSs are located at the center of the cells and MSs are uniformly distributed within the cell area. The total received power from TETRA transmitters, then the total receive interference is

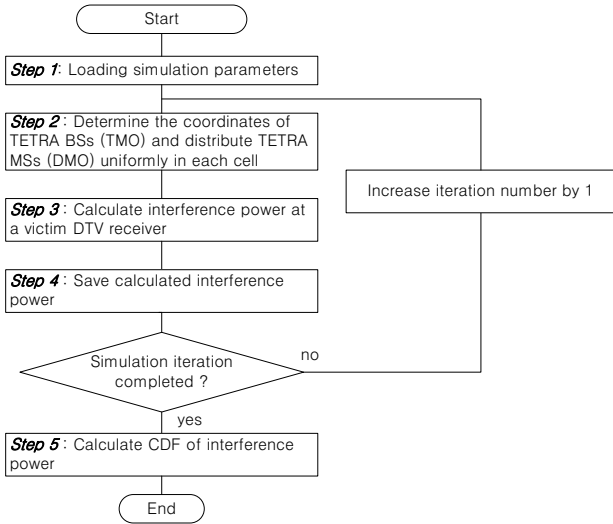


Figure 2. Simulation Flow

$$I = \sum_{j=1}^J \alpha_j d_j^{-\gamma} \quad (1)$$

$$I = I_{TMO} + I_{DMO} \quad (2)$$

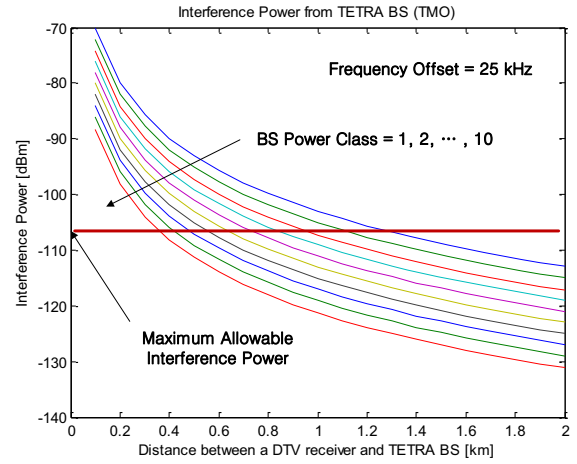
where α_j is the transmit power at the j th interferer, d_j is the distance between the DTV receiver and the j th TETRA device, and J is the number of interferers. Interference power in (1) is decomposed of deterministic values due to TETRA TMO and random values due to DMO.

The simulation flow to calculate interference power from TETRA devices is depicted in Fig. 2. First, various parameters about DTV and TETRA systems are loaded to the simulation program. Secondly coordinates of TETRA BSs and MSs are determined and then aggregated interference power is calculated in step 3. If simulation iteration is not completed, the step 2 and 3 are repeated. After simulation iteration is completed, cumulative density function (CDF) of interference power is finally calculated. Simulation parameters for calculating interference power from the TETRA system are shown in Table 2. The DTV

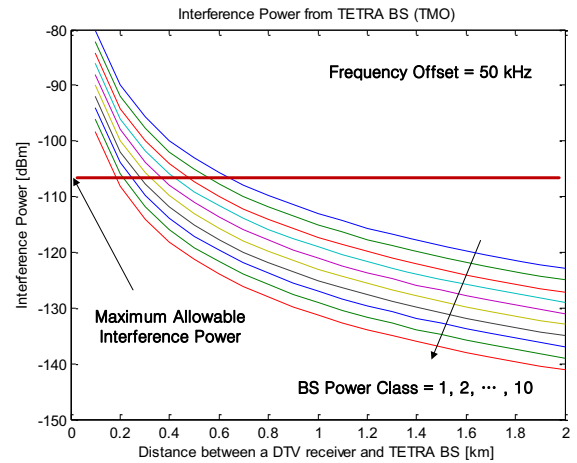
Table 2. Simulation Parameters

Parameter	Value
DTV Center Freq.	470 MHz
DTV Bandwidth	6 MHz
Required minimum field strength at a DTV receiver	38.7 dB μ V/m (-92 dBm @ 470MHz)
DTV protection ratio	15.5 dB
Maximum allowable interference power ratio	-107.5 dBm
Frequency offset	25, 50 kHz
TETRA transmit power	Table 3

system occupies a bandwidth of 6 MHz assigned with a center frequency of 470 MHz. The frequency offset means the difference between the center frequencies of DTV and TETRA. The required minimum field strength at a DTV receiver is 38.7 dB μ V/m, which is equal to -96 dBm at 470 MHz. As DTV protection ratio is 15.5 dB, then the maximum allowable interference power is -107.5 dBm (= -92 dBm - 15.5 dB). For interference calculations, the minimum distance between the TETRA MSs and the DTV receiver is assumed to be 10 m.



(a) Frequency Offset: 25kHz



(b) Frequency Offset: 50kHz

Figure 3. Interference Power from TETRA BS

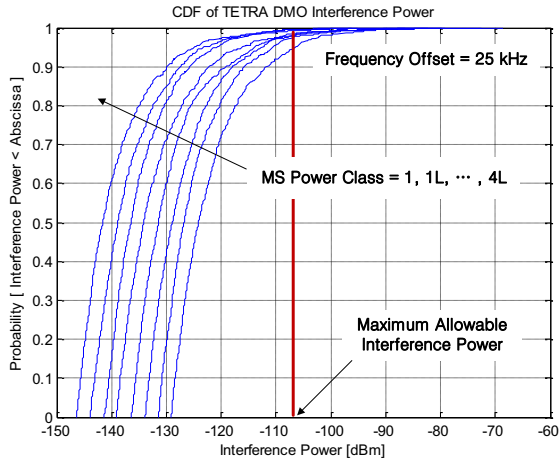
3.3. Results

Fig. 3 shows the interference power from the TETRA BS according to the transmit power class given in Table 3.

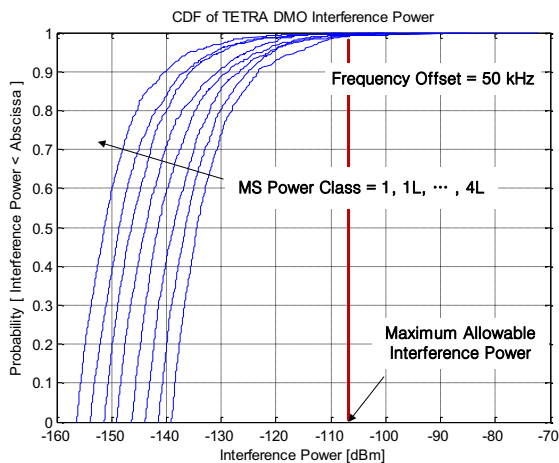
The horizontal axis represents the distance between a DTV receiver and the TETRA BS, the vertical axis interference power. The red line indicates the maximum allowable interference power. If the received interference power at a DTV receiver is below this line, it is possible for the TETRA BS to use white spaces. Fig. 3(a) shows that a 1.4 km separation distance between a DTV receiver and a TETRA BS is required in order not to interfere with DTV reception in all BS power classes. In the case of the 50 kHz

frequency offset shown in Fig. 2(b), the separation distance is reduced to below 0.7 km.

CDF graphs of interference power from TETRA MS are shown in Fig. 4. Fig. 3 shows that the possibility of interference power from the TETRA MS is less than the maximum allowable interference power. In both 25 kHz and 50 kHz frequency offsets, the interference probability is below 5 % in all TETRA MS power classes.



(a) Frequency Offset: 25kHz



(b) Frequency Offset: 50kHz

Figure 4. CDF of Interference Power in TETRA MS

Table 3. Nominal Power Transmitters [13]

Base Station(BS)		Mobile Station(MS)	
Power Class	Nominal Power	Power Class	Nominal Power
1(40W)	46dBm	1(30W)	45dBm
2(25W)	44dBm	1L(17.5W)	42.5dBm
3(15W)	42dBm	2(10W)	40dBm
4(10W)	40dBm	2L(5.6W)	37.5dBm
5(6.3W)	38dBm	3(3W)	35dBm
6(4W)	36dBm	3L(1.8W)	32.5dBm
7(2.5W)	34dBm	4(1W)	30dBm
8(1.6W)	32dBm	4L(0.56W)	27.5dBm
9(1W)	30dBm	N/A	
10(0.6W)	28dBm		

4. CONCLUSION

In this paper, we have considered applied the TETRA system to white space as a narrow band system for public safety service. In these regard, we simulated the interference power according to scenarios to investigate the possibility of coexistence with DTV. The simulation results confirmed that TETRA systems may share white spaces with DTV provided the separation distance between a DTV receiver and a TETRA BS is sufficient and the MS power class meets the maximum allowable interference power level at the DTV receiver.

Therefore, we expect that white spaces will be used in a wide range of applications by considering not currently reviewed systems, such as wireless regional area network (WRAN) [15], [16], but existing applicable system, such as TETRA. In addition, the public interest and economic improvement will be achieved with enhanced frequency efficiency. Our results demonstrate that white space utilization is plausible as a model for increasing spectrum efficiency. However, at present, the Korean government is conducting a pilot test of DTV transition and it would be completed by the end of 2012. Therefore, field tests should be conducted to complement the hypothesis of this paper.

REFERENCES

- [1] TTA, "ICT Standardization Roadmap 2009: USN", Synthesis report, 2009..
- [2] I. F. Akyildiz, W. Lee, M. Vuran, and S. Mohanty, "NeXt generation / dynamic spectrum access/cognitive radio wireless networks:A survey," ACM ComNet, 2006.
- [3] FCC, "First Report and order and Further Notice of Proposed Rulemaking", Docket No. FCC 06-156, October 2006.
- [4] F. L. Martin, S. C. Correal, R. L. Ekl, "Early Opportunites for Commercialization of TV Whitespace in the U.S.(invited paper)", CrownCom 2008, May 2008.
- [5] MIC, "For the Use of White Space and a New Wave", http://www.soumu.go.jp/main_content/000077004.pdf, September 2010.

- [6] TTA, "ICT Standardization Roadmap 2010: PPDR", Synthesis report, 2010.
- [7] M. Nekovee, "Quantifying the Availability of TV White Spaces for Cognitive Radio Operation in the UK", ICC Workshops 2009. June 2009.
- [8] D. Prendergast, Y. Wu, G. Gagnon, "The Effective of Unlicensed Cognitive Device Operation on Digital Television Performance in the VHF/UHF Band", RWS 2008, January 2008.
- [9] M. S. Song, G. Z. Ko, S. H. Hwang, "Standardization of CogNeA on TV White Spaces", ISCIT 2009, September 2009.
- [10] E. Obregon, L. Shi, J. Ferrer, "Experimental Verification of Indoor TV White Space Opportunity Prediction Model", CrownCom 2010, June 2010.
- [11] N. K. Lee, H. K. Kim, D. G. Oh, "Research on the Public Protection and Disaster Relief", ETRI Report, vol. 21, August 2006.
- [12] ITU-R, "Resolutions and Recommendations", vol. 3, 2008.
- [13] Ofcom, "Digital dividend: cognitive access", July 2009.
- [14] ETSI, "Terrestrial Trunked Radio(TETRA); Voice plus Data(V+D); Part 2: Air Interface(AI)", EN 300 392-2, V3.2.1, September 2007.
- [15] Andrew Stirling, "White Spaces – the New Wi-Fi?", international Journal of Digital Television, vol. 1, November 2009.
- [16] Liang, Y-C et. Al, "Cognitive radio on TV band: A new approach to provide wireless connectivity for rural areas", IEEE Commun. Mag. Vol.15, June 2008.

MOBILE CLOUD COMPUTING BASED ON SERVICE ORIENTED ARCHITECTURE: EMBRACING NETWORK AS A SERVICE FOR 3RD PARTY APPLICATION SERVICE PROVIDERS

Michael Andres Feliu Gutierrez, Neco Ventura

Communication Research Group
Broadband Networks and Applications
Department of Electrical Engineering
University of Cape Town, Rondebosch, 7700
Email: {mfeliu, neco}@crg.ee.uct.ac.za

ABSTRACT

The recent emergence of Cloud Computing in the IT world has opened doors for new revenue streams and business models. The movement towards a service-oriented architecture and an all-IP based communication system has led to the IP Multimedia Subsystem (IMS) being accepted as the Next Generation Networks (NGN) service control-provisioning platform. The Telco 2.0 domain can benefit from service enhancement and Web 2.0 technologies such as Cloud Computing. This can be achieved by opening gateways and APIs guarding rich underlying network resources (e.g. location) to 3rd Party ASPs, thus adopting the delivery of Network as a Service (NaaS). Telco services are faced with opportunities to make use of powerful computational power and storage services offered by cloud environments to accelerate business-processing speeds.

Keywords— Cloud Computing, NGN/IMS, *aaS, Web 2.0, OMA PEEM enabler

1. INTRODUCTION

Over the years the telecommunication industry's main revenue stream has come from voice applications, in recent years the mobile operators have been seeking for new revenue streams and business models.

There have been debates on whether mobile and broadband communication should be provided as a commodity. It is stated in [1] that the IT and Telecommunication industries are undergoing transformations of their infrastructure with the intentions to deliver services in a way similar to traditional utilities like electricity, water and gas.

The Australian government began a project [2] that will last 10 years with the intentions to deliver computing as a 5th utility. Computing is brought to your household and users consume in a "pay as go" basis.

On the other hand Cisco's Global Mobile Data Traffic Forecast [3] has predicted an overall growth of 6.3 Petabytes per month by 2015. This excess growth of data

traffic as can be observed in Figure 1 below [3] is mainly due to the mobile world with user equipment such as smartphones and laptops.

Mobile video and mobile data will contribute to almost 90% of the mobile traffic growth, with video due to its high bit rates contributing 66%. This tremendous increase in data flow brings challenges to the mobile operators (e.g. how to cope with gigantic storage and enormous amount of processing).

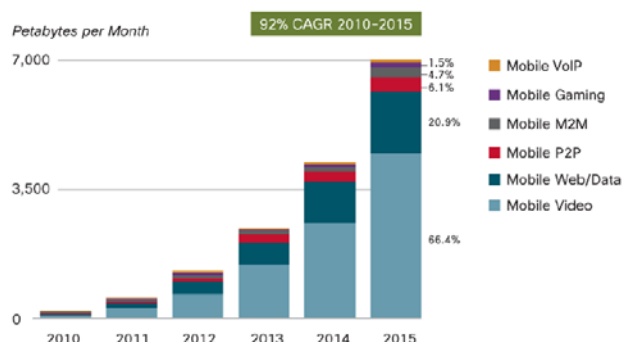


Figure 1. Showing the major contributors to the mobile traffic growth.

It can be seen that today there is an increasing demand for multimedia services and sophisticated applications, which in turn brings together internet applications with telecommunication applications. This gives the telecommunication operators great opportunities to benefit from the emerging Web 2.0 technologies. Simple protocols like HTTP, HTML description language and SOAP/REST web-based services have driven the developers community in creating sophisticated content-aware applications.

The Telco industry has to adapt to the customer's needs by upgrading their systems, installing new hardware/software and integrating new technologies into their existing infrastructure. Keeping in mind that one of the key aspects of NGN telecommunication platforms is the reuse of its

existing infrastructure for new market driven applications through dedicated application enablers, thus sharing operational expenditure across the whole architecture. This paper discusses a service creation mechanism with strong support for creating 3rd Party ASPs value-added services by building on computation and storage in a cloud environment and application services in NGN/IMS environment. Section II describes cloud computing in both worlds with the different business models and service delivery adaption. Section III discusses a NGN service-oriented architecture with focus on the IMS Application Servers (A.S.) for service delivery. Section IV proposes a possible integration scenario for a mixed Cloud/IMS environment. Section V presents some results obtained during the integration analysis. Section VI concludes this paper and states some final remarks.

2. MOBILE CLOUD COMPUTING

In the last decade new technologies have emerged, one of particular interest and gaining popularity in the Web 2.0 domain is cloud computing. The term Web 2.0 was first used by Tim O'Reilly [16] to describe a fast-growing set of web-based applications.

Cloud Computing is a concept that is currently enjoying much hype and its exact definitions is open to interpretation. Generally speaking it refers to the sharing of some kind of hosted service or application. It can be defined as a paradigm shift in the IT world where computing resources (e.g. processing power, storage) are moved away from the user's personal computer or application server to a cloud of powerful computers [1, 4].

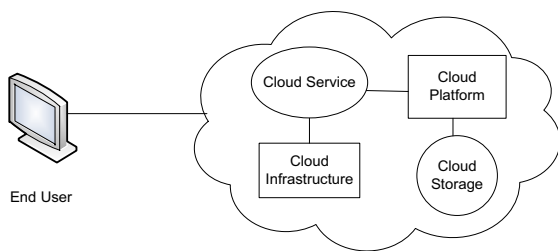


Figure 2. Showing the benefits of cloud computing

Cloud computing [6] can supply transparent and on-demand access to applications served over the Internet in a dynamic and scalable manner. It adapts the delivery of various classes of services.

2.1. Software as a Service (SaaS)

Software as a Service can be viewed as an improved version of the Application Service Provider (ASP) model in which a service provider host software applications over the network. The service is hosted/ located on a cloud environment where it can be accessed on demand by a UE anytime anywhere. SaaS can be categorized under Business-

oriented SaaS (e.g. Salesforce, Amazon S3) and Consumer-oriented SaaS (storage, Facebook).

2.2. Platform as a Service (PaaS)

Platform as a Service are hosted application environments where software developers/ 3rd party ASP who want to focus primarily on the service development cycle and the orchestration of new services, can bypass the capital expenditure that would otherwise be needed for the deployment of the underlying infrastructure. PaaS examples include the Google App Engine where developers can enhance an existing SaaS by providing mash-ups or simply develop new web-based applications.

2.3. Infrastructure as a Service (IaaS)

Infrastructure as a Service can be described as utility computing data centers that make use of cluster virtualization technology to provide powerful and flexible computing resources. IaaS provides on demand resources such as parallel computing power to process large amount of data and virtual storage to store gigantic data. Examples [5] include Microsoft Windows Azure and open source EUCALYPTUS and Nimbus.

2.4. Network as a Service (NaaS)

Network as a Service is a term used in the Mobile world [8] referring to Telco operators opening up network APIs to expose network capabilities (e.g. presence, location, billing and charging). These network resources are exposed in a standard and flexible manner to 3rd Party ASPs where they can mash-up Telco services with IT services and vice versa.

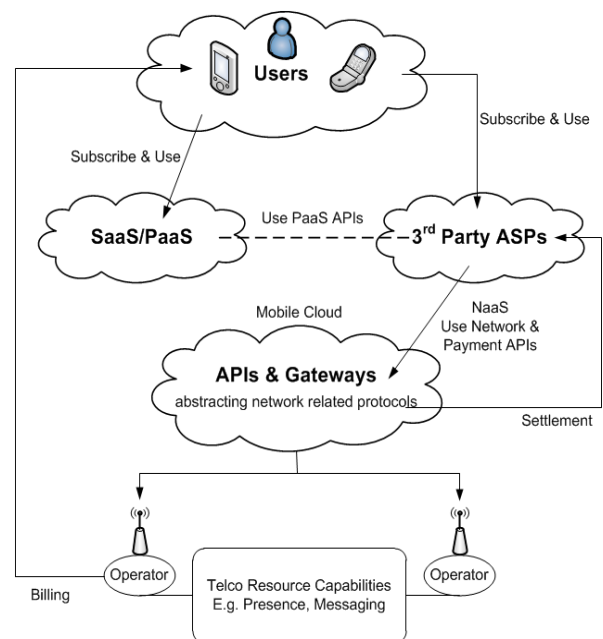


Figure 3. Showing the concept of NaaS

Telco operators can benefit from the delivery of network resources as services and give rise to the Telco 2.0 two-sided business model. The advantage would not only be revenue from their users but also from 3rd Party ASPs who use network resources to build applications and may adapt a pay as you go basis depending on the SLA.

The cloud architectures extend to the end user by allowing web browsers and software applications residing inside/outside the cloud domain to access cloud resources. The two major forms of cloud resources are computing clouds and storage clouds.

3. SERVICE ORIENTED ARCHITECTURE

The European Telecommunication Standards Institute (ETSI) defined the Internet Protocol Multimedia Subsystem (IMS) in [7] as an overlay network where the service logic has been stripped from the transport and network protocols to facilitate and standardize the service delivery mechanism.

NGN supports end-to-end services, where the QoS is independent of the underlying networking technology. Interoperability among multiple network service providers is facilitated with such architecture. A SOA thus facilitates the reuse of software and enables the creation of composite services with intent to write-once, run and sell anywhere goal.

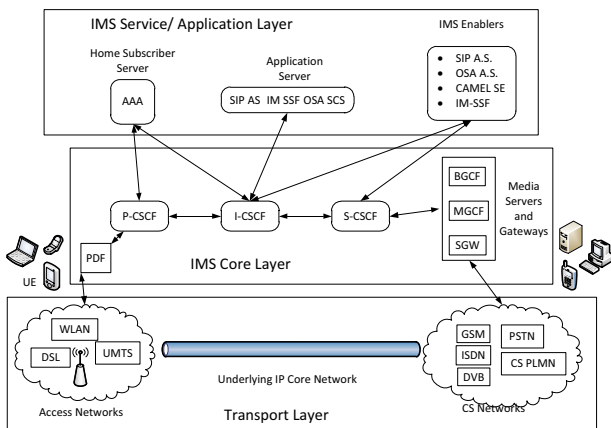


Figure 4. Showing the NGN/IMS reference architecture

3.1. Application Layer

The service logic is implemented in the A.S and is thus the service relevant part of the IMS. Programmer's are not limited to their favorite programming language, common means of implementation include C++, SIP applications, Java. A.S are considered value-added services, which can be used to enhance or mash-up already existing services. With the use of SIP application on top of SIP servers allows four mode of operation.

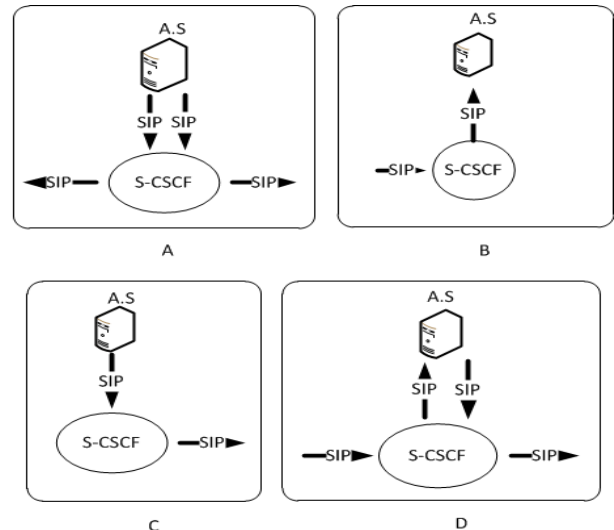


Figure 5. Showing A.S mode of operation

- A – A.S acting as a Back-to-Back UA
- B - A.S acting as a terminating UA
- C - A.S. acting as an originating UA
- D - A.S. acting as a SIP proxy

The programming flexibility and the movement towards a service-oriented programming model can be illustrated in Figure 5 which shows the different roles an A.S. can have and the different outcomes in terms of SIP signaling.

The IMS was developed as the next generation core architecture for converged voice and data services, which provides a common platform for different access technologies like Wi-Fi, Wi-Max, and DSL and aims to supply an open, standard based network that grants a wide range of multimedia services. This motivated the use of the IMS for this project in order to facilitate the deployment of services which not only uses IMS related capabilities (e.g. IM, presence) but also resources (processing, storage) from other networks residing outside the operator's network.

3.2. IMS-related Entities and Functionalities

There are 3 main Call Session Control Functions, each with its own special task, all working together to handle SIP requests, registration and session establishment.

- Proxy-CSCF is the 1st contact/entry point into the IMS environment from the client UE, it's responsible for compressing data, interacting with policy and charging rules and providing security
- Interrogating-CSCF is an intermediate point whose function is to provide topology hiding by obtaining the next signaling hop, either an A.S or S-CSCF
- Serving-CSCF is the focal/anchor point in the IMS as it is responsible for handling registration, storing service profiles and maintaining session states

The IMS entities communicate via two main interfaces forming the central session control protocols defined by IETF.

1. SIP - Session Initiation Protocol is an end-to-end client-server session signaling protocol defined by the Internet Engineering Task Force (IETF), used for session creation and termination, as well as for providing services such as presence
2. DIAMETER is used to perform AAA operations Authentication, Authorization and Accounting as well as interfaces with the HSS database to download and upload user profiles

For clarity sake only two protocols were mentioned but it is important to note that the IMS entities communicates with a vast number of well defined interfaces [11] that are out of the scope of this paper.

The Home Subscriber Server (HSS) is the main data storage database for all subscribers and service-related data (e.g. memory and bandwidth requirements). Popular databases include MySQL where registration information and service information (e.g. target resources, memory requirement) can be easily stored and obtained.

4. NGN/IMS AND CLOUD INTEGRATION SCENARIO

NGN applications running on smartphones and netbooks will operate almost entirely off mobile cloud based services. This will be achieved with the promised downlink and uplink speeds of 4G. Over time, the mobile cloud may do for the mobile devices what SaaS has done for the desktop, connecting people, services and digital data and in a flexible and standardized manner.

In order to integrate the two technologies separate standalone environment had to be deployed and tested. The environments are limited to hardware limitation.

According to the ETSI [13] there are a number of possible integration scenarios that can be achieved, the benefits and impacts of these scenarios on the NGN platform are still in its infancy and it's an open area of research currently gaining hype in both worlds.

4.1. Cloud Computing Environment

An IaaS environment was deployed using the open-source [10] Elastic Utility Computing Architecture for Linking Your Programs to Useful Systems (EUCALYPTUS). It is a simple, flexible and modular cloud computing framework that uses computational and storage infrastructure. Resource access latency is greatly reduced by using Xen virtualization technology; it supports Virtual Machines that run on top of the Xen [10] hypervisor.

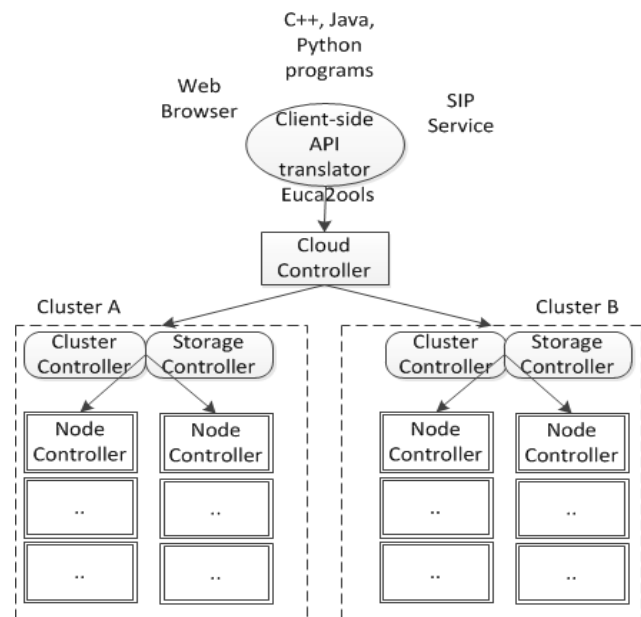


Figure 6. Showing eucalyptus architecture

The eucalyptus platform [7] is composed of the following main components

- Cloud Controller – is responsible for the management of the virtualized resources (e.g. storage, CPU instances) that are offered on demand. It is important to note that this controls the provisioning of these resources in the cloud domain, but how best to utilize these resources is up to the software developer/ 3rd Party ASPs
- Cluster Controller – is responsible for the controlling virtual machines/servers running on nodes
- Storage Controller – provides block-level storage that can be accessed on demand and can be dynamically encapsulated by virtual machines instances
- Node Controller – control underlying VM or instances by the use of a hypervisor
 - Execution, Management
 - Termination, Inspection

Xen [10] is the hypervisor of choice for controlling instances due to its compatibility with a Linux OS, where Eucalyptus is deployed.

In fact IaaS platform like Amazon place server clusters across the globe in order to improve the responsiveness, locality and redundancy of the content it hosts for users. However their price ranges are out of the reach of most and often require lengthy contract and a large usage commitment. For this reason a simple open source small-scale IaaS environment was deployed to take advantage of the processing and storage they offer.

NGN applications can be considered to be a mash-up of IT services using computational and storage resources

interacting with Telco resources capabilities (e.g. presence, messaging) in a flexible manner. These applications can be computationally demanding and the UE is limited to its inbuilt processing power and storage capacity. On demand access to storage and cluster computation can be made available through the IaaS environment where services deployed can access the resources in the cloud environment.

4.2. IMS Environment

The IMS platform is based on the FOKUS Open IMS, which has been deployed as a testbed for multimedia services at the UCT CoE. The IMS core was deployed on a Linux-Ubuntu kernel running on top of an Intel core2 duo CPU @ 2.66GHz and with 4GB RAM. All service related information is stored in a lightweight HSS (FHSS), which relies on MySQL; the database is thus populated using SQL scripts included in the IMS package.

The UCT open IMS Client [15] is a well-defined IMS end user client that communicates with the IMS entities and ultimately provides services to the customer in an abstract way. The IMS client can initiate an instant messaging services using SIP MESSAGE method, alternatively the Open Mobile Alliance (OMA) IM enabler [11] can be used which adds value to the already existing IM service by adding for example buddy list or possibly store messages.

IMS resources are expressed and interconnected via Uniform Resource Identifiers (URI), which is in the form of *sip:R1.resources@anIMSprovider.com*.

4.3. PEEM-based Serving Gateway (PSG)

The Serving Gateway sits at the border between the IMS and the Cloud environments. It acts as an intermediate logical entity, intercepting request from both worlds. It allows the exposure of IMS resources to 3rd Party ASP by acting as a SIP to HTTP protocol translator. Conversion is made simple due to the fact that SIP [12] borrows and adapts methods from HTTP [17]. It exposes web services and SIP applications to the managed, reliable and flexible IMS platform.

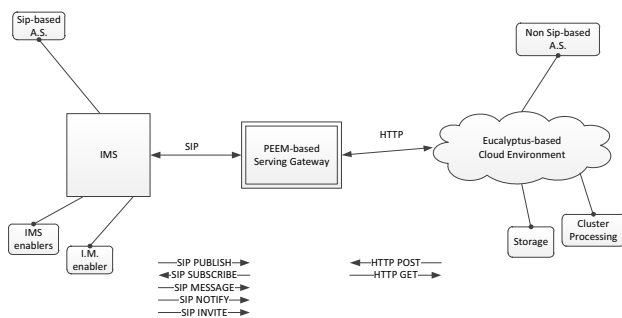


Figure 7. Showing the SG acting as an intermediate entity

The Open Mobile Alliance (OMA) standardization body has defined the Policy Enforcement Evaluation Management (PEEM) enabler. The OMA defines a number

of usage patterns for the PEEM enabler, which can be used for vast number of applications. It is important to note that PEEM proxy usage pattern [14] is the most relevant to this project as it can be used to intercept service requests from a foreign domain and from other service requester generated by 3rd Party ASP. Policies (rules) can be evaluated and enforced depending on the SLA. The PEEM can be used to expose network capabilities like presence and IM to the cloud domain. The Serving Gateway contains among others, a ContextHandler function where HTTP and SIP request are sent. It is here where rules are enforced or evaluated.

5. RESULTS

A change of media scenario between an IMS service and cloud service residing in the cloud domain is favorable. A simple use case was built around this idea to test the seamless switch between an IM session and a write/read session using storage instances provided in the cloud environment.

The size of a file usually determines the processing time required to store that file. Multithreading and parallel architecture achieved in a cloud environment allows the processing task to be sub-divided or split among independent threads or instances running concurrently. Depending on the SLA a client is allowed to use a certain amount of processing power usually on-demand access adapts pay as you go, thus you pay for exactly how much processing power you need.

```
VNET_DHCPSERVICE=/usr/sbin/dhcpd
VNET_PRIVINTERFACE=eth1
VNET_MODE=STATIC
VNET_SUBNET=137.158.125.0
VNET_NETMASK=255.255.255.0
VNET_BROADCAST=137.158.125.20
VNET_ROUTER=137.158.125.1
VNET_DNS=137.158.125.56
VNET_MACMAP=00:1D:92:60:2F:6A=137.158.125.251
VNET_MACMAP=00:1D:92:60:2F:6B=137.158.125.252
VNET_MACMAP=00:1D:92:60:2F:6C=137.158.125.253
```

Figure 8. Showing hypervisor snippet while testing instance availability in the cloud infrastructure deployed.

The snippet shows three IP addresses that are available for instances. A 3rd Party ASP can use 137.158.125.251-3 to obtain on demand access to three cluster processing instances.

The diagram below illustrates the adoption of two SLAs and shows the processing time where two clients with the same file size have different amount of processing power. Thus one client uses one instance of processing while the other uses three instances to execute the same task.

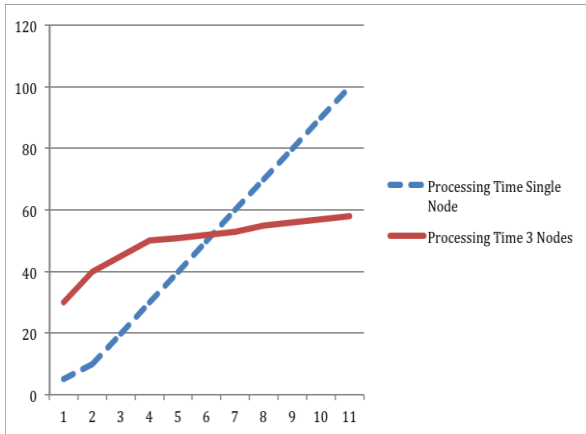


Figure 9. Showing the concept of on demand processing power depending on two different SLA

Consider an IM session where Alice generates a message consisting of text and fills the request-URI with the address of Bob. The IMS infrastructure (P/I/S-CSCF) forwards the message to the recipient. Once bob receives the message, he replies with a 200 OK message, as it can be observed in Figure 11 acknowledging the delivery. It is important to note that each instant message is an independent transaction and is not related to the previous requests. The protocol responsible for conveying messages within an instant messaging (IM) session is called Message Session Relay Protocol [18], which sits on top TCP and allows messages of arbitrary size, due to messages being sent in small chunks.

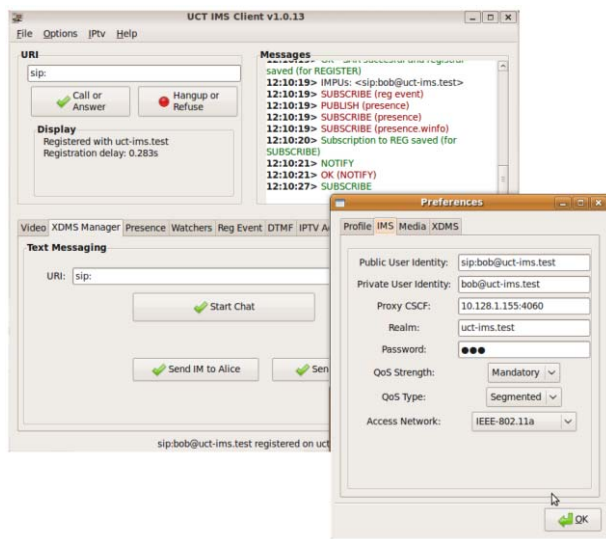


Figure 10. Showing the UCT IMS client instant messaging registration and configuration

To discover IMS resources a SIP SUBSCRIBE is sent by the SG and to discover cloud resources an HTTP GET can be sent by the SG. Alice later decides to access on demand storage from the cloud and give rights to bob to access her file (e.g. photo, video, letter). The SG is responsible for delegating services between the two worlds.

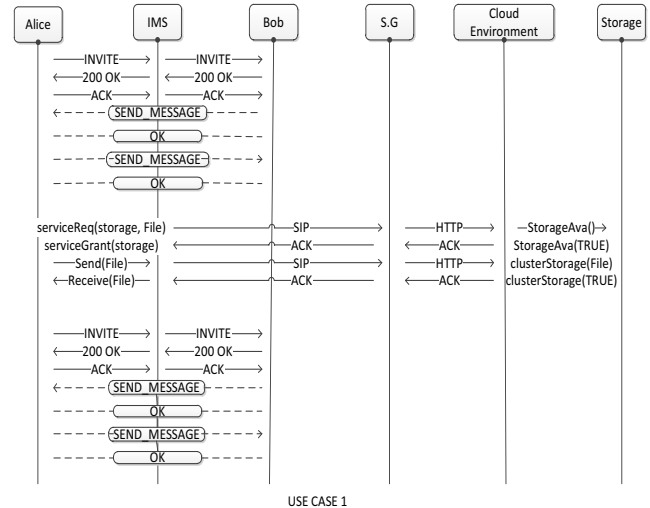


Figure 11. Showing signaling diagram for a simple use case

This use-case illustrates how we seamlessly expose the IMS presence and instant messaging done within the operator domain to access a storage instances provided by an external party on the Cloud infrastructure. This enables the provision of a consistent and efficient user experience, wherever the resource is stored and is network connection independent.

Currently many telecom service systems could be running in the “cloud”, which make the services easy to manage, update and use. It can also make full use of the powerful processing ability and storage.

6. CONCLUSIONS AND FUTURE WORK

In the Web 2.0 domain, cloud computing has adapted the delivery of IT service and infrastructure as computing utilities. Telco operators are expecting that cloud-enabled application can improve their internal network operation in terms of allowing 3rd Party ASPs to enhance their existing services or develop new innovative services by mashing-up the IT and Telecom services in a flexible and attractive manner.

This paper briefly presents a possible scenario in which cloud application (on demand storage access) can be merged with a Telco application (instant messaging using presence).

Resource allocation algorithms are responsible for delegating resources to service request depending on request need, availability and allocation, thus providing as much resources as possible while preventing deadlock stages.

A system in practice consists of a finite number of resources that are shared among a number of competing processes. Our next step is to develop a resource-allocation algorithm like the Banker's Algorithm for multiple instances of each resource type including and not limited to disk storage, processing instances, and memory allocation. The aim is to increase system utilization when resources limitations are faced due to perhaps peak time usage or abnormal increase in resource demand or simply service attractiveness.

REFERENCES

- [1] Rajkumar Buyya, Chee Shin Yeo, Srikumar Venugopal, James Broberg, Ivona Brandic, "Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility" *Future Generation Computer Systems Journal Elsevier* 2010
- [2] The Broadband Advisory Group's Report to Government. "Australia's Broadband Connectivity" Summary Document
- [3] Cisco White Paper, "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2010 - 2015" February 1, 2011
- [4] John W. Rittinghouse, James F. Ransome. "Cloud Computing Implementation, Management and Security" CRC Press Taylor and Francis Group, 2010
- [5] Asma BEN LETAIFA, Amel HAJI, Maha JEBALIA, Sami TABBANE, "State of the Art and Research Challenges of new services architecture technologies: Virtualization, SOA and Cloud Computing" *International Journal of Grid and Distributed Computing* Vol. 3, No. 4, December, 2010
- [6] Vânia Gonçalves, Pieter Ballon, "Adding value to the network: Mobile operators' experiments with Software-as-a-Service and Platform-as-a-Service models" *Telematics and Informatics Journals, Elsevier* 2010
- [7] ETSI ES 282 007, "Telecommunications and Internet converged Services and Protocols for Advanced Networking (TISPAN); IP Multimedia Subsystem (IMS); Functional architecture" V2.0.0 (2008-03)
- [8] AEPONA White Paper "Network as a Service and Mobile Cloud Computing." February 2010.
- [9] Fabricio Gouveia, Sebastian Wahle, Niklas Blum, Thomas Magedanz. "Cloud Computing and EPC/IMS Integration: New Value-Added Service on Demand" *Mobimedia '09: Proceedings of the 5th International ICST Mobile Multimedia Communications Conference*, September 2009.
- [10] Daniel Nurmi, Rich Wolski, Chris Grzegorzczak, Graziano Obertelli, Sunil Soman, Lamia Youseff, Dmitrii Zagorodnov, "The Eucalyptus Open-source Cloud-computing System" University of California, Santa Barbara
- [11] Miikka Poikselka, Georg Mayer. "The IMS – IP Multimedia Concepts and Services" Wiley Book 2009
- [12] IETF RFC 3261, "SIP: Session Initiation Protocol". 2002
- [13] ETSI TR 102 767, "Grid Services and Telecom Networks; Architectural Options" V1.1.1 (2009-02)
- [14] Open Mobile Alliance, "Policy Evaluation, Enforcement and Management Architecture" V 1.0.0 Aug 2008.
- [15] David Waiting, Richard Good, Richard Spiers and Neco Ventura, "The UCT IMS Client" Testbeds and Research Infrastructure for the Development of Networks & Communities and Workshops, TridentCom 2009
- [16] Tim O'Reilly, "What is Web 2.0" O'Reilly Media Release September 2005.
- [17] IETF RFC, 2616, "Hypertext Transfer Protocol - HTTP" 1999.
- [18] IETF RFC 4975, "Message Session Relay Protocol - MSRP" September 2007.

RBAC FOR A CONFIGURABLE, HETEROGENEOUS DEVICE CLOUD FOR WEB APPLICATIONS

Hannes Gorges¹, Robert Kleinfeld²

¹²Franhofer FOKUS, Berlin, Germany

E-Mail: ¹hannes.gorges@fokus.fraunhofer.de, ²robert.kleinfeld@fokus.fraunhofer.de

ABSTRACT

A key challenge during the development of Web applications on top of multiple heterogeneous devices is to discover and get access to device-specific resources.

This paper shows an architecture which enables and virtualizes the resources of user devices whereupon the user controls the access to his resources with a user-configurable rule-based system.

User resources are mapped to a RESTful API, so that Web applications can easily use them. In doing so, the communication protocols or APIs used by the underlying resource are completely hidden from the user of the RESTful API. This increases the interoperability for a looser coupling between the parts of distributed devices, because the user can replace resources with new ones without an update of the Web applications, which use these resources. This facilitates the creation of mashups, which combine traditional Web 2.0 services with resources from the user.

Based on this assumption this paper presents in detail a user-configurable rule-system for controlling access to user resources.

Keywords— Device Cloud, Access Control, Distributed Applications, Web of Things, Cloud Services, Rule-based Systems, Mobile Web Applications

1. INTRODUCTION

Today, there is a new distinct trend called Social Media. Kaplan and Haenlein described Social Media as “*a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, which allows the creation and exchange of user-generated content.*” [1]

The boundaries between local data storage and centralized data storage in the Web faded with Web 2.0 (and so with Social Media). Users can store their pictures on Flickr, their holiday videos on YouTube or their diploma thesis on the file storage and sharing service Windows Live SkyDrive. The same happens with native applications and Web applications. It is quite common for applications today, to download updates and additional modules from the Internet. Furthermore it gives Web applications for a host of problems and they achieve in some cases the complexity and quality of their native counterparts.

A widespread style of software architecture is the component-based style. This style emphasizes the

separation of concerns, the reusability and the maintainability, which decreases complexity from applications. Components are encapsulated. They provide an interface which specifies the offered services. Other components can utilize these services.

Web services are also components and many Web applications use them. To build a Web application with Web services decreases the development effort and increases the possibilities for a Web application. For example, the Web service Google Maps offers a great amount of data and functions to use the data (typically a map) in a Web site or in an application.

Web services are services, which provide data and functionalities via Web technologies. They commonly use the HTTP as transport protocol and XML or JSON as data-format. Web services which conform to the REST architecture [2] are prevalent nowadays. A Web service exists for almost every use-case. But one group of Web services is greatly underrepresented. Those are Web services which operate on device-specific resources, like camera, message services (for example SMS and MMS) or localization services like GPS. These resources are provided by the operating system of the device.

What are the reasons for this underrepresentation?

There are very different groups of user devices, which can be accessed direct or indirect via Web technologies. Devices with access to the Web are desktop PCs, notebooks, 3G mobile phones, handhelds (Nintendo DSi), modern TV-sets, gaming consoles (for example PlayStation 3), accessible sensors and many more. Currently there are no broadly accepted standards, which specify access and maintenance of device-specific resources over web-based technologies in Web applications. These resources cannot be controlled and monitored without using proprietary interfaces and dedicated software. As a consequence, device-specific resources are hard to integrate into composite applications, which severely hinder the realization of a flexible ecosystem of devices that can be reused. There are a few solutions with a limited scope for specific device groups, like Wholesale Applications Community (WAC) [3], a framework for mobile phones with similar objectives and UPnP [4], which is a solution for Home Entertainment to connect seamlessly with the environment.

If a Web application wants to use these and other similar technologies to enable access to resources from

heterogeneous devices, it must also take care of the support, because additional and changing interfaces or protocols force the developer to modify his Web application. Therefore a uniform interface is required, that abstracts the underlying heterogeneous user devices and offers their resources to Web applications. For a satisfactory solution regarding the problem statement described above, various targets have to be implemented successfully.

The shared resources of the user devices have to be accessible through a uniform interface so that Web applications can easily use them. User devices and their resources will be mapped to a RESTful API, which abstracts from the executing resource. In doing so, the communication protocols or API's used by the underlying resource will be completely hidden from the client. This increases the interoperability for a looser coupling between the parts of distributed devices, because the user can replace resources with new ones without an update of the Web application. This facilitates the creation of mashups, which combine traditional Web 2.0 services with resources from the user device. The user controls the access to his resources with a user-configurable rule-based system. This system provides fine-granular adjustments for the user resources. An overview for the described objective:

1. Enable services on heterogeneous and distributed devices for Web applications
2. Increase interoperability for a looser coupling between the parts of distributed devices
3. Make services on devices available as REST resources
4. Abstract the proprietary communication protocols or API's of devices and offer their accessible functionalities via a RESTful API
5. Introduce a user-configurable rule-based system for service access control

The project <Device Cloud> at Fraunhofer Institute for Open Communication Systems (FOKUS) covers these objectives, which will be tackled in the following sections of this paper. The focus aside from enabling access to resources from heterogeneous devices lies in a user-configurable rule-based system to control the access for device resources.

2. RELATED WORK

The introduction already mentioned WAC. This framework enables device-specific capabilities via a JavaScript API. But in contrary to <Device Cloud>, Web applications do not extend on distributed devices.

Web of Things (WoT) [5] is the vision bringing embedded devices into the Web by using Web standards. The project takes the same direction like <Device Cloud>. Therefore it is no surprise that they cover the objectives one till four. It is a goal of WoT to enable services on heterogeneous and distributed devices for Web applications (objective one). Furthermore it increases interoperability for a looser coupling between the parts of distributed devices (objective two).

The WoT architecture abstracts the communication with several types of sensors and offers their accessible functionalities via REST (objective three and four). However the main stress of the project lies on smart things like sensors and this project has a wider focus. User devices like mobile phones, multimedia-components and notebooks are the subject. The devices build the center point in WoT and in <Device Cloud>, it is the user. Users in WoT can just connect their devices through the WoT-server and make them web-enabled. But they have no capability to restrict the access to these devices. If the devices are enabled to the Internet, every user and application can use these devices. However this assumption is acceptable for smart devices like sensors. In contrast to WoT, the user has in <Device Cloud> the decision who gets access to these devices. Ergo WoT does not cover objective five.

Another project with the same direction is webinos. *"Webinos will define and deliver an Open Source Platform which will extend existing Web technologies to enable Web applications and services to be used and shared consistently and securely over a broad spectrum of converged and connected devices, including mobile, PC, home media (TV) and in-car units."* [6]

Webinos is a Seventh Framework Program (FP7) project funded by the EU. The webinos project has over twenty partners from across Europe spanning academic institutions, industry research firms, software firms, handset manufactures and automotive manufactures. Webinos is coordinated by Fraunhofer FOKUS. In context of previous research studies, webinos enhances the research findings of this paper.

The presented approaches cover one or more objectives from this paper and show possible solutions to gain these. All approaches are still working and there are applications, which build on top of these infrastructures. The presence of such a plurality from similar approaches shows the actuality of this topic.

3. STATE OF THE ART

A rule-based system for access control is one of the objectives in this paper. Some approaches which deal with this subject are described in this paragraph.

3.1. RBAC

RBAC stands for Role Based Access Control and is an approach to restricting system access to authorized users. RBAC is widely accepted as a best practice for managing user privileges within a single system or application. It was developed in 1992 by D.R. Kuhn and D.F. Ferraiolo [7]. Four years later Sandhu et al. present a family of reference models to simplify the work with RBAC [8]. In 2000 the three published a paper, in which they described a unified model for RBAC [9]. This model was adopted as an ANSI/INCITS standard in 2004. Until today NIST supports RBAC research and standards [10].

The revised RBAC model from 2000 is organized in four levels, which are cumulative in the case that each includes the requirements of the previous. In the first level (Flat

RBAC), a user U has one or more roles R whereat every role owns multiple permissions P. Permissions allow an access to one or more objects from the system. One role can be assigned to more than one user. *“The requirement that users acquire permissions through roles is the essence of RBAC.”* [9] The further levels expand this model with role hierarchy and separation of duties (SoD). The Flat RBAC is still the same model like the first approach from 1992. This paper adopts the Flat RBAC model. The entities user, role and permission can be found in the same constellation.

3.2. X-GTRBAC

Rafae Bhatti et al. developed the X-GTRBAC Model [11], which is based on the GTRBAC model [12]. GTRBAC stands for Generalized Temporal Role Based Access Control. GTRBAC is based on TRBAC [13], which is an extension of the famous RBAC approaches. It allows expressing role hierarchies and SoD constraints for specifying fine-grained temporal semantics. SoD and role hierarchies are not new for RBAC, but the fine-grained temporal semantics with them.

In [14] they extend this model and create a Context-Aware Access Control Model for Web-Services. This XML-based approach allows defining a wide range of context-aware rules and observing these. It also provides a context-aware access control system. This system checks the user-defined rules for every request and decides, whether it can pass.

The extensions from GTRBAC to RBAC are very useful for our system, because of the lack of a rule engine in the RBAC model. With approach [14] there are also context-aware rules with observer-patterns for these. All the theoretical parts of these approaches fit in well with this paper. Specially, the later explained rules with fast-changing values (location- and time-based) benefit from these concepts. But the implementation has never reached a status of a well-formed framework. For example, there are no methods for generating and saving all required entities (user, roles, and complex permissions).

3.3. Java Rule Engine

The Java Specification Request 94 (JSR-94) [15] defines the Java Rules Engine API. The Java Community Process (JCP) for this JSR started in 2000 and ended with the final release on 15 September 2003.

The specification defines a lightweight-programming interface to access a business rule engine from a Java Platform. It constitutes just a standard API for acquiring and using a rule engine, but not the rule engine itself. The interfaces contain mechanisms for invoking rule execution sets by runtime clients and for loading rule execution sets from external resources. The execution sets are usable for runtime clients of a compliant implementation.

“One of the most common classes of rule engines is the forward-chaining rule engine. Forward-chaining rule engines implement an execution cycle that allows the action of one rule to cause the condition of other rules to become met. In this way, a cascade of rules may become activated and each rule action executed. Forward-chaining rule

engines are suitable for problems that require drawing higher-level conclusions from simple input facts.” [15]

With business rules you have the ability to separate program code and the rule description. But the rules are not independent from the program code, because they refer to the methods in the *then* portion. The JSR-94 has no specifications for role-based access controls as in RBAC. This does not belong in a business rule specification and has to be modeled separately with the rules or outside of the rule chain. A more sophisticated approach is to use a RBAC implementation and use a set of business rules for the compliance of SoD and the permissions assigned to a role.

3.4. Further Approaches

There are many more approaches, which deal with the topic access control. Many access control approaches are based on the RBAC model. Because there is no general specification for rule evaluation inside the RBAC model, various approaches present solutions for this topic. There are also a lot of approaches, which deal with rule engines. Business rule engines like the Java Rule Engine are only one possibility for making decisions with rules. The following section presents a little glimpse over the rim of the proverbial tea cup.

A universal policy model and an access protection framework to secure data space are described in [16]. It provides policy indexing, parallel searching for policies and resources and building of policy hierarchies. This makes the framework efficient for a large number of users. Policies are saved as 6-tuple, whereby the access right can be dependent on many different factors. The approach does not provide any roles or user-models, because it is designed for data sources with many rules. But it is not the focus of this paper because resources aren't shared with a lot of users.

The two mashup frameworks [17] and [18] use rules for auto-generating mashups. But the rules are not created by the user. The two frameworks decide on a base of some inputs, which service or resource may be used for a certain task. The user can indirectly influence the system's choice. [17] uses the five of the user weighted QoS parameters as basis for the choice. [18] combines several models and semantics to evaluate, which mashups a user wants to see in this point of time.

4. CONCEPT

To achieve the established goals from the introduction we need a server infrastructure. Figure 1 shows the <Device Cloud> server and the interaction with the three groups: Web Application, devices and device owner.

The server has to fulfill these four tasks:

1. Interfaces for Web applications
2. User Interfaces for setup
3. User-Configurable Rule-Based System for Service Access Control
4. Communication between user devices and Web applications

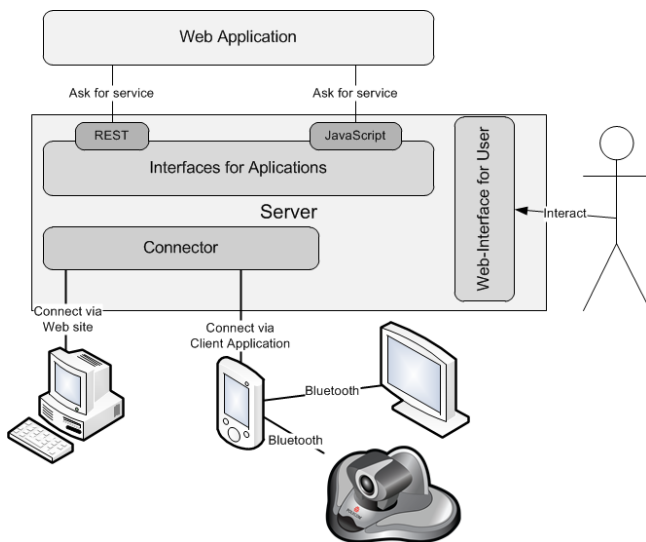


Figure 1. Interaction with the <Device Cloud> server

4.1. Basics

4.1.1. Setup

The registration process of third-party Web applications is based on common authentication concepts. A provider needs to register these applications at <Device Cloud>. After the registration process the provider gets a token for authentication.

Before a Web application use resources from a device, the user has to register this device at the <Device Cloud> server via a GUI interface.

In a second step the GUI shows the related resources to this device. The user marks the resources, which he wants to offer.

The services from the user devices will be mapped to a uniform interface. Objective three says: make services on devices available as REST resources. And this is what we do. All user services are accessible via a RESTful API. The API abstracts the device capabilities and offers these as REST resources. A Web application, which uses a resource, does not know the real executing service and the operating device.

4.1.2. Connect a user device to the server

Register the user devices is only one side of the coin, connect it to the server the other. There are different ways to connect a device with the server. The more common way is the connection over a Web site. The user logs in with his device via this site. The <Device Cloud> server can now use the resources from this device (given that the user registered the device before). The other way is via an installed application. This application connects to the <Device Cloud> server and offers the resources. Both ways have advantages and it depends on the device class, which is the better one. The log-in via the Web site has the

advantage, that the user device needs only a Web browser. No further software is required.

But an installed application to communicate with the <Device Cloud> server is for some device classes the better choice. Devices like sensors have per default no Web browser. The Connector has to communicate with them via proprietary interfaces and protocols. If there is a possibility to install further software on the sensor, it would be a server (cf. Web of Things).

4.2. Property Framework

The system covers the objectives from the introduction, which is described until this stage. Furthermore, four of the five above mentioned objectives are already covered at this point. Users add heterogeneous and distributed devices to their personal device cloud. And Web applications use the services from these devices (objective one). Objective two is due to this fact. This increases the interoperability for a looser coupling between the parts of distributed devices. Because all devices and their services are abstracted behind a RESTful API, objective three and four are also fulfilled.

So far, this approach differs from the previous discussed approaches WoT and Pachube [19] (a platform to share sensing resources) in one fact. WoT and Pachube concentrate on sensors and this approach focus user devices. This includes sensors, but contains much more device classes. Objective five is still uncovered and the continuation of the user story will show the necessary of this last missing objective.

As mentioned above, the user Kevin registers his desktop PC and his mobile phone at <Device Cloud>. Both devices have a camera, which is also known by <Device Cloud>. The PC camera is the better one. Mango + has a video chat functionality and Kevin wants to use this. If Kevin at home, he wants to use his PC camera for the video chat and at the same time the mobile phone for the event recommendations from Mango +.

The <Device Cloud> server above offers no possibility to manage this for Kevin. What he needs is a set of rules to control the access to his resources. In this special case a *priority rule* to prefer a camera is required. The Property Framework fulfills this task.

4.2.1. Rule creation GUI

Error! Reference source not found. shows a mockup of the GUI. In the left upper corner, we see a drop-down list with "Camera" selected. With the help of this list the user chooses one of the provided abstract services. An abstract service could be SMS, Location or Camera. The real user resources are related to this. We use the term abstract services in the further context in the described manner.

Kevin has two devices with a camera. A Web application uses one of Kevin's cameras via a request for the abstract service Camera. But the Web application does not know which one of the two cameras is chosen to fulfill the task. The shown rules are only related to the abstract service Camera. If the user chooses another item from this list, the view will display rules related to it.

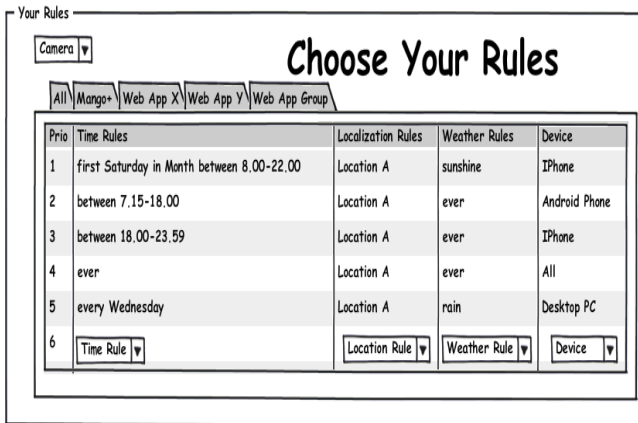


Figure 2. Rule Creation

The tabs show the roles, which the user has been awarded for the Web applications. The roles assign to one or more Web applications. Per default, every role contains only one Web application. But the user has the ability to combine more Web applications to one role. The defined rules are valid for all Web applications which applied to a role. Similar to the drop-down list the rules are only related to the chosen role. If the user chooses another role, the view will display rules related to it. The big list in the middle represents the rules. The complete list in **Error! Reference source not found.** contains five rule sets (rows). Every rule set has a number of rules and devices which are affected to these rules (column). The first column Prio represents the priority of the rules.

The cells present a textual representation of user generated rules. If the user clicks in a cell, a menu pops out. This menu offers all supported options from a rule type to regulate this.

4.2.2. Types of rules

We already know one required rule type - the priority rule. With the help of this rule the user orders his personal resources which are mapped to the same abstract service. Kevin has two devices with a camera and wants to prefer the iPhone camera. The Property Framework decides with the constructed list of precedence, which device fulfills the task. If the device with the resource on place one passes all rules in this row, the Property Framework takes this. Otherwise it looks for the next one.

Kevin adds the mobile phone from his girlfriend to his profile. Because the mobile phone contracts are different, he wants that Mango+ and Web-App X send SMS after 6pm via the mobile phone from his girlfriend. Before it, the applications should use his Android phone. What Kevin needs is a *time rule*. The user has several time criteria to control the access to his resources. He can adjust a time interval in which a resource is usable from a Web application. He can also choose one or more days of the week or a date interval to restrict the access to his resources.

Kevin uses Web-App X everywhere, but he wants to hear the nice music tracks from this application only at home.

He does not want to annoy anybody with the music. For this case Kevin needs a *localization rule*. He declares an area on a map as the validity area for a resource. The simplest case for doing that is a point with a radius. A more advanced method is a polygon with an unlimited number of edges. These three rule types are fundamental for creating rules to control the access to user resources. Of course there is enough space for more rule types. The Property Framework is built to handle with any number of rule types. A developer should easily add further rules to the already exist rule types. Now we have a set of three basic rules:

- Priority Rule
- Time Rule
- Localization Rule

4.2.3. Rule Classes

The rules can be divided in two classes. All rules need an information base to decide, if a request is allowed or denied. The location of this information base is the divisor of these groups. One location is clear. All rule types need the previous taken decision. The other part of the information base is different at the rule types. There are two classes:

- Rules, which use Web Services or server resources
- Rules, which based on user resources

The *priority rule* and the *time rule* use server resources. The *priority rules* need only data from the data storage to evaluate a request. And the *time rule* read the system time from the server. The second group, where the *localization rule* type belongs to, raises questions. The *localization rule* needs a user resource, which returns the current position from the user. This causes some problems. What is, if no actual online device supports the needful resource? The most obvious option is to deny the request for this rule. How we respect the user created rules in this case? Should we evaluate the required rule-sets, if the server needs the resource for the rule evaluation? We leave these design questions open.

4.2.4. Realization with RBAC

For the realization of the Property Framework we use the concepts from RBAC and X-GTRBAC [11]. We applied the typically level one RBAC design (compare [9]) in our schema for the stored rules. Our Web applications are the RBAC users and they assigned to certain roles. Every Web application is assigned to at least one role for every <Device Cloud> user. In contrast to the most other systems which implement the RBAC model we have many more roles than RBAC users. Every role is assigned to a set of permissions, we call it rule sets. A rule set contains rules and is assigned to one or more user devices. A role is only viewable for the <Device Cloud> user, who has created this one. This user creates rule sets and assigns them to the role. The user has also the ability to group RBAC users (our Web applications) and form a role which is assigned to more than one user.

4.2.5. Rule evaluation

The RBAC model lacks of permission handling. The model deals with it like a black box. The approaches around GTRBAC fill this gap by describing concepts to process context aware rules. All our rules stand in a context and therefore the found principles are relevant for this work.

If a Web application asks for an abstract service from a user, the Evaluation Engine from the Property Framework has to check the related rules. The evaluation of the rules decides if the Web application is authorized to use the abstract service. The following paragraphs help us to explain how the Evaluation Engine interprets and processes the user generated rules. The Web application Mango+ wants to use the abstract service Camera in our scenario.

Which steps have the Evaluation Engine to make a decision about a request?

At first the Evaluation Engine needs the necessary data. It gets the stored user generated rules. It also gets context-based data from other sources (for example from a Web service, a user device or from the system itself). The Evaluation Engine creates an internal representation of the rule creation GUI form **Error! Reference source not found.** with this data.

The next step is to evaluate the rules. The Evaluation Engine has two result types. It returns a Boolean value for the Web application (accepted or denied access request) and a device identifier for the Device Manager. The purpose of the Device Manager is the registration of user devices and the management of their connections. The user pretends the order of rule evaluation with the *priority rule*. The rule set with highest priority (lowest number) is the first to be evaluated. All rules in the first row will be evaluated. If the result of all is true, the Evaluation Engine returns the ID of the device from the last column. If the result is false, the Evaluation Engine goes further with the next row and so on. If the result of all rows is false, the engine returns no device ID and the Web application gets no access rights to any user device. The engine evaluates, if the related devices are online. A row, where all devices are offline, returns always the value false. These requirements relates to the following equation, which is based on the Boolean algebra:

$$\mathcal{AR} = \begin{matrix} [r_1(x_{11}) \wedge r_2(x_{12}) \wedge \dots \wedge r_n(x_{1n})] \\ \vee [r_1(x_{21}) \wedge r_2(x_{22}) \wedge \dots \wedge r_n(x_{2n})] \\ \vdots \\ \vee [r_1(x_{m1}) \wedge r_2(x_{m2}) \wedge \dots \wedge r_n(x_{mn})] \end{matrix}$$

4.3. Communication

This paragraph describes the basics behind a resource access request from a Web application. We distinguish between three phases:

- access request phase
- connection establishment
- communication phase

4.3.1. Access Request Phase

The Abstract Interfaces for Web applications follows the REST design principles. Web applications ask for access to user resources via these interfaces. The Abstract Interfaces for Web applications have the following design:

URI: {<Device Cloud> URI}/{resource}/{user}

The URI consists of three patterns. The first is {<Device Cloud> URI}. This is just the base URI from the <Device Cloud> server. The second pattern is {resource}. This is one of our provided abstract services (e.g. Camera). {user} stands on the end of the URI. It is the user ID, which is used from the Web application. This value is mapped to the <Device Cloud> user ID. The user has to adjust this ID during the selection of his preferred Web applications. If a Web application sends an access request to the above described URI, the <Device Cloud> server checks this request. Abounded of the previously created rules from the user the server permits or denies the request.

4.3.2. Connection Establishment and Communication Phase

The previous paragraph described how a Web application gets access rights for a certain user resource. But there was still no communication between Web application and a user device. This paragraph illustrates several communication methods. We differentiate between three communication styles:

- 1) The Web application sends a request via an interface to the <Device Cloud> server. The server forwards the request to the appropriate user device. The device processes the request and if necessary it sends a response back to the server. The server forwards it to the Web application.
- 2) After the Web application got the access rights, the server initiates a connection between Web application and client. The two entities speak direct without a server in the middle.
- 3) Like at the second style, the server establish a connection between client and Web application. But instead of a duplex channel, there is only a simplex. The Web application sends requests directly to the client, but the response takes the way through the server. The other direction is also possible. The Web application sends the requests via the server to the client and gets the responses directly from it.

All three styles have advantages and disadvantages. The advantage from the first communication style is simplicity for the Web application. The Connector abstract all devices and the Web application needs only the REST interface to communicate with user devices. The prize for this is the high server load. All data go through the server and it has to manage several connection states.

The second style has also advantages. The server load is lower than in the first communication style, because the data does not pass the server. The <Device Cloud> server has to establish only the connection between Web application and user device. But the Web application has to

deal with different kinds of communication for the heterogeneous devices. There is no part, which abstract this. The third communication style is a mix of one and two. It unifies the advantages and disadvantages in one style. Depending in which direction is the data flow, there are more pros than cons or not. Based on the classification of this three communication styles we have several communication methods.

REST request and response via server: In this communication method Web applications use only REST. They send the REST request to the server and get a REST response. The underlying communication is fully abstracted for the Web application.

REST with redirect: A Web application gets with access granted response from the server a new REST-based URI. This is the URI from the REST-enabled Web server, which runs on the user device.

Web Sockets: The client of a user device creates a Web Socket with a specified URI. The <Device Cloud> server knows the URI and sends the Web application this. The Web application calls the URI and establishes a full-duplex communication with the client for any kind of data.

Both REST methods are based on widespread and well proven technologies. But both have a deficit. *REST via server* is an example for communication style one, where the complete traffic goes through the server. This increases enormous the server load. *REST with redirect* solves this problem with a redirection from the traffic directly to the client. But this requires a running Web server on the user device. This is an assumption, which does not fit for all device classes.

Web Sockets and *REST with redirect* are booth communication style two methods. Both do not route the traffic through the server. But Web Sockets are a technique, which is still in its infancy. In the near future we will choose Web Sockets for our purpose, but today there are the REST-based communication methods. The <Device Cloud> server will provide both methods. The server prefers the *REST with redirect* communication method and chooses *REST via server* as fallback.

5. REALIZATION

The core of this system is the server, which is implemented with Java Enterprise Edition 6 (Java EE 6) and runs on a GlassFish v3 server. A MySQL 5.1 database is used for the data storage. The user interfaces are built with JavaServer Faces 2.0 and the component library IceFaces in version 2.0.2. We use Jersey for the implementation of REST. Jersey is the reference implementation of the Java Specification Request 311, also known as JAX-RS, which stands for Java API for RESTful Web Services [20].

6. CONCLUSION

Now, we can say, that the work presented here reached all of its goals. We have a base, where users can connect their devices and offer their resources. Web applications have the ability to use these resources via a REST interface. On top

users have the ability to restrict the access to their resources via a configurable rule-based system.

The projects WoT, Pachube [19] and SenseWeb [21] have nearly the same goals. The three approaches virtualize sensors and offer them over as API. But this paper does not concentrate on sensors. There is a wider range with user devices. Pachube has also an active community, which creates applications with this API. This shows us, how actual the topic *virtualizing of web-enabled devices* is.

This paper goes another step further and developed an access control system for the user resources. This is the real improvement over the other approaches. We have seen in our concept, that a completely abstraction of the user device is not possible. The problems with the two REST-based communication methods are already known. For the connection via Web Sockets needs the Web application information about the used devices. A complete abstraction is perhaps only possible with high effort. There have to distribute new protocols and extension on server- client-side. But there is no way for doing that.

6.1. Future Work

It is commonly known, that the work on a software system is never complete. In this section, we present some possible extensions for the <Device Cloud>.

6.1.1. Indirect connected devices

We spoke always from direct connected user devices. But the Private Area Network (PAN) contains also devices, which have not the ability to connect with the server. The resources of these devices can be also enabled and virtualized.

To share these resources with the server, the user needs an additional application on a device, which is already connected. The application searches for other devices in the surrounding area and shows the user a list of results.

A possible scenario is a mobile device with such an application. The application searches for Bluetooth-enabled devices in the near area and shows the user a list of these devices. The user marks a listed device as trustable and adds it to his personal device cloud. These devices will be always automatically available for Web applications, when they are located in the PAN of the activated client. A new search and adding are not necessary, because the application knows already the previous found devices. The communication with Bluetooth devices is specified through Bluetooth profiles. The application has to handle several profiles in order to interact with the devices.

6.1.2. Community Features

The <Device Cloud> addresses Web applications as users of the enabled resources. We can expand the group of requester with members of a community. These members could be in a buddy list of the resource owner. It is possible to share the resources also with other users of the community. The rule system should be extended, so that the

resource owner has the ability to create rules for restricting the access also for community members. In the actual system the resource owner can only restrict the access for a complete community, which is represented through one or more Web applications. WoT made this step already [22].

6.1.3. Quality of Service

Another type of rules could be the Quality of Service parameters. But these parameters should not be set from the users, because it requires advanced skills to handle with these. A normal user has not such skills.

The approach [17] shows, how a mashup creation engine can use these parameters to choose a well suited mashup for the user. The ideas from this approach can be adopt from this paper to offer a QoS rule. The user can only choose, if he wants to activate this rule or not. The QoS rule adjusts itself automatically.

REFERENCES

- [1] Andreas M. Kaplan and Michael Haenlein, "Users of the world, unite! The challenges and opportunities of Social Media," *Business Horizons*, no. 53, pp. 59-68, 2010.
- [2] Roy Thomas Fielding, "Architectural Styles and the Design of Network-based Software Architectures," University of California, Irvine, doctoral dissertation 2000.
- [3] WAC. (2010, July) Wholesale Applications Community. [Online]. <http://www.wholesaleappcommunity.com>
- [4] UPnP-Forum. (2010, November) UPnP-Forum. [Online]. <http://upnp.org>
- [5] Vlad Trifa, "Content Creation on the Web: Mashing Up the Real World With the Internet," in *Proceedings of The First International Workshop on Contents Creation Activity Support by Networked Sensing (CCASNS)*, Kanazawa, Japan, 2008.
- [6] webinos Consortium, "webinos report: Use Cases and Scenarios," report 2011.
- [7] D. Richard Kuhn and David F. Ferraiolo, "Role-Based Access Controls," in *15th National Computer Security Conference*, Baltimore, Maryland USA, 1992, pp. 554 - 563.
- [8] Ravi S Sandhu, Edward J. Coynek, and Hal L. Feinstein, "Role-Based Access Control Models," *IEEE Computer*, vol. 29, no. 2, pp. 38-47, February 1996.
- [9] Ravi Sandhu, David Ferraiolo, and Richard Kuhn, "The NIST model for role-based access control: towards a unified standard," in *RBAC '00: Proceedings of the fifth ACM workshop on Role-based access control*, Berlin, Germany, 2000, pp. 47-63.
- [10] National Institute of Standards and Technology. (2010, November) Role Based Access Control And Role Based Security. [Online]. <http://csrc.nist.gov/groups/SNS/rbac/>
- [11] R. Bhatti, "X-GTRBAC: An XML-based Policy Specification Framework and Architecture for Enterprise-Wide Access Control," Purdue University, West Lafayette, Indiana, USA, Master Thesis 2003.
- [12] J. B. D. Joshi, Elisa Bertino, Usman Latif, and Arif Ghafoor, "Generalized Temporal Role Based Access Control Model (GTRBAC) (Part I) - Specification and Modeling," Purdue University, West Lafayette, Indiana, USA, Technical Report 2001.
- [13] Elisa Bertino, Piero Andrea Bonatti, and Elena Ferrari, "TRBAC: A temporal role-based access control model," *ACM Transactions on Information and System Security*, vol. 4, no. 3, pp. 191--233, 2001.
- [14] R. Bhatti, E. Bertino, and A. Ghafoor, "A trust-based context-aware access control model for Web-services," in *International Conference on Web Services (ICWS'04)*, San Diego, California, USA, 2004, pp. 184-191.
- [15] Sun Microsystems, Inc. (2003, September) JSR 94: Java Rule Engine API. Specification.
- [16] Jin Lei, Zhang Yawei, and Ye Xiaojun, "Secure Dataspace with Access Policies," in *International Symposium on Parallel and Distributed Processing with Applications (ISPA)*, Sydney, Australia, 2008, pp. 701-706.
- [17] Xu Huiyang, Song Meina, and Luo Xiaoxiang, "A QoS-oriented management framework for reconfigurable mobile mashup services," in *11th International Conference on Advanced Communication Technology (ICACT)*, Myeong-Ri, Bongpyong-Myeon, Pyeongchang-Gun, Gangwon-Do, Korea , 2009, pp. 2001-2005.
- [18] Fedor Bakalov, Birgitta König-Ries, Andreas Nauerz, and Martin Welsch, "Ontology-Based Multidimensional Personalization Modeling for the Automatic Generation of Mashups in Next-Generation Portals," in *First International Workshop on Ontologies in Interactive Systems (ONTORACT)*, Liverpool, United Kingdom, 2008, pp. 75-82.
- [19] Connected Environments Ltd. (2010, December) Pachube. [Online]. <http://www.pachube.com/>
- [20] Sun Microsystems, Inc. (2008, September) JSR 311 - JAX-RS: The Java™ API for RESTful Web Services. Specification.
- [21] Microsoft Corporation. (2010, December) Microsoft Research: SenseWeb. [Online]. <http://research.microsoft.com/en-us/projects/senseweb/>
- [22] Dominique Guinard, Mathias Fischer, and Vlad Trifa, "Sharing Using Social Networks in a Composable Web of Things," in *Pervasive Computing and Communications Workshops (PERCOM Workshops), 8th IEEE International Conference*, Mannheim, Germany, 2010, pp. 702-707.

ABSTRACTS

Session 1: ICTs helping Africa¹	
S1.1	<p>The Role Of ICTs In Quantifying The Severity and Duration Of Climatic Variations - Kenya's Case*</p> <p><i>Muthoni Masinde, Antoine Bagula (University of Cape Town, South Africa)</i></p> <p>For the last 2 decades, Kenya has consistently contributed the highest number of people affected by natural disasters in Africa. This is especially so for disasters triggered by climatic variations. The Kenya Meteorological Department has provided regular weather forecasts since the 60s. One of the shortcomings of this Department's approach is the fact that their forecasts provide conceptual indications of droughts/floods without giving operational indicators. This makes it difficult for key stakeholders to develop solid strategic plans. Innovative use of ICTs can turn around this situation by realigning the forecasts to aid in answering questions such like, how long and how severe the predicted climatic variations will be. Use of cheaper wireless sensors can also help readdress the current poor coverage by weather stations. Based on analysis of 31 years of historical daily precipitation data from three weather stations, we prove that the Effective Drought Index can be used to quantify droughts/floods. We also present an effective web-based system that policy makers can use to monitor droughts/floods on daily basis. In the discussion, we explain how an on-going initiative aimed at integrating wireless sensor networks and mobile phones will further improve drought monitoring.</p>
S1.2	<p>ICT use in South African Microenterprises: An assessment of Livelihood outcomes</p> <p><i>Frank Makoza, Wallace Chigona (University of Cape Town, South Africa)</i></p> <p>This paper reports on a study on the impact of using Information and Communication Technologies (ICT) on the livelihoods of microenterprises. The study used a qualitative approach and focused on the case of South Africa. Microenterprises play an important role in socio-economic development and in bridging the gap in the segments of the economy of South Africa. The study findings confirm that ICT use and support of institutions and organisations have a positive impact on the livelihoods of microenterprises. However, ICT use in microenterprises is curtailed by challenges beyond access and ownership of ICTs. Chief among these problems is lack of awareness of application of ICT in business activities and awareness of support services provided by business development organisations.</p>
S1.3	<p>SM2: Solar Monitoring System in Malawi</p> <p><i>Mayamiko Nkoloma (Malawi Polytechnic, Malawi); Marco Zennaro (ICTP - The Abdus Salam International Centre for Theoretical Physics, Italy); Antoine Bagula (University of Cape Town, South Africa)</i></p> <p>This paper describes recent work on the development of a wireless based remote monitoring system for renewable energy plants in Malawi. The main goal was to develop a cost effective data acquisition system that continuously presents remote energy yields and performance measures. A test bed comprising of a solar photovoltaic (PV) power plant has been set up at Malawi Primary School and a central management system at Malawi Polytechnic. The project output gives direct access to generated electric power at the rural site through the use of wireless sensor boards and text message (SMS) transmission over cellular network. The SMS recipient at the central site houses an intelligent management system based on FrontlineSMS for hosting SMSs and publishing remote measurement trends over the Internet. Preliminary experimental results reveal that the performance of renewable energy systems in remote rural sites can be evaluated efficiently at low cost.</p>

¹ Papers marked with an “*” were nominated for the three best paper awards.

Session 2: Connecting rural regions	
S2.1	<p>Proposal of a Wired Rural Area Network with Optical Submarine Cables <i>Yoshitoshi Murata (Iwate Prefectural University, Japan); Hiroshi Mano, Hitoshi Morioka (Root Inc., Japan)</i></p> <p>The lack of access to fast Internet services is a serious digital divide for future networks. Most areas outside fast Internet service coverage are rural areas. Many kinds of wireless systems have been proposed for rural areas because of their low establishment cost per residence. Where there are several residences in a small area, a wireless system is effective for establishing a network at a low cost. Our investigation of residence plots in rural areas around Morioka city, Japan, revealed 15 residences on average at intervals of 50-200 m along roads. For such areas, there are no suitable wireless systems. In this paper, we propose a wired rural area network system that uses an optical submarine cable instead of a wireless system. We name it OSC-RAN. One of the goals for the OSC-RAN is to reduce the total cost, which includes both establishment and maintenance costs. We conducted the OSC-RAN field trial and provided trial services for about six months. We verified that residents and/or home appliance installation workers could establish the network, and they longed for the Internet.</p>
S2.2	<p>Development of an ICT road map for eServices in rural areas <i>Nobert Rangarirai Jere, Mamello Thinyane (Telkom Centre of Excellence in ICTD, South Africa); Alfredo Terzoli (Rhodes University, South Africa)</i></p> <p>ICTs, driven by the convergence of computers, telecommunications and traditional media, are crucial for the knowledge-based economy of the future. The rapid technological changes have resulted in different ideas being suggested for the expected ICT applications. As a result, different e-Service applications have being developed as a way to foster ICT developments. However, ICT applications deployed at the moment may not be able to sustain the rural communities in maybe 10years or more to come. The paper considers the past, analyzes the present and conduct surveys to gain insight into the future. Based on all of this information, the research tries to provide an ICT road map for what is to come. What kind of applications can we develop now to cater for the technological changes, so that the ICT applications developed today would still be compatible with those developed in years to come? The Siyakhula Living Lab (SLL) is used as the case study in this paper and some interviews and literature review are done to get different ideas on the future of ICTs.</p>
S2.3	<p>Investigating implementation of communication networks for advanced metering infrastructure in South Africa <i>Monontši Nthontho, S P Chowdhury, Simon Winberg (University of Cape Town, South Africa)</i></p> <p>Advanced metering infrastructure (AMI) is a relatively new field in South Africa. The first standard to govern the envisaged AMI implementation was released in 2008 in NRS 049 document. This paper reports on the investigation of supporting communication networks for AMI. There are several approaches that the South African utilities (Eskom and Municipalities) can follow to implement AMI communication networks. Two broad options are constructing their own private network or connecting with existing network service providers. The communication networks can either use wired or wireless media technologies. They can use mobile broadband networks for a wireless wide area network. Moreover, they can build on and use their legacy optic fibre and PLC networks used for supervisory control and data acquisition (SCADA) application. This paper investigates both wired and wireless technologies that can be considered. Furthermore, it discusses different factors such as bandwidth capacity that would influence the approach chosen.</p>

Session 3: Reflections on a fully networked society	
S3.1	<p>Invited paper: Cooperative Wi-Fi-Sharing: Encouraging Fair Play <i>Hanno Wirtz, René Hummen, Nicolai Viol, Tobias Heer, Mónica Alejandra Lora Girón, Klaus Wehrle (RWTH Aachen University, Germany)</i></p> <p>Cooperation enables single devices or applications to establish systems that exceed the capabilities of single entities. A prime example for cooperation are Wi-Fi-sharing networks, in which multiple parties cooperatively share their resources, such as wireless access points and Internet uplinks, to form a large-scale Wi-Fi network that offers access to mobile users. Mobile users benefit from this network by gaining free network access at every access point of the network. However, such cooperation needs to be established in the first place by providing incentives to users to join the network. Furthermore, in an established network, users need incentives to behave cooperatively when using the network. Frameworks to provide incentives and to regulate user behavior in the presence of malicious parties can exist at multiple levels: The technical level inside the given network, a contractual level that regulates the operation of the network and the legislative level that establishes general rules for the operation of Wi-Fi-sharing networks. In this paper, we analyze requirements and mechanisms to establish such frameworks at each level and discuss possible solutions and existing examples.</p>
S3.2	<p>Making things socialize in the Internet - Does it help our lives?*</p> <p><i>Luigi Atzori (University of Cagliari, Italy); Antonio Iera (University "Mediterranea" of Reggio Calabria, Italy); Giacomo Morabito (University of Catania, Italy)</i></p> <p>Current communication and computation technologies make it possible to embed intelligence and communication capabilities in most of the things surrounding us; this leading to the Internet of Things (IoT) concept. To really exploit the potential of the IoT, objects and provided services should be easily discoverable and usable by humans and by other objects. Besides, trustworthiness of the billions of members of the IoT should be a key element in service selection. Existing solutions for service discovery in IoT do not scale with the number of nodes that is expected to be order of magnitude larger than in the current Internet. In this paper we propose to build a social network, that we name the Social Internet of Things (SIoT), that can be used to provide a navigable structure to the IoT. We also provide a framework that can be applied to socially tie things together and a preliminary architecture to be used as a baseline for the implementation of the SIoT. Our work demonstrates that standards should support establishment and management of federations of objects (ruled by social relationships) that represent "communities" of things in the SIoT.</p>
S3.3	<p>Net-Centric World: Lifestyle of the 21st Century <i>Daniel Kharitonov (Juniper Networks Inc, USA)</i></p> <p>In this paper, we research the potential of information communication technologies (ICTs) for changing our society from a commute-centric to a network-centric environment. We propose to formalize the key attributes of ICT-based telecommuting experiences from both economic and human interactivity perspective. We introduce the notion of network-eligible transactions and disclose the link between degree of network centrality and worker settlement radius, postulating that media-rich network services have a strong potential to increase the physical distance between work and home locations. We also highlight notable technology challenges and opportunities of migration from location-based to mobile living, signifying the needs for new services and standards development.</p>

S3.4 Reflexive Standardization of Network Technology

Ian Graham (University of Edinburgh, United Kingdom)

This paper investigates a JTC1 working group to identify how formal standards processes are evolving in response to globalization and the emergence of consortium standardization. It is found that being part of the formal standards development systems provides a source of legitimacy, but also limits the freedom for the process to replicate the structures of consortia. Their standardization process is deeply reflexive, with focuses on maintaining legitimacy and negotiating the boundaries where their activities impinge on other processes. It is argued that the structure of committees of multiple national standards bodies feeding national requirements into the global processes by responding to ballots resolutions and nominating representatives is increasingly anachronistic in a world of global communications, more open standards development, global technology companies and the weakening of the ability of states to identify a national interest in technology policy.

Session 4: Frequency and Spectrum Management

S4.1 Radio Resource Management in OFDMA-CRN Considering Primary User Activity and Detection Scenario

Dhananjay Kumar (Anna University, India); Kanagaraj Nachimuthu Nallasamy (Alcatel-Lucent India Limited, India)

In this paper an adaptive radio resource allocation scheme is developed for OFDMA based cognitive radio network (OFDMA-CRN), which not only considers the dynamic nature of primary users but also includes the detection scenario of unused licensed spectrum. In contrast to the existing research for OFDMA-CR systems which consider either of these issues independently, our approach tackles both of these concerns jointly and finds optimal solution. The proposed sub-carrier and power allocation (SPA) algorithm optimally selects and computes optimal power loading for each sub-carrier thereby increases sum data rate and overall throughput of the system.

S4.2 Optimal Pilot Patterns Considering Optimal Power Loading for Cognitive Radios in the Two Dimensional Scenario

Boyan Soubachov, Neco Ventura (University of Cape Town, South Africa)

In Orthogonal Frequency Division Multiple Access (OFDMA) based Cognitive Radio (CR) systems, optimal power loading schemes are devised such that the transmission rate is maximized while maintaining interference from Secondary Users (SUs) to Primary Users (PUs) below a specified threshold. The power loading algorithms however do not distinguish between data and pilot symbols. A similar situation exists for optimal pilot patterns where pilots are placed in the optimal positions to achieve the lowest Mean Squared Error (MSE) but no consideration is given to power loading schemes. This paper investigates this scenario and proposes an optimal solution based on the Least Squares (LS) estimator which could be applied to future CR algorithm implementations as well as a possible standardization aspect to ensure an optimal solution to a crucial problem in practical implementations.

S4.3 Optimal Spectrum Hole Selection & Exploitation in Cognitive Radio Networks

Mahdi Pirmoradian, Christos Politis (Kingston University London, United Kingdom)

Future networks, especially in the framework of smart cities will be populated with various kinds of equipment accessing wireless communication channels. A much higher device variety will increase spectrum scarcity. Cognitive radio networks will be a key enabling technology in order to cope with the availability of the allocated radio spectrum bands. Cognitive Radio (CR) technology significantly utilizes current static spectrum bands assignment in an opportunistic manner. In this paper, we propose two spectrum opportunity (or spectrum hole) selection schemes; Minimum Collision Technique (MCT) and Maximum Residual Lifetime Technique (MRLT). The proposed techniques are evaluated by average channel utilization, average channel collision and successful secondary transmission bytes over licensed channels in a specific period of time (100s). The numerical results confirm that the MRLT scheme provides higher channel utilization and transmission bytes as well as decreases channel collision compared with the MCT scheme.

Session 5: Optimisation of Layers 1 – 3	
S5.1	<p>Transmission Analysis of Digital TV Signals over a Radio-on-FSO Channel* <i>Chedlia Ben Naila, Kazuhiko Wakamori, Mitsuji Matsumoto (Waseda University, Japan); Katsutoshi Tsukamoto (Osaka University, Japan)</i></p> <p>Recently, Radio frequency on free-space optical (RoFSO) technology is regarded as a new universal platform for enabling seamless convergence of fiber and FSO communication networks, thus extending broadband connectivity to underserved areas. In this paper, an experimental demonstration of the newly developed advanced RoFSO system capable of transmitting the Japanese integrated services digital broadcasting-terrestrial (ISDB-T) signals over 1km FSO link. Our innovative system combines a new generation full optical FSO system with radio over fiber (RoF) technology. The obtained results can be used for designing, predicting and evaluating the RoFSO system capable of transmitting multiple wireless services over turbulent FSO link.</p>
S5.2	<p>A Hybrid MAC with Intelligent Sleep Scheduling for Wireless Sensor Networks* <i>Mohammad Arifuzzaman, Mohammad Shah Alam, Mitsuji Matsumoto (Waseda University, Japan)</i></p> <p>In this paper, we present Intelligent Hybrid MAC (IHMAC), a novel low power with minimal packet delay medium access control protocol for wireless sensor networks (WSNs). IH-MAC achieves high energy efficiency under wide range of traffic load. It ensures high channel utilization during high traffic load without compromising energy efficiency. IH-MAC does it by using the strength of CSMA and TDMA approach with intelligence. The novel idea behind the IH-MAC is that, it uses both the broadcast scheduling and link scheduling. Depending on the network loads the IH-MAC protocol dynamically switches from broadcast scheduling to link scheduling and vice-versa in order to achieve better efficiency. Furthermore, IH-MAC uses Request-To-Send (RTS), Clear-To-send (CTS) handshakes with methods for adapting the transmit power to the minimum level necessary to reach the intended neighbor with a given BER target or packet loss probability. Thus IH-MAC reduces energy consumption by suitably varying the transmit power. The analytical results corroborate the theoretical idea, and show the efficiency of our proposed protocol. Considering the importance of a unique MAC protocol for WSNs, we propose a study for standardization work in the ITU as an initiative which can lead to its rapid adaptation.</p>
S5.3	<p>Route Optimization Based On The Detection of Triangle Inequality Violations <i>Papa Ousmane Sangharé, Bamba Gueye, Ibrahima Niang (Université Cheikh Anta Diop de Dakar, Senegal)</i></p> <p>During the last decade, new services networks and distributed applications have emerged. These systems are flexible insofar as they can choose their ways of communication among so much of others. However, this choice of routing is based on a large number of measurements of times (Round Time Trip (RTT)) which are sources of overload in the network. Network Coordinate Systems (NCS) allow to reduce measurements overhead by mitigating direct measurements. However, NCS encounter inaccuracies with respect to distance prediction, when the measured distances violate the principle of the triangular inequality (TIV-Triangle Inequality Violation). Firstly, we propose a new metric, called "RPMO", which is based on the Ratio of Prediction and the Average Oscillations of the estimated distances, to detect the potential TIVs. The obtained results show that the "RPMO" metric gives better performance compared to metrics presented in former work. Secondly, we propose to use the existence of TIVs to optimize the routing in Overlay Network. To achieve this goal, we present a new approach that enables to detect the best shortened paths offered by the existence of potential TIVs.</p>

Session 6: Architectures to support a fully networked society	
S6.1	<p>Invited Paper: Effective Collaborative Monitoring In Smart Cities: Converging Manet And Wsn For Fast Data Collection</p> <p><i>Giuseppe Cardone, Paolo Bellavista, Antonio Corradi, Luca Foschini (University of Bologna, Italy)</i></p> <p>Ubiquitous smart environments, equipped with low-cost and easy-deployable Wireless Sensor Networks (WSNs) and with widespread Mobile Ad-hoc NETWORKS (MANETs), are opening brand new opportunities in urban monitoring. Ur-ban data collection, i.e., the harvesting of monitoring data sensed by a large number of collaborating sensors in a wide-scale city, is still a challenging task due to typical WSN limitations (limited bandwidth and energy, long delivery time, ...). In particular, effective data collection is crucial for classes of services that require a timely delivery of urgent data, such as environmental monitoring, homeland security, and city surveillance. This paper proposes an original solution to integrate and to opportunistically exploit MANET overlays that are impromptu and collaboratively formed over WSNs in order to boost data collection: overlays are used to dynamically differentiate and fasten the delivery of urgent sensed data over low-latency MANET paths. The reported experimental results show the feasibility and effectiveness (e.g., limited coordination overhead) of our solution for MANET overlays over WSNs. In addition, our proposal can easily integrate with the latest emergent WSN data collection standards/specifications, thus allowing immediate deployability over existing smart city environments.</p>
S6.2	<p>SOA Driven Architectures for Service Creation through Enablers in an IMS Testbed*</p> <p><i>Mosiua Tsietsi, Alfredo Terzoli, George Wells (Rhodes University, South Africa)</i></p> <p>Standards development organisations have long been in agreement that the most appropriate and cost effective way of developing services for the IP Multimedia Subsystem (IMS) is through the use – and re-use – of service capabilities, which are the building blocks for developing complex services. IMS specifications provide a theoretical framework for how service capabilities can be aggregated into large service applications. However, there is little evidence that mainstream IMS service development is capability-based, and many services are still designed in a monolithic way, with no re-use of existing functionality. Telecommunication networks are well positioned to stimulate the Internet services market by exposing these service enablers to third parties. In this paper, we marry the two issues by defining an extended IMS service layer (EISL) that provides a service broker that is the central agent in both service interaction management and the execution of external requests from third parties. A prototypical implementation of the service broker is described that was developed using a converged SIP servlet container, and a discussion is also provided that details how third party developers could use HTTP APIs to interact with a service broker in order to gain access to network capabilities.</p>
S6.3	<p>A Virtualized Infrastructure for IVR Applications as Services*</p> <p><i>Fatna Belqasmi, Christian Azar (Concordia University, Canada); Mbarka Soualhia (ETS, University of Quebec, Canada); Nadjia Kara (École de Technologie Supérieure, Canada); Roch Glitho (Concordia University, Canada)</i></p> <p>Interactive Voice Response (IVR) applications (e.g. automated attendant) are ubiquitous nowadays. Cloud computing is an emerging multi-faceted paradigm (Infrastructure as a Service - IaaS, Platform as a Service - PaaS, and Software as a Service - SaaS) with several inherent benefits (e.g. resource efficiency). Very few, if any, IVR applications are offered today in cloud settings despite all the potential benefits. This paper introduces a novel architecture for a virtualized IVR infrastructure and demonstrates its potential with a case study. The architecture proposes IVR substrates that are virtualized, composed, and assembled on the fly to build IVR applications. It relies on a business model which introduces the IVR substrate provider as a new role in the cloud business model. In the case study, which includes a prototype, IVR service providers develop and manage IVR applications using a simplified platform that adds a level of abstraction to the substrates available in the virtualized infrastructure. The applications are offered as SaaS to end-users.</p>

S6.4 Seamless Cloud Abstraction, Model and Interfaces

Masum Z. Hasan, Monique Morrow, Lew Tucker (Cisco Systems, USA); Sree Lakshmi D. Gudreddi; Silvia Figueira (Santa Clara University, USA)

An enterprise, as a Cloud Service Consumer (E-CSC), may acquire and consume (off-premises) resources in one or more Public or Community Clouds owned and operated by one or more Cloud Service Providers (CSP). A CSP (as a CSC: S-CSC) may itself consume resources from other CSPs on behalf of an E-CSC. For seamless manageability an E-CSC may want to combine a select set of on-premises (intranet or private Cloud) resources with off-premises Cloud resources to create a Seamless Cloud (SCL). The E-CSC may also include in the SCL a select set of its (branch and DC) sites. Based on the definition an SCL subsumes various categories of Cloud, such as private, public, community, hybrid and inter-Cloud. A CSP can offer a service, which we call the Seamless Cloud service that will facilitate creation, deletion and update of an SCL on-demand. In a multitenant Cloud environment SCLs of each tenant should be isolated from each other end-to-end (from CSC enterprise to on-demand acquired resources in CSP DC). The SCL service will facilitate such isolation. By adding proper QoS capability to the SCL service, a CSP will be able to offer (what we call) Differentiated Quality of Seamless Cloud Services (DQSCS). In this paper we describe abstraction, model and interfaces (CSC to CSP) for SCL. It is expected that the interfaces will be standardized.

Session 7: Service Quality for a fully networked society

S7.1 Regulation of Bearer / Service Flow Selection between Network Domains for Voice over Packet Switched Wireless Networks

Nikesh Nageshar, Rex Van Olst (University of the Witwatersrand, South Africa)

With the evolution of wireless systems from traditional circuit switch technology to packet based technology there is a requirement that voice be maintained to an acceptable level of quality such that user experience does not become compromised. All next generation wireless networks have been specified as packet switched radio networks which imply that the flaws of traditional packet based networks now also apply to voice over the wireless medium. This combined with the dynamics of a traditional air interface provides a further challenge to voice over a packet switched wireless network. The following paper proposes the facilitation of regulation that will predefine the handover of Quality of Service (QoS) metrics for voice from one predefined QoS network domain to subsequent network domains for wireless system handover. It is the intention of this paper to highlight the advantages of providing a voice QoS regulated admission control, bearer / service flow selection and mobile transport backhaul so as to ensure the successful transmission of quality voice packets.

S7.2 Accessibility support for persons with disabilities by Total Conversation Service Mobility Management in Next Generation Networks*

Leo Lehmann (OFCOM, Switzerland)

This paper describes the principles and concepts necessary to support total conversation service mobility within a fixed/mobile converged telecommunication network. Regarding the network platform, this paper considers the functional architecture of the Next Generation Network (NGN), as the International Telecommunication Union (ITU) standardizes it. The presented procedure shall enable persons, who are disabled by deafness, speech disabilities and/or vision disabilities, to use the advantages of fixed/mobile converged telecommunication networks not only in a stationary situation but also when they are mobile. A strong focus is given on the support of context based service performance adaptation. Different handover scenarios are considered, including devices and network access of different capabilities.

S7.3 LabQoS: A platform for network test environments

Luis Zabala, Armando Ferro, Cristina Perfecto, Eva Ibarrola, Jose Luis Jodra (University of the Basque Country, Spain)

This paper proposes the deployment of a network software platform for experimentation called LabQoS that will allow the scientific community to establish scientific experiments relating to measure the performance of applications and services on the Internet and other network environments. Previous experiences have already deployed measurement systems, but they don't incorporate the concept of control of the scenario to enable the performing of experiments. This LabQoS capacity makes it an innovative platform unique in its approach. A module called Test Environment Builder has been designed in order to monitor the operating parameters, and to propose adaptation strategies to handle the scenario control variables. We define an experimental data model that identifies the entities to be considered in the management of the platform. Assessment mechanisms are proposed to study the dynamic sensitivity of the control variables with respect to the parameters. LabQoS is based on QoS METER architecture that addresses technological aspects such as user management, deployment tools, data collection, reporting, security, etc.

Poster Session: Showcasing innovations for future networks and services

P.1 A Trust Computing Mechanism for Cloud Computing

Mohamed Firdhous, Osman Ghazali, Suhaidi Hassan (Universiti Utara Malaysia, Malaysia)

Cloud computing has been considered as the 5th utility as computing resources including computing power, storage, development platform and applications will be available as services and consumers will pay only for what consumed. This is in contrast to the current practice of outright purchase or leasing of computing resources. When the cloud computing becomes popular, there will be multiple vendor offering different services at different Quality of Services and at different prices. The customers will need a scheme to select the right service provider based on their requirements. A trust management system will match the service providers and the customers based on the requirements and offerings. In this paper, the authors propose a trust formulation and evolution mechanism that can be used to measure the performance of cloud systems. The proposed mechanism formulates trust scores for different service level requirements, hence is suitable for managing multiple service levels against single trust score. Also the proposed mechanism is an adaptive one that takes the dynamics of performance variation along with cloud attributes such as number of virtual servers into computations. Finally the proposed mechanism has been tested under a simulated environment and the results have been presented.

P.2 The Energy Label A Need To Networks And Devices

Virgilio Puglia (Italtel, Italy)

The energy consumption of Information and Communication Technologies (ICTs) is today relevant also compared with the other industries. The evolution of ICT will determine enormous improvements in our daily lives, but will also increase energy need. So energy consumption becomes one of the key aspects in the evolution. Today every vendor manifests a trend to improve energy saving, but the lack of agreed indexes, terms, definitions and procedures makes it difficult for the customer to realize a relevant comparison. This paper presents a holistic approach in order to identify energy key indexes in all the Life-Cycle Assessment (LCA) of devices and networks. This study starts with the analysis of what was already implemented in other sectors and it proposes a devices energy label and a network energetic classification. A case study on the proposed method application is described. The results of the study demonstrates the need to adopt regulatory energetic indexes in order to ensure competition and energy saving.

P.3 A distributed mobility management scheme for future networks

Ved P. Kafle, Yasunaga Kobari, Masugi Inoue (National Institute of Information and Communications Technology, Japan)

Unlike Mobile IP protocols, which specify centralized mobility management schemes, the future network is envisioned to embrace distributed approaches to mobility so that it can avoid a single point of failure and triangular routing problems as well as optimize the handover process. This report presents a distributed mobility management scheme of HIMALIS (Heterogeneity Inclusion and Mobility Adaptation through Locator ID Separation) architecture where mobility signaling takes place through the control network composed of end hosts and dedicated functional nodes in the network. Moreover, the signaling functions are network layer independent; therefore, the proposed scheme can be applied to a future networking environment where multiple protocols do coexist in the network layer. It discusses the architectural components, mobility protocol, and an analysis of performance results obtained from an emulation system.

P.4 Toward Global Cybersecurity Collaboration: Cybersecurity Operation Activity Model

Takeshi Takahashi (National Institute of Information and Communications Technology, Japan); Youki Kadobayashi (Nara Institute of Science and Technology, Japan); Koji Nakao (KDDI Corporation, Japan)

The importance of communication and collaboration beyond organizational borders is increasingly recognized with regard to maintaining cybersecurity. Yet organizations still face difficulties communicating and collaborating with external parties. Among these difficulties is the absence of a common vocabulary, as organizations do not always share the same terminology in describing operations, and this consumes unnecessary time and can lead to miscommunication. This paper addresses the problem by introducing a cybersecurity operation activity model that provides the foundation for defining such vocabulary. The model also facilitates understanding and review of cybersecurity operations and their associated activities. This paper demonstrates the model's usability by visualizing the domains of cybersecurity operations and services and concludes that the model has sufficient usability as a foundation for building vocabulary and as a tool for visualizing cybersecurity issues, which will help expedite communication beyond organizational borders.

P.5 Context Representation Formalism and Its Integration into Context as a Service in Clouds

Boris Moltchanov (Telecom Italia, Italy)

Context Management technology itself is not novel and ICT companies are already trying to find a technically feasible solution and appealing marketing usage of the context-awareness. However, after many years of technology scouting and academic scrutiny within this still innovating area, the usage of the personalized and context-aware services is still limited due to totally new business models that shall be put in place. The context information available in the real world from many potential context sources shall be handled as a near real-time, efficiently processed by many devices and be interoperable among different actors dealing with the context. Therefore among a comprehensive context management framework and its efficient representation the context information shall be exposed in an way easy to use and consume. Even more better is if the context information and data are embedded within service clouds using frontier technology of cloud computing embedding not only the tools for a reach, flexible and scalable service creation but also integrating context knowledge and an efficient real-time data management. A solution integrating the context information in the clouds with its efficient context representation and publish-subscribe web service based interface is described in this paper.

P.6	<p>Supporting technically the Continuity of Medical Care: Status report and perspectives <i>Vasileios B. Spyropoulos, Maria Botsivaly, Aris Tzavaras (Technological Education Institute of Athens, Greece)</i></p> <p>The purpose of this paper is to present the status of the R&D efforts of our Laboratory concerning the development and the improvement of hardware and software means, appropriately designed to ensure Continuity of Medical Care among Primary Health-care Agencies, Hospitals and Home Care, according to existing or emerging National, European and International regulations and standards. Our R&D is presently focused on the development of an integrated prototype system, including first, improved equipment facilitating the Continuity of Care, second, software supporting Medical Decision Making during emergency-care delivery, as well as, real and virtual audio-visual monitoring of chronic and terminally ill patients at home-alike conditions, third, a Continuity of Care Record (CCR), complying with the major ANSI E2369-05 CCR, ISO 13606-1:2008 and prEN 13940 Standards, and finally, linking the CCR to appropriate semantically annotated Web-Services, providing for enhanced technical interoperability and medical clarity and simplicity.</p>
P.7	<p>Coexistence of a TETRA System with a Terrestrial DTV System in White Spaces <i>Heejoong Kim, Hideki Sunahara, Akira Kato (Keio University, Japan)</i></p> <p>In this paper, we have investigated the possibility of coexistence of terrestrial truncated radio (TETRA) as a narrow band system with digital television (DTV) in TV white spaces. Based on the system operation mode of a TETRA system, trunked mode operation (TMO) and direct mode operation (DMO), the interoperable power range of fixed and mobile terminals is obtained according to their frequency offsets and power classes.</p>
P.8	<p>Mobile cloud computing based on service oriented Architecture: embracing network as a service for 3rd party application service providers <i>Michael Andres Feliu Gutierrez, Neco Ventura (University of Cape Town, South Africa)</i></p> <p>The recent emergence of Cloud Computing in the IT world has opened doors for new revenue streams and business models. The movement towards a service-oriented architecture and an all-IP based communication system has led to the IP Multimedia Subsystem (IMS) being accepted as the Next Generation Networks (NGN) service control/provisioning platform. The Telco 2.0 domain can benefit from service enhancement and Web 2.0 technologies such as Cloud Computing. This can be achieved by opening gateways and APIs guarding rich underlying network resources (e.g. location) to 3rd Party ASPs, thus adopting the delivery of Network as a Service (NaaS). Telco services are faced with opportunities to make use of powerful computational power and storage services offered by cloud environments to accelerate business-processing speeds.</p>
P.9	<p>RBAC for a configurable, heterogeneous Device Cloud for Web Applications <i>Hannes Gorges, Robert Kleinfeld (Fraunhofer FOKUS, Germany)</i></p> <p>A key challenge during the development of Web applications on top of multiple heterogeneous devices is to discover and get access to device-specific resources. This paper shows an architecture which enables and virtualizes the resources of user devices whereupon the user controls the access to his resources with a user-configurable rule-based system. User resources are mapped to a RESTful API, so that Web applications can easily use them. In doing so, the communication protocols or APIs used by the underlying resource are completely hidden from the user of the RESTful API. This increases the interoperability for a looser coupling between the parts of distributed devices, because the user can replace resources with new ones without an update of the Web applications, which use these resources. This facilitates the creation of mashups, which combine traditional Web 2.0 services with resources from the user. Based on this assumption this paper presents in detail a user-configurable rule-system for controlling access to user resources.</p>

INDEX OF AUTHORS

Index of Authors

A lam, Mohammad Shah	123	H asan, Masum Z.	165
Arifuzzaman, Mohammad	123	Hassan, Suhaidi.....	199
Atzori, Luigi	67	Heer, Tobias.....	59
Azar, Christian.....	157	Hummen, René	59
B agula, Antoine	9, 25	I barrola, Eva.....	189
Bellavista, Paolo	141	Iera, Antonio	67
Belqasmi, Fatna	157	Inoue, Masugi	215
Ben Naila, Chedlia.....	115		
Botsivaly, Maria	239	J ere, Nobert Rangarirai	41
		Jodra, Jose Luis.....	189
C ardone, Giuseppe.....	141	K adobayashi, Youki.....	223
Chigona, Wallace	17	Kafle, Ved P.....	215
Chowdhury, S P.....	49	Kara, Nadjia.....	157
Corradi, Antonio.....	141	Kato, Akira	247
		Kharitonov, Daniel	75
F eliu Gutierrez, Michael Andres.....	253	Kim, Heejoong.....	247
Ferro, Armando	189	Kleinfeld, Robert	261
Figueira, Silvia	165	Kobari, Yasunaga	215
Firdhous, Mohamed.....	199	Kumar, Dhananjay	91
Foschini, Luca	141		
G hazali, Osman	199	L ehman, Leo	181
Glitho, Roch	157	Lora Girón, Mónica Alejandra.....	59
Gorges, Hannes.....	261		
Graham, Ian	83		
Gudreddi, Sree Lakshmi D.....	165		
Gueye, Bamba	131		

M akoza, Frank.....	17	T akahashi, Takeshi.....	223
Mano, Hiroshi.....	33	Terzoli, Alfredo	41, 149
Masinde, Muthoni.....	9	Thinyane, Mamello.....	41
Matsumoto, Mitsuji	115, 123	Tsietsi, Mosiuoa.....	149
Moltchanov, Boris	231	Tsukamoto, Katsutoshi	115
Morabito, Giacomo.....	67	Tucker, Lew	165
Morioka, Hitoshi.....	33	Tzavaras, Aris.....	239
Morrow, Monique.....	165		
Murata, Yoshitoshi	33	V an Olst, Rex.....	175
		Ventura, Neco	99, 253
N achimuthu Nallasamy, Kanagaraj.....	91	Viol, Nicolai	59
Nageshar, Nikesh.....	175		
Nakao, Koji	223	W akamori, Kazuhiko.....	115
Niang, Ibrahima	131	Wehrle, Klaus.	59
Nkoloma, Mayamiko.....	25	Wells, George	149
Nthontho, Mononts'i.....	49	Winberg, Simon.....	49
		Wirtz, Hanno.....	59
P erfecto, Cristina.....	189		
Pirmoradian, Mahdi.....	105	Z abala, Luis.....	489
Politis, Christos	105	Zennaro, Marco.....	25
Puglia, Virgilio	207		
S angharé, Papa Ousmane	131		
Soualhia, Mbarka.....	157		
Soubachov, Boyan.....	99		
Spyropoulos, Vasileios B.	239		
Sunahara, Hideki	247		

