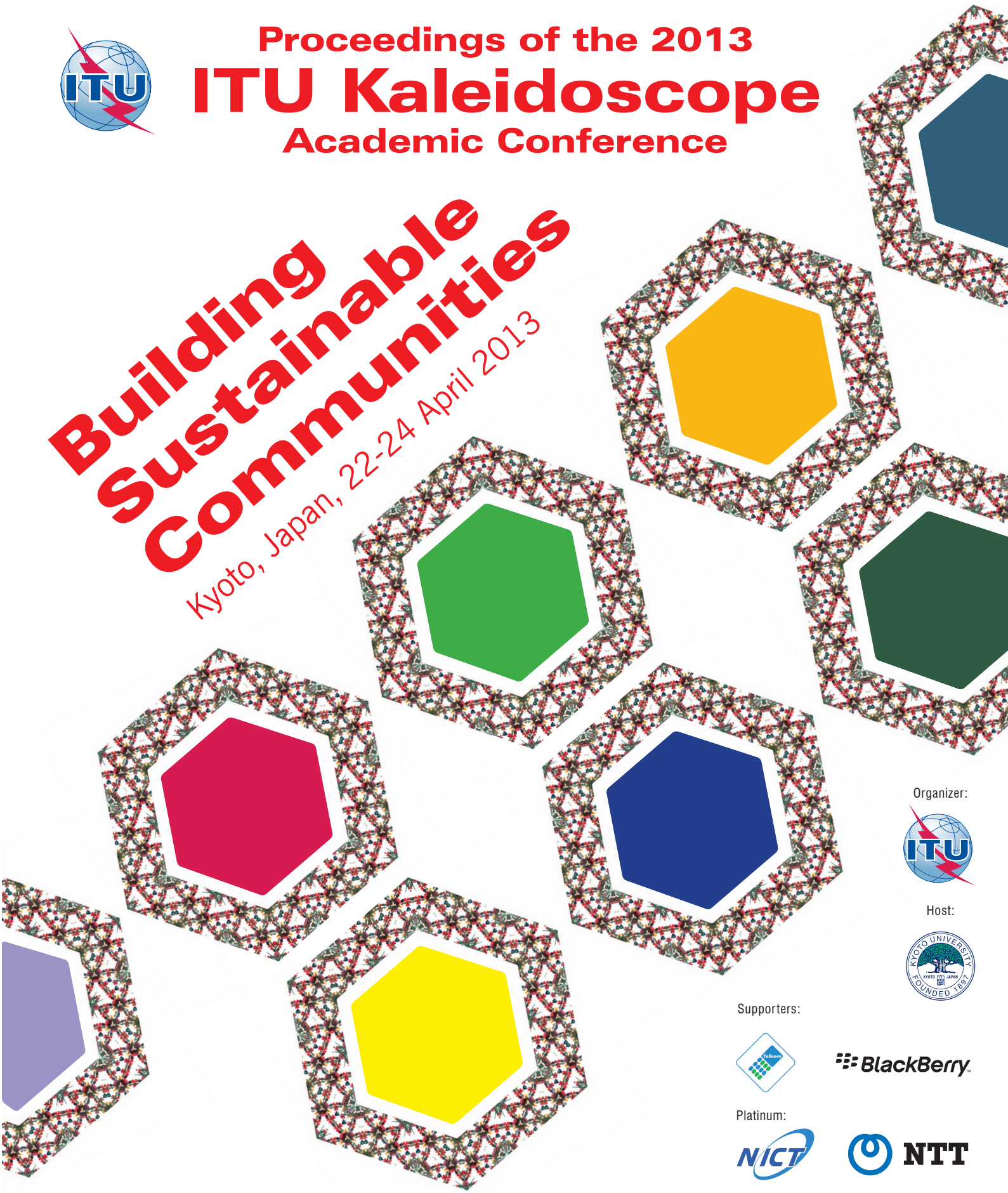
International Telecommunication Union

Proceedings of the 2013 ITU Kaleidoscope Academic Conference

Building Sustainable Communities

Kyoto, Japan, 22-24 April 2013



Organizer:

Host:

Supporters:

Platinum:

Silver:

Technical co-sponsors:

International Telecommunication Union

Proceedings of the 2013 ITU Kaleidoscope Academic Conference

Building Sustainable Communities

Kyoto, Japan, 22-24 April 2013

Supporters:

Platinum:

Partners:

Technical co-sponsors:

Silver:

Host:

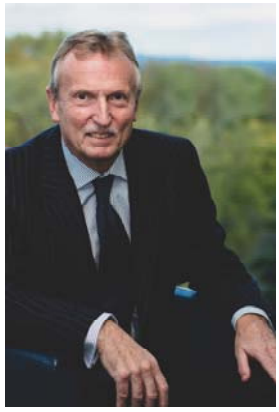
Organizer:

IEEE Catalogue Number: CFP1138E-ART

Disclaimer

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

Foreword



Malcolm Johnson
Director
ITU Telecommunication Standardization Sector

Academia's long-term approach to research is a major contributor to innovation and has played a significant role in the success of information and communication technologies (ICTs).

From ITU's beginnings in 1865 standardizing international telegraph service, academics have played a key role in ITU's mission to "Connect the World". As members of national delegations or advisors to private-sector organizations, academics have always contributed to ITU. With the creation of an academic membership category in November 2010, this contribution can now be given more visibility.

Academic members work alongside leading ICT engineers from industry, policy makers and business strategists from around the world. In just over two years our academic membership has grown to 55 universities of which 38 are members of ITU's Telecommunication Standardization Sector (ITU-T).

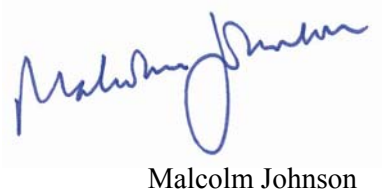
Kaleidoscope, ITU's flagship academic event, was launched in 2008 and aims to identify ICT research at an early stage so as to promote it and address the standardization needs that can launch the innovation onto the global market via internationally recognized ITU-T Recommendations.

The fifth Kaleidoscope conference took place at Kyoto University in Japan, a country that is the third largest global investor in research and development (R&D). Japan has long understood the importance of the ICT industry to its future economy; prioritizing R&D and reaching a stage today where it boasts the most advanced broadband infrastructure in the world and the world's highest penetration of Fiber to the Home (FTTH) subscribers.

Japan accounted for 11 of the 30 papers selected for presentation at the conference, a very impressive achievement in light of the competition faced in a total of 99 submissions from 37 countries. The submissions highlighted a wide range of inventive ICT applications reflecting the extent to which ICTs can address the challenge of building sustainable communities. It is clear that ICTs will continue transforming business processes and consumer behavior, and the standardization community will continue to rely on ITU as the place to develop global consensus on the technical framework upon which the Information Society is built.

Kaleidoscope 2013 further solidified ITU's relationship with academia. ITU is immensely grateful for the valuable contribution academia continues to make to the work in ITU, and Kaleidoscope's participants can be very proud of the role they played in making the conference a success.

On behalf of ITU my sincerest thanks go to Japan's Ministry of Internal Affairs and Communications (MIC) for making this event possible; our gracious hosts, Kyoto University; our generous sponsors, NICT, NTT, OKI, KDDI, NEC, Hitachi, Fujitsu, Mitsubishi, Huawei Japan, Telkom SA, and RIM; our tireless Steering Committee and Technical Programme Committee members; and of course our distinguished Chairman, Professor Hiroshi Matsumoto, President of Kyoto University.

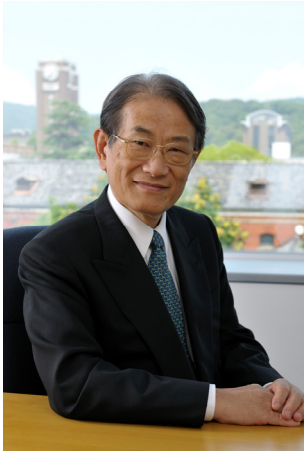


Malcolm Johnson

Director
ITU Telecommunication Standardization Sector

Chair's Message

Hiroshi Matsumoto
General Chair

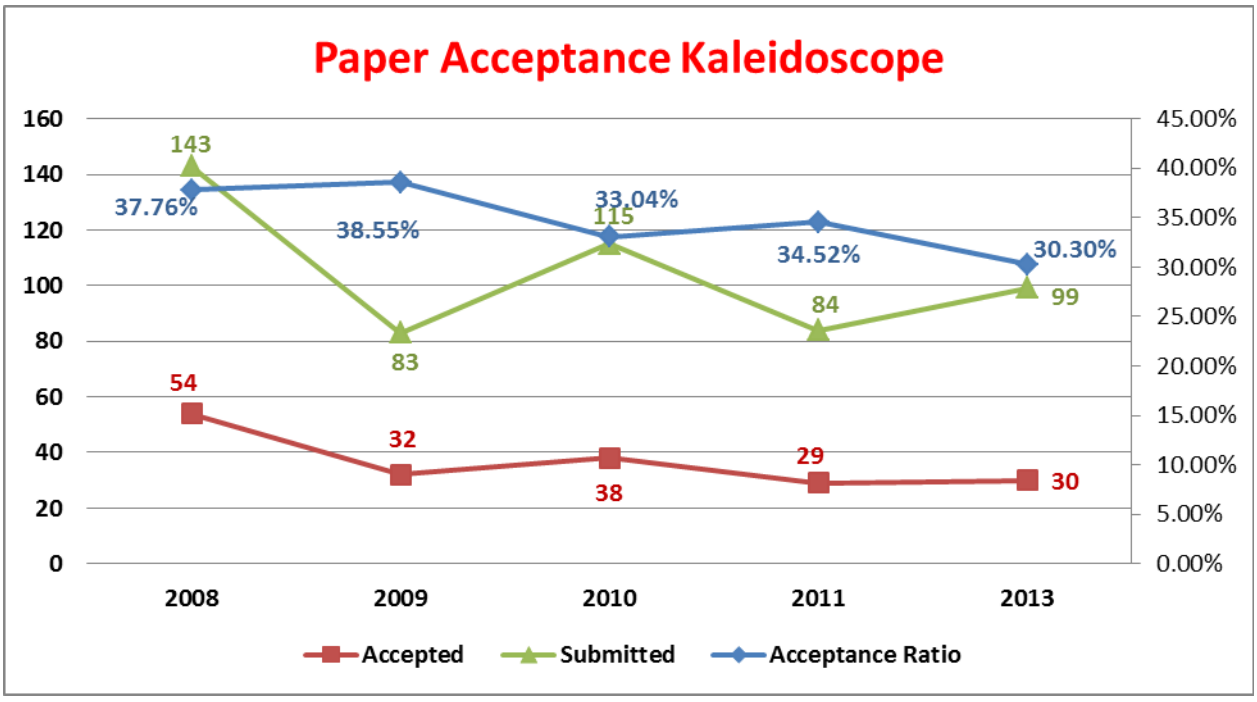


ITU initiated its Kaleidoscope series of conferences in 2008 to provide a forum for practitioners and researchers on the standardization of information and communication technologies (ICTs). The ITU academia membership category, established in 2010, gave further impetus to the Kaleidoscope series, and today the conference is an important event on the calendar of the academic community engaged in ICT research. I would like to express my appreciation to ITU for selecting Kyoto University as this year's host as well as my gratitude for the great volume of work ITU has undertaken to make this event possible.

I am honored to chair Kaleidoscope 2013, and in my view the event's theme, *Building Sustainable Communities*, could not be more appropriate given the social, economic, and environmental challenges of our time. Economies throughout the world must contend with global economic and financial crises, high unemployment rates, increasing pressure on energy resources, climate change, environmental disasters, and infrastructure straining under the weight of growing populations. ICT can however catalyze the transformations needed to meet these challenges.

ICT cuts across all industry sectors and thus has the potential to enact transformation across the whole spectrum of economic activity. The study of ICTs is today very much a multidisciplinary one, and this was clearly evident at Kaleidoscope 2013 where the papers presented touched on the activities of a wide range of industry sectors.

The Kaleidoscope 2013 Technical Programme Committee chaired by Kai Jakobs (RWTH Aachen University, Germany) selected 30 papers from the 99 submissions received from 37 countries. The committee selected papers on the basis of double-blind reviews with the help of 135 international experts, and also worked diligently to identify candidate papers for awards. As seen in the figure below, the high quality of submissions made the selection challenging, and I must express my gratitude to all reviewers and members of the Technical Programme Committee for their generous contribution of time and effort.



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

Countries	Lecture	Poster
France		1
Germany	2	
Japan	7	4
Jordan		1
India	1	
Italy	2	
Spain	1	
P.R. China	1	1
South Africa	1	2
Thailand	1	
Turkey		1
Ukraine		1
United Kingdom	1	
USA	1	
Uzbekistan		1

Alongside selected papers, Kaleidoscope 2013 also featured two distinguished keynote speakers, three invited papers, two special sessions, and two side events covering topics closely related to the conference. The Jules Verne's corner special session, now a staple of the event, took the theme, "Technological tsunami: Imagining a world without communications," offering a space to science fiction writers and futurists to imagine the consequences of a network collapse, the resource we rely on to support every aspect of our lives.

Our keynote speakers delivered insightful critiques of the ICT landscape and the importance of standardization. Makoto Nagao (Kyoto University, Japan) spoke on a *Digital Library for Creative and Sustainable Society.* Akihiro Nakao (Tokyo University, Japan) outlined *Deeply Programmable Network; Emerging Technologies For Network Virtualization And Software Defined Network (SDN).*

The authors of our three invited papers presented on sustainability, the intersection of ICTs and broadcast media, and the role of open standards as an important public resource to citizens of the Information Society. Shinichiro Haruyama (Keio University, Japan) provided an overview of his work on *Visible Light Communication Using Sustainable Led Lights.* Hisayuki Ohmata (NHK, Japan) introduced the audience to *Hybridcast: A New Media Experience by Integration of Broadcasting and Broadband.* William H. Melody (LIRNE.NET, Aalborg University Copenhagen, Denmark) sought to answer an interesting question: *Open Standards: a Shrinking Public Space in the Future Network Economy?*

Two side events tied in to Kaleidoscope 2013's theme with interactive discussions on future networks and education about standardization. In parallel with the conference, attendees were treated to a showcase of the future network R&D set to play a critical role in the *Building Sustainable Communities* event. Additionally, a Joint ITU-IEICE-CTIF-GISFI Workshop on Education about Standardization on April 25, 2013, which included the 2nd meeting of the TSB Director's Ad hoc Group on Education about Standardization, provided an overview of how standardization is currently approached in academic curricula, and fostered an exchange of ideas on how universities could scale up the production of standards-minded graduates across academic disciplines.

In line with an ITU agreement with IEEE Communications Society (ComSoc), selected papers from each year's conference are considered for publication in a special feature section of IEEE Communications Magazine. The Kaleidoscope 2013 papers are tentatively scheduled for publication in the August 2013 issue. In addition, special issues of the International Journal of Technology Marketing (IJTMKT) and the International Journal of IT Standards and Standardization Research (IJITSR) are interested in publishing revised versions of Kaleidoscope papers.

All accepted papers are accessible through the IEEE Xplore digital library. The Proceedings from 2009 onwards can be downloaded free of charge from http://itu-kaleidoscope.org.

In closing, I would like to thank our hosts, Japan's Ministry of Internal Affairs and Communications (MIC) and Kyoto University; the IEEE, IEEE ComSoc, and the Institute of Electronics, Information and Communication Engineers of Japan (IEICE), for their technical co-sponsorship. I would also like to thank Telkom SA for the cash awards offered to the best papers; RIM for the blackberry devices presented to the authors of the two best papers among recipients of the Young Author Recognition Certificate; NTT, OKI, KDDI, NEC, Hitachi, Fujitsu, Mitsubishi and Huawei Japan for their logistical support; NICT for both logistical support and coordination of the future R&D showcase; our supportive partners, TTC, ITU Association of Japan, Waseda University, The Institute of Image Electronics Engineers of Japan, EURAS; and finally, Alessia Magliarditi and her team from ITU-T for playing the leading role in Kaleidoscope 2013's remarkable success.

Hiroshi Matsumoto

General Chair

TABLE OF CONTENTS

Page

Poster Session

COMMITTEES

Steering Committee

- General Chairman: Hiroshi Matsumoto (President, Kyoto University, Japan)

- Christoph Dosch (IRT GmbH, Germany)
- Kai Jakobs (RWTH Aachen University, Germany)
- Mostafa Hashem Sherif (AT&T, USA)
- Alfredo Terzoli (Rhodes University, South Africa)

Host Committee

- Chairman: Tatsuro Takahashi (Kyoto University, Japan)

- Tohru Asami (University of Tokyo, Japan)
- Yoshikazu Ikeda (Otani University, Japan)
- Yasuyuki Koga (NICT, Japan)
- Yoichi Maeda (TTC, Japan)
- Mitsuji Matsumoto (Waseda University, Japan)
- Tetsutaro Uehara (Ministry of Internal Affairs and Communication, Japan)

Secretariat

- Alessia Magliarditi, Project Head

- Martin Adolph, Project Technical Advisor
- Leslie Jones, Administrative support
- Pablo Palacios, Administrative support
- Simão Campos Neto, Project Advisor

Technical Programme Committee

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

- Marcelo F. Abbade (Pontifical Catholic University in Campinas, Brazil)
- Martin Adolph (ITU-T, Switzerland)
- Sachin Agrawal (Samsung, India)
- Denis Andreev (ITU-T, Switzerland)
- Chaodit Aswakul (Chulanlongkorn University, Thailand)
- Jonathan Bachens (Old Dominion University, USA)
- Abdelmalik Bachir (Imperial College, United Kingdom)
- Bartosz Balis (AGH University of Science and Technology Krakow, Poland)
- Roxana Barrantes (Instituto de Estudios Peruanos, Peru)
- Bernd Bechow (Franhover FOKUS, Germany)
- Abdelmoula Bekkali (QU Wireless Innovation Center, Qatar)
- Rudi Bekkers (TU Eindhovel, The Netherlands)
- Paolo Bellavista (University of Bologna, Italy)
- Vitor Bernardo (University of Coimbra, Portugal)
- Jose Everardo Bessa Maia (State University of Ceará, Brazil)
- Mauro Biagi (Sapienza University of Rome, Italy)
- Knut Blind (TU Berlin, Germany)
- Niklas Blum (Fraunhofer Institute FOKUS, Germany)
- Mario Bourgoult (École Polytechnique de Montréal Canada)
- Michael Bove (MIT, USA)
- Cagatay Buyukkoc (AT&T, USA)
- Marco Carugi (ZTE Corporation, France)
- Marcelo Carvalho (University of Brasilia, Brazil)
- Zehra Cataltepe (Istanbul Technical University, Turkey)
- Isabella Cerutti (Scuola Superiore Sant'Anna, Italy)
- Omar Cheikhrouhou (Higher Institute of Technological Studies, Tunisia)
- 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)
- Laurence Delina (University of Massachusets, USA)
- Marc De Leenheer (Ghent University, Belgium)
- Alvaro Augusto de Medeiros (Federal University of Juiz de Fora, Brazil)

- Ilker Demirkol (Universitat Politecnica de Catalunya, Spain)
- José María Díaz Batanero (ITU, Switzerland)
- Antoine Dore (ITU, Switzerland)
- Tineke Mirjam Egyedi (Delft University of Technology, The Netherlands)
- Dmitry Epstein (Cornell University, USA)
- José Ewerton Farias (Federal University of Campina Grande UFCG, Brazil)
- Ahmed Fathy Atya (University of California Riverside, USA)
- David Faulkner (Climate Associates, United Kingdom)
- Armando Ferro Vázquez (ETSI de Bilbao, Spain)
- Erwin 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)
- Ivan Ganchev (University of Limerick, Ireland)
- Molka Gharbaoui (Scuola Superiore Sant'Anna, Italy)
- Katja Gilly (Miguel Hernandez University, Spain)
- Visvasuresh Victor Govindaswamy (Texas A&M University, USA)
- Ian Graham (University of Edinburgh, United Kingdom)
- Chris Guy (The University of Reading, United Kingdom)
- Ruan He (Orange Lab, France)
- Richard Heeks (University of Manchester, United Kingdom)
- Courtney Humphries (University of Mississippi, USA)
- Eva Ibarrola (University of the Basque Country, Spain)
- 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)
- Patrick Kalas (FAO, Switzerland)
- Kamugisha Kazaura (Tanzania Telecommunications Company Limited, Tanzania)
- Tim Kelly (World Bank, USA)
- Adrian Kliks (Poznan University, Poland)
- Masafumi Koga (Oita University, Japan)
- Junko Koisumi (ITU-R, Switzerland)
- Stephan Kopsell (TU Dresden, Germany)
- Andrej Kos (University of Ljubljana, Slovenia)
- Katarzyna Kosek-Szott (AGH University of Science and Technology, Poland)
- Ken Krechmer (University of Colorado, USA)
- Sajeesh Kumar (University of Tennessee, USA)
- Richard Labelle (The Aylmer Group, Canada)

- 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)
- Jean-Paul Lemaire (Université Denis Dideró, France)
- Fidel Liberal (ETSI de Bilbao, Spain)
- Sang-Kyu Lim (ETRI, Korea)
- Luigi Logrippo (Université de Québec en Outaouais, Canada)
- Waslon Araujo Lopes (Federal University of Campina Grande, Brazil)
- Jose Giovanny López Perafán (University of Cauca, Colombia)
- Giovani Mancilla (Universidad Distrital, Colombia)
- Didier Marquet (Orange Labs, France)
- Mitsuji Matsumoto (Waseda University, Japan)
- Venkatesen Mauree (ITU-T, Switzerland)
- Arturas Medeisis (Vilnius Gediminas Technical University, Lithuania)
- Pablo Menoni (ANTEL, Uruguay)
- Thomas Meuser (Krefeld University of Applied Sciences, Germany)
- Anne Mione (Université Montpellier 1, France)
- Werner Mohr (Nokia Siemens Networks, Germany)
- Antonella Molinaro (Università degli Studi Mediterranea di Reggio Calabria, Italy)
- Ashwinkumar Motagi (Visvesvaraya Technical University, India)
- Yoshitoshi Murata (Iwate Prefectural University, Japan)
- Tae Oh (Rochester Institute of Technology, USA)
- Fumitaka Ono (Tokyo Polytechnic University, Japan)
- David Palma (University of Coimbra, Portugal)
- Mukaddim Pathan (CSIRO ICT Center, Australia)
- Henrique Pequeno (Federal University of Ceará, Brazil)
- Francisco Portelinha (University of Campinas, Brazil)
- Louis Pouzin (Eurolinc, France)
- Francisco Ramos (Universidad Politécnica de Valencia, Spain)
- Greg Ratta (ITU-T, Switzerland)
- Ramona Rednic (Coventry University, United Kingdom)
- Felipe Rudge Barbosa (University of Campinas, Brazil)
- Chiara Sammarco (University 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)
- Reijo Savola (VTT Technical Research Centre of Finland, Finland)

- Sanaa Sharafeddine (Lebanese American University, Lebanon)
- Helmut Schink (Nokia Siemens Networks, Germany)
- Ulrich Schoen (Germany)
- Florian Schreiner (Fraunhofer Institute FOKUS, Germany)
- DongBack Seo (University of Groningen, The Netherlands)
- Robert Shaw (ITU-D, Switzerland)
- Mostafa Hashem Sherif (AT&T, USA)
- Irina Sineva (Moscow Technical University of Communications and Informatics, Russia)
- Pierre Siohan (France Telecom, France)
- Stanimir Stojanov (Plovdiv University, Bulgaria)
- Ewan Sutherland (University of Narum, Belgium)
- Szymon Szott (AGH University of Science and Technology, Poland)
- Kenzo Takahashi (University of Fukui, Japan)
- Alfredo Terzoli (Rhodes University and University of Fort Hare, South Africa)
- Andrea Tonello (University of Udine, Italy)
- Ualsher Tukeyev (Al-Farabi Kazakh National University, Kazakhstan)
- Kurt Tutschku (University of Vienna, Austria)
- Hiromi Ueda (Tokyo University of Technology, Japan)
- Manuel Urueña (Universidad Carlos III de Madrid, Spain)
- Geerten van de Kaa (University of Delft, The Netherlands)
- Jari Veijalainen (University of Jyvaskyla, Finland)
- John Visser (Canada)
- Marc Waldman (Manhattan College, USA)
- Wilson Yamaguti (University of Mogi das Cruzes, Brazil)
- Ahmed Zeddam (Orange Labs, France)

KEYNOTE SUMMARIES

DIGITAL LIBRARY FOR CREATIVE AND SUSTAINABLE SOCIETY

Makoto Nagao

Prof. Em. Kyoto University, Japan

Broadband networks are becoming common worldwide, and people can communicate with each other overcoming the language barrier using machine translation and other tools. Libraries are becoming digital libraries where information can be obtained searching the contents of books and journals stored all over the world. Digital library technology contributes to improving the sustainability of society by reducing the amount of paper consumed and, most importantly, by preserving cultural assets in digital archives.

Today, reading devices can handle multimedia information and allow people to interact with each other and even with authors of books and journals. Digital libraries provide a “virtual common space” where people learn, communicate, create ideas, and share mutual interests. They are essential for creating a sustainable worldwide community. For example, the Japanese National Library (Diet Library) is collecting and connecting not only collected records of the Great East Japan Earthquake but also restoring/reconstructing records in digital format in close cooperation with governmental ministries. The resulting information archive will be open to anyone and will help reduce the aftermath of similar disasters in the future.

DEEPLY PROGRAMMABLE NETWORK; EMERGING TECHNOLOGIES FOR NETWORK VIRTUALIZATION AND SOFTWARE DEFINED NETWORK (SDN)

Akihiro Nakao

Associate Professor, The University of Tokyo, Japan

This presentation introduces the recent trends in network virtualization and software defined network (SDN). We expect these trends to help establish sustainable communities to meet challenges of today's and tomorrow's communication infrastructures. These infrastructures, in turn, will support various human activities and thus contribute to the creation of human-oriented technologies.

Research and development in these fields are considered as "high growth areas" for realizing the future Internet worldwide. We observed that these research areas have not only been promoted in academia, but that some of them have also been rapidly commercialized and have triggered standardization activities in several standardization bodies, such as ITU-T for the network virtualization framework, ETSI for network function virtualization and ONF for openflow.

This presentation gives an overview of various research activities on network virtualization and SDN in the world, clarifies the difference and the close interaction between them, and discusses the recent research direction towards merging and extending them into enabling "deep programmability" within the network. It also introduces global scale international joint research trials such as "slice around the world" among US, South America, Europe, and Japan. In these trials, we reserve multiple "slices" of computational, storage and network resources across the world and enable deep programmability inside these slices. This allows us to design, deploy and experiment with new communication technologies in each slice without interference between slices. We expect these series of international activities to eventually formulate the standardization of federating various technologies emerging from all over the world.

SESSION 1

INFRASTRUCTURES AND PLATFORMS TO SUPPORT COMMUNITIES

SUSTAINING LIFE DURING THE EARLY STAGES OF DISASTER RELIEF WITH A FRUGAL INFORMATION SYSTEM: LEARNING FROM THE GREAT EAST JAPAN EARTHQUAKE

Mihoko Sakurai — Keio University — sakuram@sfc.keio.ac.jp
Richard T. Watson — University of Georgia — rwatson@terry.uga.edu
Chon Abraham — The College of William & Mary — Chon.Abraham@business.wm.edu
Jiro Kokuryo — Keio University — jkokuryo@sfc.keio.ac.jp

ABSTRACT

Important lessons for responding to a large-scale disaster can be gleaned from the March 11, 2011 Great East Japan earthquake and tsunami. The failure of the electrical power system and the resultant loss of information communication and processing capability severely constrained the recovery work of many municipalities. It was difficult for supporting organizations to collect and share information. A frugal Information System (IS) designed around the four U-constructs is suggested as a solution to handle the early stages of disaster relief. This paper focuses on the most frequently available device, the cellular phone, as the foundation for a frugal IS for disaster relief. Familiar and available tools place minimal stress on an already stressed system.

Keywords— Disaster, Frugal Information System, U-constructs, Cellular phone, Emergency

1. AN INFORMATION SYSTEM FOR DISASTERS

March 11, 2011 (3.11) marked a day of devastation for East Japan. An earthquake and tsunami immobilized much of the technologically advanced nation. Nature managed to subdue the prowess of some of humans' greatest inventions over the course of the disaster. In its wake of a catastrophe, the first few hours and days are critical but typically thwarted by communication and power interruptions that threaten the sustainment and sanctity of life immediately and long-term through the creation of lasting environmental hazards, such as radiation leaks or chemical spills. Thus, attending to life and death tasks such as confirming residents' whereabouts and conditions, dispatching aid and supplies, and attending to the most critical environmental threats become the pre-imminent objectives.

Information systems, relying on a continuous power supply, are typically not operational immediately following a disaster. Yet, the information accessible via these systems regarding where people reside is vital to a disaster relief effort. Imagine a team trying to enter an impacted area and not knowing which establishments are regularly occupied, how many inhabitants to expect in a household, or the environmental impact of a chemical leak. Such information is not always accessible to relief teams immediately following a disaster. Displaced families often have no means of relaying information about their status to family and friends. Such information is also invaluable in helping relief teams decide where to search for survivors, assessing what basic supplies are needed to support the survivors, and where to deliver them.

The lack of communication capabilities and information processing resources needs to be swiftly handled by deploying a system with basic capabilities and sufficient bandwidth to meet the most pressing needs while leveraging a basic communication tool most people possess. It is what some scholars identify as a frugal Information System (IS) [1], which embodies a set of characteristics that enables swift and effective deployment of a very limited IS designed, in this case, to gather and distribute the minimal but critical information to maximize survival rates.

After reviewing a comprehensive report on the impact of 3.11 on Information and Communication Technologies (ICT) at the Municipal Government level [2], we propose a design for a cell phone based frugal IS to be deployed during the first phase of disaster relief. The article is structured as follows: (1) an overview within 13 municipalities of the 3.11 crisis and its impact on ICT and emergency response capability; (2) analysis of the implications and a broad statement of requirements for a better solution, (3) conceptualization of a frugal IS and presentation of the four information drives to describe the design of an IS for first phase emergency response, (4) matching the design against the specific needs of 3.11 in the prior section, and (5) conclusion.

2. REACTION OF THE DISASTER

The Great East Japan Earthquake occurred at 14:46 Japan Standard Time on March 11, 2011. At a Richter scale of 9.0, it was the largest earthquake on record for Japan. More damaging than the quake itself, a tsunami of up to 40 meters hit the coastline, devastating cities and towns. The Fire and Disaster Management Agency reported 16,131 deaths, 5,994 injuries and 3,240 missing as of January 2012. It also reported 128,497 houses totally lost and more than 900,000 partially destroyed.

The tsunami also destroyed all power supply to the cooling systems of the nuclear power plant in Fukushima causing a meltdown. As of January 1, 2012, 159,124 people from Fukushima had still not returned to their homes. From an ICT perspective, disruption of communication and loss of IS capabilities for operations were a significant hindrance to effective and rapid recovery. People and organizations were deprived of the information and processing systems required to deal with the situation. The effect was particularly noticeable at the municipal government level because it is their responsibility to support their citizens in an emergency.

Loss of power, termination of telecommunications services (primarily due to the power loss to base stations), and data loss were the key problems. The extent of the damage far exceeded any predictions, and recovery efforts had to be made outside of prepared procedures. Rescue workers had to rely on their judgment and the capabilities of personnel on the scene. Loss of power and communication especially affected all initial relief efforts.

Some local governments were equipped with emergency power generators, but they covered only basic needs and were not sufficient to enable ICT to function at the needed level. Recovery of commercial power required more than four months in some areas.

Telecommunication was also disrupted as a result of the power outage. Many switching facilities were lost and cables were damaged. Recovery time varied depending on the damage. Recovery of fixed line telephony and Internet reconnection required from one week to one month.

The Government Disaster Management Radio Communication Network and the satellite phone system survived. Municipal governments used these to request support and to organize relief. Their usefulness was, however, limited in that the former connected few destinations and use of the latter was limited due to the high cost. In addition, the high power usage of satellite phones became problematic under constrained power supplies.

Among individuals, cellular phones were the most widely used communications tools. The service was available in most areas until the batteries of the base stations died. Conversation was mostly impossible, but packetized mail systems could be used immediately after the quake. Many municipal government officials learned about the coming of the tsunami with TV tuners on their private phones. Base station batteries ran out by the following day (3.12). They were restored quicker than fixed line communication lines, but nevertheless were out of service for one to two weeks.

In summary, ICT infrastructure destruction severely limited communications bandwidth. Under such circumstances, the primary operations of municipal governments following the disaster were:

1. Confirming the whereabouts and safety of residents
2. Establishing and operating evacuation centers
3. Transport and management of relief goods
4. Support of evacuees, and creating evacuee lists
5. Issue of Disaster-victim Certificates

These tasks are different from the daily operations of municipal government, and other tasks were suspended to meet the impending and mounting needs of citizens. The most important task was to establish lists of residents to guide rescue operations, which is done manually because of power and communication failures. Creating manual systems was burdensome for overworked municipal government officials.

3. ANALYSIS OF THE CASES, IMPLICATIONS, AND A BROAD STATEMENT OF REQUIREMENTS FOR A BETTER SOLUTION

Constrained by meager ICT resources, shortage of personnel and the impossibility (and thus failure) of information sharing among the agencies were the main problems facing municipal governments as they took on the task of saving and supporting citizens.

In the initial phase, each evacuation center had to make its evacuees' lists. Each city had up to several hundred evacuation centers and fragmentation of databases became an issue. Although there was a strong need for sharing the manually created data lists, this was very difficult because some of the necessary infrastructure had been obsoleted years ago (e.g., carbon paper). This lack of a single view of the data was particularly problematic as many of the evacuees moved in search of better conditions and food.

Information sharing between municipal governments was another major bottleneck. The lack of residential record data hindered police and self-defense troops lifesaving operations. Similar communication failures existed among logistics organizations. The inland city of Tono, which played a major role as a neighboring city to many of the most seriously affected coastal cities, suffered from inaccurate and/or outdated information on the supply needs of its neighbors. Many donated supplies sat unused while people lacked food and clothing.

To enable delivery of relief to a disaster's refugees, evacuee lists should include a person's residential status to check eligibility. (The fraudulent receipt of relief was a sad reality that had to be addressed.) While tentative relief is provided even for those who cannot be verified, other more costly operations such as the allocation of temporary housing units and loans for housing reconstruction require eligibility confirmation. The baseline data are the residential records maintained by municipal governments. While they are responsible for issuing relief certificates, it was not a critical first response issue, but something to handle later. However, there was a need to collect some data during the initial response to avoid later fraud.

The temporal integration of the victim database was another big issue. Initially, there was simply a need to record life or death information together with a residential address. Later, information such as damage to a person's home had to be added. This database had to be maintained as evacuees relocated and various parties in dispersed locations amended it.

In spite of the importance of maintaining an integrated database, there was fragmentation and duplication at many locations and the establishment of an ad hoc database. In addition to the power and communication loss, stringent security hindered the use of official residential records.

Shortage of personnel was another major issue. Resident service counters, such as those for the issue of Disaster Victim Certificates, often received many applications at the same time and municipal personnel were not always able to handle the workload expeditiously. In some municipalities, there were days when over 500 residents queued at temporary disaster response counters, and municipal government employees worked indefatigably.

In retrospect, it is easy to identify the need for openness and compatibility among the various systems so that data from them can be integrated rapidly when the need arises. The reality in a chaotic scene is that people are forced to use whatever is available, including pen and paper.

The availability of resources varied among locations. Some needed to function totally without power and communications. Others were fortunate to have access to working ICT equipment. Consequently, we cannot assume a single system can deal with such diverse situations. Redundancy is necessary to handle unanticipated circumstances both in form and scale.

Integrated databases, nevertheless, are critically important. If access to the official residential records could have been established more quickly in many locations, data fragmentation would have been reduced. The idea of supporting relief operations by ICT has existed since the 1995 Kobe earthquake. The city of Nishinomiya created an integrated package for emergency response support, and it was freely shared with other municipal governments. The town of Minami-Sanriku originally allowed each department to develop its relief support systems, but it later switched to Nishinomiya's package when it recognized the importance of database integration.

4. A DESIGN FOR A FRUGAL IS FOR THE FIRST PHASE OF EMERGENCY RELIEF

As the prior analysis has shown, immediately following a disaster there is often a lack of communication capabilities and information processing resources. This situation might be best handled by quickly deploying a frugal IS, which is defined as, "…*an information system that is developed and deployed with minimal resources to meet the preeminent goal of the client"* [1]. Because of the likelihood of limited communication and processing resources, the IS should use minimal bandwidth to send and receive a few critical messages. Furthermore, the frugal IS should focus on the dominant problems of the first few days of a disaster – determining the location and condition of the victims, helping them to get emergency support, and identifying the missing. Such information is essential to managing the relief effort and informing concerned relatives and friends.

4.1 Structural factors

Many parties have a potential role in emergency relief, from neighbors to international agencies, such as the Red Cross. The major players in most cases are the national government and local governments, such as municipalities, because they have prime but differing responsibilities for their citizens. This was the case in Japan, where there are 47 prefectures and 1742 municipalities. The size of Japan's municipalities varies considerably, with Osaka and Yokohama having a few million and small villages less than one thousand residents.

The varied responsibilities exemplifies the classic trade-off between efficiency and effectiveness that are reflected in all organizational questions related to the degree of centralization (e.g., [3]). Many organizations have found that the creation of an enterprise IS architecture involves working out a design that strives to balance global efficiency with local effectiveness [4].

The extent of the damage far exceeded the capabilities of municipal governments, and it took a national and international mobilization of resources (the US military played a major role). At the same time, municipal governments played a critical role as they are the agencies closest to the residents. They have firsthand knowledge of the people and resources in their area, they are already at the disaster scene because they live there, and they are familiar with local customs and dialects. Such knowledge is critical for outside rescuers to function effectively.

The role of the national government, in our assessment, includes supporting the municipal governments to play their critical role by establishing a national disaster recovery infrastructure and standards. It needs to ensure that as soon as possible after a disaster, municipalities have the minimal resources they need for a local response. As the focus is on sustaining life immediately after a disaster, we concentrate on what is required to create an efficient national frugal IS that can be deployed to support effective local relief.

4.2 Foundations of a frugal national emergency IS

The four u-constructs [5, 6] are a basis for establishing system design principles (Table 1). We consider each of the drives, starting with the universality construct because settling on a common communication platform is the foundation for satisfying the other u-constructs.

Table 1. The information drives [6]

Drive	Definition
Ubiquity	The drive to access information unconstrained by time and space
Uniqueness	The drive to know precisely the characteristics and location of a person or entity
Unison	The drive for information consistency
Universality	The drive to overcome the friction of information systems' incompatibilities

4.2.1 Universality

A frugal IS must be compatible with existing systems that disaster victims are likely to possess and should require them to acquire minimal new technology, if any. Smart phones, which are becoming very common, are portable, battery powered, often within the owners' easy reach, incorporate a GPS, and have the capability to run apps. Because global cell phone adoption is 87 percent and 79 percent in the developed and developing world, respectively [7], it is appropriate to make the mobile phone the standard platform for a frugal emergency relief IS.

To further enhance universality, all phones should have a pre-installed emergency app that can be remotely updated as required. Such an app should be created and maintained by the appropriate national emergency agency.

Language is a major form of information system friction. People in a disaster zone might well speak a different language from some relief workers. Thus, we envisage an emergency app that transmits a code (Table 2) as well as the GPS location, and the sender's phone number. A code can be readily converted into the language of the receiver. A code is frugal in terms of bandwidth requirements. Similarly, there can be another set of frugal messages to support communication with victims.

Table 2. Universal message coding

Message	Code
I need water	1
I need food	2
I am injured and cannot move	3
:…	…

Cell phones require power, but most should have sufficient battery power to continue to operate for a day or so. In addition, we advocate that a hand-powered phone battery charger be added to the recommended household emergency kit. Municipalities should also establish an emergency phone charging resource because their relief workers would likely be using cell phones during the first stage of recovery.

4.2.2 Ubiquity

Victims and relief teams need ubiquitous access to essential information wherever they might be in the disaster zone. Because it is likely that essential infrastructure has been destroyed, there is a need to quickly deploy an alternative wireless communication system that is compatible with the victims' cell phones and can cover the relief area. A cell phone network must provide both coverage and bandwidth to service customers. The GSM specification supports communications up to 35 kilometers, and a frugal IS need provide only minimal coverage because frugal messaging minimizes the need for bandwidth.

Aerial deployment of base stations is probably the quickest way to create coverage. The aerial options might include drones, tethered balloons, blimps, and aircraft. Another alternative is to deploy offshore ships or river barges. Such resources can be pre-commissioned and ready to position and activate. Consideration also needs to be given to the power required to operate the temporary base stations and their connection to the operational remnants of the existing cell network. The goal is create a minimal ubiquitous access system as soon as possible.

It is beyond the scope of this paper to consider and cost the communication alternatives, but it is important to keep in mind that the disaster area for 3.11 covered over 800 kilometers. Sustaining human life and the environment is expensive for mass scale disasters.

4.2.3 Uniqueness

The identity, location, and condition of victims must be established as soon as possible. In addition, there is a need to identify locations that are hazardous (e.g., a radiation leak) or require a rescue team (e.g., a demolished building). Because the cell phone is typically a personal device, there is a one-to-one mapping of a cell phone number to a person.[1] What is often missing, as 3.11 demonstrated, is a national, or even a local, government-based, integrated database that records this mapping along with the person's home and work addresses and some minimal personal (e.g., date of birth) and family information (e.g., details of children), and critical health information (e.g., diabetic). Such information means that the general and unique needs of a disaster area can be quickly determined (e.g., the number of doses of insulin). While a disaster is often seen as a national tragedy, it is a combination of many individual calamities that frequently have unique needs.

4.2.4 Unison

Information about victims and where they live can be scattered across multiple databases (automated and manual within the community, in regions, and in national registries) and during an emergency there is often a need to integrate

[1] The full identifier, country code plus phone number, needs to be used to allow for visitors.

these data. Such integration, however, should be prepared prior to the need, as identified in the prior discussion on uniqueness. The tension between the general need for privacy and the specific needs of a disaster needs to be accommodated in line with existing or new national laws.

Unison also comes into play in another way. Keeping details of victims is critical to managing the recovery. Each relief station needs to keep track of who is housed where, receiving assistance, and so forth. Relying on paper records can result in a lack of information consistency. Written names can be miskeyed and databases then become unreliable. The recording process needs to be fast, efficient, and designed to maintain unison across recording events as victims move.

The cell phone is the key to maintaining unison because it is a person's electronic identifier. The emergency app needs a feature that transmits its phone number to the phone (e.g., Bluetooth) of an appropriate relief person in the field or at a relief center. A relief worker's list of phone numbers and coordinates can be uploaded automatically to a computer system.

Unison is maintained by the creation of a single integrated database and ensuring consistent identification of entered data. In the case of an emergency, the key identifier is the phone number of the victim.

We also need to consider the case of people who do not have a cell phone with them or do not own one (e.g., children). Those in the first category should be able to remember their phone number and this can be manually captured. The second category can be identified by adding a digit to the phone number (e.g., +1 999-999-999-1 for the youngest child). A similar workaround needs to be developed for handling those who do not own a cell phone or who are dependent on others for use of communication devices such as some elderly people.

A pre-established schema for the integrated database on survivor information should exist. In particular, preparation for integration of official residential data and carrier user data will be essential, but will require legal framework in many countries if that is to happen. Collaboration of public and private sectors thus becomes an important element of applying frugal systems for disaster relief.

The lessons from 3.11 underscore the importance of unison, and this is perhaps the u-construct that should be pursued with the most vigor. However, without the other constructs in place, it is not feasible to have unison.

5. APPLYING THE PROPOSED FRUGAL IS TO THE GREAT EAST JAPAN DISASTER

While hindsight is not the best method of testing a design, it is the best alternative at this point. Consequently, we iteratively refined the design by examining the performance of five key tasks, and we now discuss how the proposed frugal IS can handle these critical issues.

5.1 Confirming the whereabouts and safety of residents

Achievement of this task requires the rapid creation of a low bandwidth ubiquitous communications network so that residents can use an app on their cell phones to self-report their condition and environmental hazards. These reports will be self-identified by phone number and GPS coordinates and as such can be linked to residential data.

5.2 Establishing and operating evacuation centers

Once a municipality has physically established evacuation centers, it needs a frugal IS to record data about who arrives at or leaves a center. Again, the universal nature of the cell phone means it becomes a key device for recording such data by Bluetooth (for example, data exchange between survivors' and relief workers' phones). These data can be transmitted to the integrated database as bandwidth becomes available. Local centers should be able to draw on this database using a preset small number of queries that require minimal ICT resources.

As power and communication systems are established, there must be a smooth migration to computer-based, rather than phone-based, information systems.

5.3 Transport and management of relief goods

The response effort needs to ensure that it delivers with precision what is required rapidly. As disruption of the road system is likely, those delivering supplies need a separate frugal IS to be able to report failures in the road network and get advice on new routes. The same u-constructs can guide design of this system. We imagine a system where a copy of the existing road network is dynamically updated with the latest availability data and used to create dynamically new routes.

5.4 Support of evacuees and creating evacuee lists

The power of a phone-based system is that data on citizen whereabouts and condition can be collected in the field as well as at evacuation centers. Thus, rescue teams might well include a data officer, carrying a high capacity battery, who can identify each evacuee's needs before they reach a relief center so the center has time to gain some advanced warning of the types of support required. In an emergency, a few minutes extra can be life saving.

The actions taken to confirm the whereabouts of citizens become the foundation of evacuee lists. As these lists will be transmitted to the cloud as soon as bandwidth is available, they can be made generally available, with appropriate privacy safeguards, to inform relatives.

5.5 Issuing of disaster victim certificates

The frugal IS captures essential data about each victim and tracks movement between evacuee centers. In effect, it creates an audit trail that can be used to distinguish the genuine victim from the fake. Once a person's right to a certificate has been verified, this could be issued electronically (e.g., send a QR code to their phone) as well as recorded in the database for subsequent reference. The issuing of victim certificates illustrates the power of frugal thinking. Use what the victim is very likely to already have and avoid additional resources (e.g., printers).

Of course, there needs to be a workaround for those who do not have a phone, and in this case a possible frugal solution could be based on small printers that can generate a paper-based QR code.

6. CONCLUSION

In summary, we propose

- Use what people already have, cell phones, as the foundation. Unfamiliar devices especially prepared for disasters were not used.
- Governments should invest in the resources to create quickly the infrastructure to sustain a frugal IS.
- The public and private sectors must collaborate to ensure the effective integration of the government disaster infrastructure and privately owned mobile devices.

It is worth noting that no single solution solved the problem for 3.11, but we believe a phone-based solution, while maybe not the complete answer, will be a major advance in disaster relief. There is, however, a need for compatibilities at various levels so that various frugal systems (including handwriting) can be integrated to serve the ultimate task of effective and efficient relief.

A disaster creates a massive disruption of human life, physical structures, and ICT assets. Information about the state of this disruption is usually critical if the inevitably scarce rescue and recovery resources are to be used effectively. A frugal IS is an initial step towards providing such information and should be a central part of both national and local disaster plans. Japan's experience in building and deploying such a frugal IS could greatly enhance disaster recovery across the globe.

REFERENCES

[1] Watson, R.T., K.N. Kunene, and M.S. Islam, "Frugal IS," *Information Technology for Development,* forthcoming.

[2] Sakurai, M. and J. Kokuryo, "Municipal Government ICT in 3.11 Crisis: Lessons from the Great East Japan Earthquake and Tsunami Crisis," Berkman Center for Internet & Society in partnership with Keio University: Cambridge, MA., 2012

[3] Bartlett, C.A. and S. Ghoshal, "Managing across borders: new strategic requirements," *Sloan Management Review,* vol.28 no.4, pp. 7-17, 1987

[4] Smith, H.A., R.T. Watson, and P. Sullivan, "Delivering Effective Enterprise Architecture at Chubb Insurance: A Case Study," *MISQ Executive,* vol.11 no2, 2012.

[5] Watson, R.T., et al., "U-commerce: expanding the universe of marketing," *Journal of the Academy of Marketing Science,* vol.30 no.4, pp. 333-347, 2002

[6] Junglas, I.A. and R.T. Watson, "The U-constructs: Four information drives," *Communications of AIS,* vol.17, pp. 569-592, 2006

[7] International Telecommunications Union, "The World in 2011: ICT Facts and Figures," Geneva, Switzerland, 2011

A MODEL FOR CREATING AND SUSTAINING INFORMATION SERVICES PLATFORM COMMUNITIES: LESSONS LEARNT FROM OPEN SOURCE SOFTWARE

Sulayman K. Sowe, Koji Zettsu, Yohei Murakami

National Institute of Information and Communications Technology,
3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0289, Japan
sowe@nict.go.jp, zettsu@nict.go.jp, yohei@nict.go.jp

ABSTRACT

Many research institutions are building cloud-based information services platforms (ISPs) that enable their researchers, scientists, and the general public use information assets, share knowledge and experience, and create sustainable communities. However, there is no guarantee that when you build an ISP this will happen. Part of the problem is because ISP providers lack the model to help them facilitate the building of sustainable communities. In this paper, we present a model for creating and sustaining communities on the ISP being developed by the National Institute of Information and Communications Technology (NICT) of Japan. Inspired by the way Open Source software communities operate, we describe the model concept, its settings, and the tools ISP communities may need to support their contribution towards the development of products and services. Our experience in the design and implementation of the model provides useful insights into emerging ICT trends and the means for ISP providers to identify, at an early stage, the requirements for creating successful products and services ecosystem.

Keywords— Information Services platform, Cloud computing, Information assets, Web Services, Prosumers, Open Source Communities, ICT

1. INTRODUCTION

A plethora of web platforms are available that offer service consumers and providers opportunities to search, filter, select, use, and carry out considerable transactions on the internet. Since its introduction in the late 90‘s, cloud computing is becoming increasingly popular [1, 2, 3, 4] as a *de facto* for deploying a host of services. This popularity has brought new trends in ICT.

First, businesses, governments, and research institutions are increasingly relying on cloud computing systems to deploy essential services. These service delivery platforms (SDP) accommodate anything ranging from telecommunication services, e-government, e-learning to e-business, e-agriculture, software and data, to mention a few. “Anything as a Service” or XaaS [3, 4] is appropriately used to describe this ecosystem of cloud services. SPDs leverage the Software, Platform and Infrastructure (SPI) [3] framework to create three service models [4], commonly referred to as infrastructure as a service or IaaS (e.g. Amazon.com‘s EC2), software as a service or SaaS, (e.g. salesforce.com), and platform as a service or PaaS (e.g. Google Apps). Information services platforms or ISPs that provide composite services to satisfy user requirements can utilise one or a combination of any of these models.

Big Data [5,20] represents the second ICT trend in the currently flourishing service oriented, high performance (HPC), and grid computing era. For example, [5] reported that we create 2.5 quintillion bytes of data every day and about 90% of that heterogeneous data has been created in the last two years (2010-2012) alone.

These trends are both exciting and challenging for the development and deployment of ISPs. The excitements stems from the fact that we can now, practically, manage and deploy essential services to more consumers. Technically, however, ISPs must overcome challenges associated with, but not limited to, standardization [38] infrastructure and resource allocation [4, 6, 7], security and risks [3], integration of heterogeneous data source [8, 9, 10, 19], efficient service delivery [11], sustainable business models [12], and, most importantly, how to create sustainable communities. With regards to sustainability, [2] noted that cloud computing systems have the potential to be more sustainable, when compared to traditional service centers.

However, to date, diminutive research, experience reports, best practices, frameworks, or models are available to help us understand how communities emerge and interact on ISPs. Maybe, the lack of literature is because, as [1] noted, most cloud platforms are proprietary and are built upon infrastructure that is invisible to the community. We posit that there is little consideration on the part of ISP providers to consider or integrate community aspects into their platform requirements. The exception we could find is the Eucalyptus cloud computing IaaS [1].

1.1. Contribution and research questions

Given the emerging ICT research trends relating to ISPs, our contribution is to address the gap between the design or infrastructure requirements of ISPs and community involvements.

We believe that this is important in that it will help future cloud-based infrastructure developers identify and provide means for building sustainable communities, at an early stage in their design process.

Furthermore, we believe that technology standardization bodies and focus groups such as the ITU-T Focus Group on Cloud Computing [39] should strongly integrate community sustainability elements.

However, these can only be achieved when we have better understanding of the community dynamics in various service areas. And, this is what we intend to achieve in this paper, by addressing two research questions, thus:

In the absence of ample literature and empirical data (e.g. case studies, surveys, experimentation) relating to community involvement in ISPs,

Q1. Are there lessons to be learnt and best practice to be adopted from Free and Open Source Software projects and communities that can help ISP providers create sustainable communities?

Once we have lessons to learn from and best practices to adopt to help create sustainable communities, then we can ask our second question,

Q2. How can ISP providers create a conducive environment for communities to collaborate and share their experience in the production and consumption of information assets?

To answer these questions, we present and describe the design and implementation of a model for creating and sustaining communities on the Japan National Institute of Information and Communications Technology (NICT) ISP. The model is inspired by our experience and understanding of the ways Free and Open Source Software (FOSS) projects and communities operate [10, 14, 15, 37]. We describe the tools and how service providers and consumers can use them to share their knowledge and experience, and contribute towards the co-evolution of products and services on the ISP.

The rest of the paper is organized as follows. In section 2 we present the background and work related to our research. Section 3 highlights the research settings of our model. In section 4, we describe the concepts that lead to the development of our model and our vision of a sustainable ISP community. Detail description of the model for creating and sustaining ISP communities is presented in section 5. Our concluding remarks and future research direction are summarized in section 6.

2. BACKGROUND AND RELATEDWORK

However versatile an ISP may be, the engagement and satisfaction of a community of service consumers and producers will immensely ensure its long-term sustainability. For example, [13] found out that satisfaction with the Google Android platform increased the level of consumer participation in that platform's community.

In an ISP community, service providers can interact with service consumers and discuss the design and enhancements of new or existing services. A community can act as advocates or service promoters, test the functionalities of services, act as opinion leaders, and provide insights into future service areas.

This kind of interaction is similar to the way self-learning and self-organizing FOSS developer and user communities operate [14, 15]. We conjecture that ISP providers can learn from FOSS projects and incorporate the community dynamics into the design of ISPs.

The general concept behind FOSS is making the human-readable source code of software accessible to anyone who wants to obtain it. Users can freely download, share, distribute copies to other people, customize and adapt the software to their local needs, and publish improved versions so that the whole FOSS community can benefit [37].

The FOSS development process [16] continues to produce a number of successful applications (e.g. Linux, Apache, Firefox, MySQL, and Symbian, to mention a few). Furthermore, most cloud computing infrastructure (e.g. Twitter, Facebook, and Amazon) is supported by FOSS technologies such NoSQl, Apache Hadoop and Cassandra.

Inherently, what makes the Open Source software paradigm more remarkable is not so much about the software itself, rather the way the communities in various projects operate [37]. Enable by the Internet, FOSS communities are typified by voluntarily (some paid) contributions to FOSS projects. Extensive peer collaboration allows project participants to write code, debug, test, integrate software, and help newbies [15]. The most typical characteristic of the FOSS development process is that developers are themselves users of the software, giving rise to the concept of *ugrammers* or users + programmers.

2.1. Collaboration and Coordination in FOSS

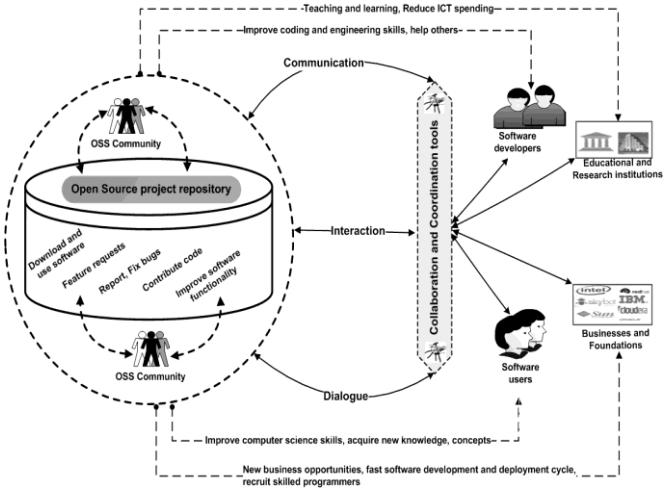


Fig. 1: Open Source Communities ecosystem.

The model we present in this paper is inspired by the dynamics of FOSS communities. As shown in figure 1, software developers, users, IT businesses, educational and R&D institutions use various tools (Versioning Systems or CVS/SVN, mailing lists, bug tracking systems, etc.) to

enable the software development and community building processes to proceed. Assisted by collaboration and coordination tools, the community gets involved in continuous dialogue, interaction, and communication. Participants can download, use communal products (software) and services, request new features, report and fix bugs, take part in discussions in various mailing lists, forums. In turn, the entire community benefits, as shown by the feedback loop (dotted lines in figure 1). Thus, it is not hard to see how the activities of FOSS developers and users are akin to that of service consumers and producers on an ISP.

3. RESEARCH SETTING

The model for creating and sustaining communities is an integral part of the ISP Lab cloud computing infrastructure (figure 2). The ISP lab is one of the six R&D labs belonging to the Universal Communication Research Institute (UCRI) of the Japan National Institute of Information and Communications Technology (NICT). UCRI promotes R&D, and conducts research in many fields including cloud computing and content service infrastructure [18].

The objective of the ISP lab is to build an information services platform that will support service consumers and providers in their use and provision of information assets, first for researchers and scientists locally, and then the global community.

The ISP consists of three technology platforms that offer mashups or composite web services [40] to users;

- *KGL Platform*: the Knowledge-Language Grid Testbed (KLG) [19] allows service consumers and data providers to publish, share and analyze Big Data [5] information assets for disaster information analysis and data-intensive science [8]. Big Data, according to [20], is "the aspiration to build platforms and tools to ingest, store and analyze data that can be voluminous, diverse, and possibly fast changing". The target users of this platform are NICT researchers and scientists, at first, and then the general public.
- *CPSenS Platform*: Cyber-Physical Sensing Information System (CPSenS) is a participatory sensing cloud platform for collecting and processing both physical and social sensor data. The target users of this platform are the general public.
- *WDS Platform*: the Cross-Database Search platform is a Data Citation Wiki for the World Data Systems or WDS (http://www.icsu-wds.org/). The platform serves as a search engine for discovering correlating between datasets from multi-domain, heterogeneous, very large science databases of the WDS [8]. For this platform, the ISP lab provides search and mining engines [21] on top of the participatory WDS repository. The target users of this platform are NICT researchers and climate and environmental scientists, at first, and then the general public.

The model described in this paper is meant to enhance the participatory nature of these platforms by, for example, supporting the use of information assets (right of Fig. 2) such as data, programs and devices (e.g WISDOM and Ikkyu), Web archives, and Tiled Displays.

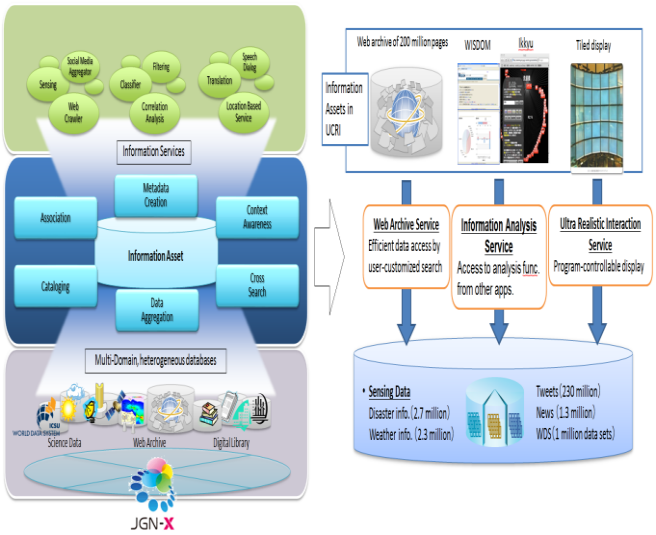


Fig. 2: Overview of ISP lab Cloud computing infrastructure.

3.1. Information assets and services on the NICT ISP

The NICT platform currently contains web archive of two billion pages. This is expected to double by 2015. The data is harvested using crawlers and text and natural language processing applications developed by the ISP lab.

The science data from WDS is an archive of environmental science data in the magnitude of 1.5 Exabyte, collected from over one hundred sites, spanning twelve countries.

The physical and social sensing data consists of aggregated data about the weather, natural disasters, traffic, census, news, and socio-economic indicators.

Service consumers can collect information from the ISP using crawlers and aggregators; analyze information using applications developed by the lab; and deliver information services using mobile Location Based Services (LBS), maps, speech dialogs, digital signage, and ultra-realistic displays.

4. THE ISP COMMUNITY MODEL CONCEPT

Before presenting and describing our model for creating and sustaining communities on the information services platform, we first introduce our vision of a sustainable community, and then conceptualize where on the NICT ISP we need to concentrate in order to maximize the chance of creating such a community.

Our vision of a sustainable information services platform community is one in which service or information assets

consumption is balanced by service production. This balance could be achieved when the community is intrinsically or extrinsically motivated to produce its own information assets from available sources through peer-production and information assets curation [40].

In this ecosystem, NICT (ISP provider) provides the technical infrastructure (servers and networks) to support the community platform, the tools and user guides or *informalisms* needed for the community to operate, and information assets (data and programs). As shown in the example use cases in figure 3, service consumers can then use, browse or search, review, annotate, contribute or import data and/or programs, and assetize user data. Service providers can perform the same function, but in addition to providing service on top of available data and programs, they can also maintain and provide support for the services they provide.

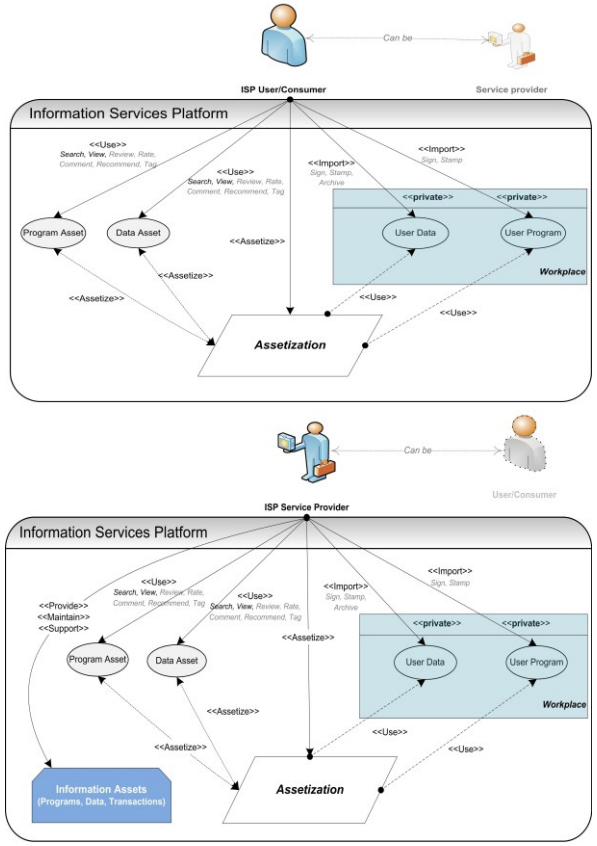


Figure 3: ISP Service Consumer and Provider Use Cases

These service consumers and providers will interact with system administrators to create a community of service curators.

The information assets curation process creates a closed-loop (shown in figure 4) in which the community use available assets, modify the assets to suit their own needs, and contribute their modified assets (with some added value) back to the ISP asset repository for the entire community to benefit. The cycle of assets use, modification, and contribution may create a sustainable service creation and consumption ecosystem on the platform.

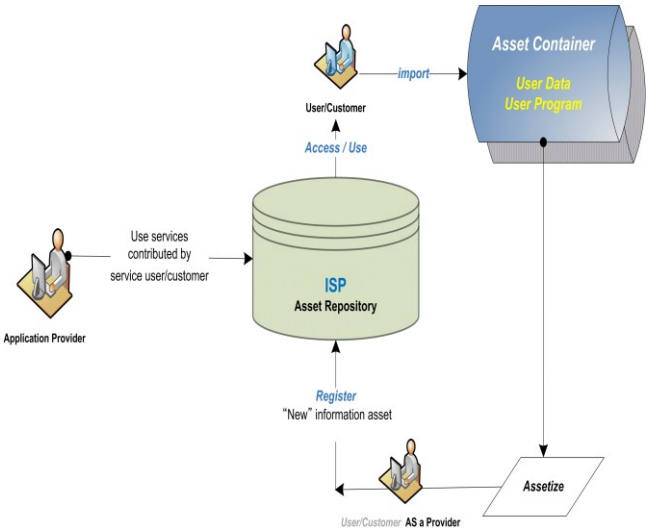


Figure 4: Service creation and consumption community loop

4.1. Conceptualizing community participation on ISPs

We conceptualized the NICT platform as consisting of four interrelated layers. As shown in figure 3, the community layer (4) is in constant engagement with layers 1 and 2. The community provides feedback to the engineers and architects in layer 1 so that they can improve the infrastructure and design of the platform. The community can also get involved in the information assets layer (2) by making suggestions and requesting features to improve products and services.

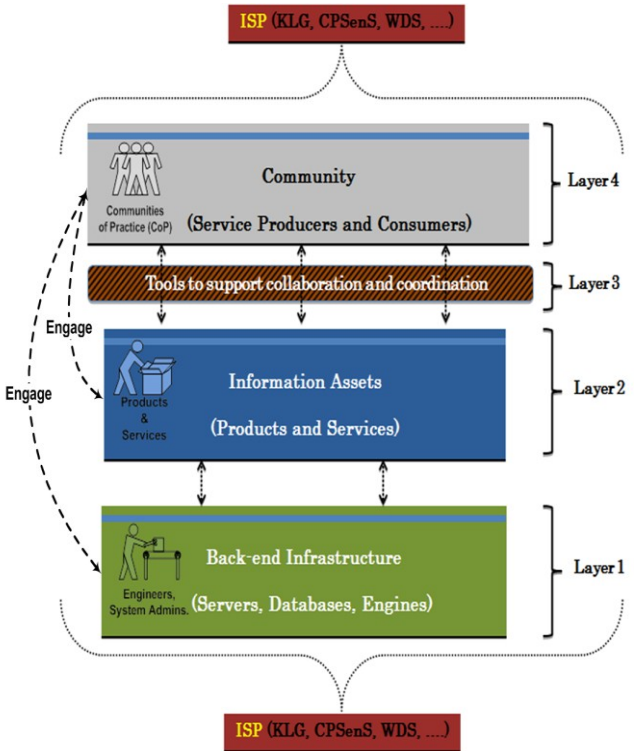


Fig. 5: High-level concept for creating sustainable communities on ISPs.

In order to support the ISP community in their service production and consumption activities, we envisage a tool layer (3) that will provide the medium for platform communities to collaborate and coordinate their activities. The platform community will be supported and mentored on the use of the tools by the *ISP community facilitator*. The community facilitator will further be responsible for moderating and guiding discussions and alerting community members when there are events and news about services.

Below is a detail description of what each layer entails:

1. **Layer 4**: The community layer consists of service consumers and producers or *prosumers* (producer + consumer) [25, 26, 27, 28] who will interact with available products and services in layer 2. Martin et al. [29], defined prosumers as "web users that also produce their own content". Prosumers may use, produce new services, and improve existing ones. Note that prosumers are synonymous to our *ugrammers* concept in Section 2.
2. **Layer 3**: Like FOSS communities [22, 14, 17], ISP communities need tools to help them engage in peer production of products and services. This layer consists of the tools communities will need in order to collaborate and coordinate their service production and consumption activities.
3. **Layers 1 and 2 (core components)**: The back-end infrastructure of the platform (Layer 1) consists of servers, database engines, and APIs. The information assets layer (2) consists of all the platform's information assets (data, meta-data, applications, and programs) that can be accessed by the community.

4.2. Tools to support collaboration and coordination

Many factors contribute to the success of FOSS communities in various projects. Factors include, but not limited to; trust [23], the quality of the software [products] or services on offer, usability, community management [5] and knowledge brokerage strategies [15]. The frequency of response to participants' queries has also been found to contribute to the success of FOSS communities [15].

Furthermore, lightweight tools to help members coordinate their activities are fundamental for the long-term success of FOSS communities. According to [24], collaboration tools enable groups of service consumers and providers to work as a team, sharing information and communicating as needed, without being co-located. In our model, the tools we envisaged will support this kind of participation and collaboration are the following:

- *Forums - discussion*: Forums are important in archiving all discussions relating to the platform's products and services. The platform community can see an entire history and evolution of both the community and technological artifact. Forums have the benefits of a pull-factor instead of being a push-factor. That is, service consumers and providers have to visit and take part in discussion in a forum of their choice, instead of discussions or notifications being sent or pushed to them by email. For the NICT platform, we have three forums.
    I. Forum for the announcements of new products and services, arrival of new members, platform news and events;
    II. Forum for general discussion about the platform; anything from products and services, management, governance, to technical issues;
    III. A developers forum where people can discuss about the development of products and services, coding and testing, security and quality assurance (QA), release milestones, blueprints, etc.
- *Mailing Lists - communication*: The platform community also has an option to subscribe to three mailing lists.
    I. General discussions mailing list dealing with anything from products and services, management, governance, to technical issues;
    II. Products mailing list for all communications relating to products;
    III. Services mailing lists for all communications relating to services.
- *Wikis*: The platform will maintain a wiki for documentation and general information about information assets (name, description, contributors, partners, screenshots, demos, FAQs).

5. MODEL DESCRIPTION

The model for creating and sustaining communities on the NICT ISP is schematically shown in figure 6. Three stakeholders (service consumer, service provider, system administrator) converge on the platform. The stakeholders have complementary roles. For instance, service consumers, who may be NICT researchers or scientists, can also become service producers. NICT researcher, as service provider, may also develop text mining and visualization software (shown in figure 6 as "asset") for his own use or as NICT research requirement, and host the software on the platform as software as a service (SaaS) [30].

Once software, data, or applications are uploaded to the platform by the system administrator, other NICT researchers or scientists, as service consumers, can download and use the asset. This cycle continues for all information assets hosted on the NICT platform, creating an ecosystem of service creation and service usage.

An important parameter we have added to the role of a service provider is a service maintainer. A term borrowed from the Open Source Debian project [31], a package maintainer is someone who maintains one or more Debian software packages and is privileged to upload their packages to the Debian repository.

In our model a service maintainer is a service provider who coordinates the development, maintenance, and usage of

that service on the platform. He may write documentation explaining how the service is structured, response to queries from the platform community about that service, and announce any enhancements or improvements made on the service.

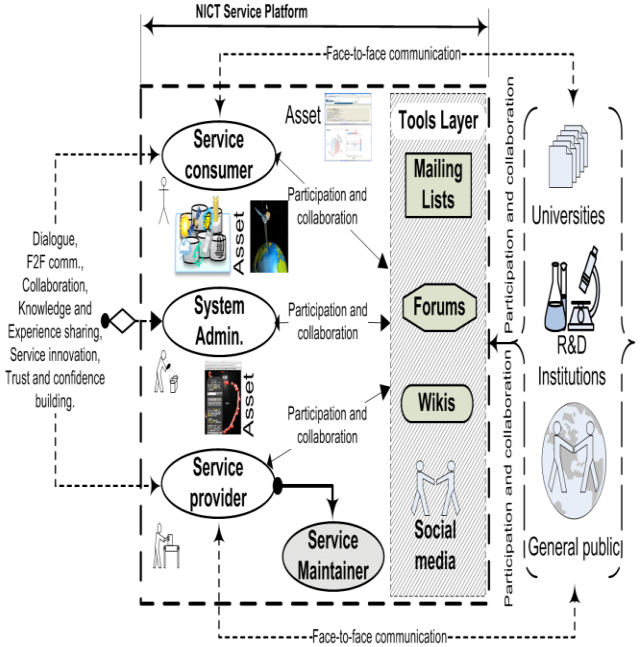


Fig. 6: Model for creating and sustaining ISP communities.

5.1. Community Participation

In order to address our first research question, the model depicts the ISP as a successful FOSS project [22, 17]. As shown in figure 6, tools layer is introduced in the NICT ISP. The community can use the tools to enhance their participation and collaboration. The tools support dialogue, trust and confidence building in the community. Each stakeholder can use the tools to collaborate and exchange knowledge and experience about the information assets. The community can also be involved in service innovation [32] by using the tools (e.g. forums) to discuss improvements and development status of products and services.

An important parameter in the model is the recognition of a face-to-face or F2F communication channel that bypasses the tools layer. In a research institution such as NICT, seminars, workshops and Agile and sprint [33] software development practices are very common. Our model accommodates these idiosyncrasies by providing the paths for the platform community to be involved in off-line, F2F communication. We also expect the model to support participation and collaboration with a multitude of other stakeholders, including the general public, R&D institutions, and Universities.

6. CONCLUSIONS

In this paper, we presented a model for creating and sustaining a community of service consumers and providers on the NICT cloud-based ISP. We began by conceptualizing where on the platform should an ISP community facilitator give due consideration in order to maximize the potentials for building a sustainable community [34]. We went on to describe the tools need to support collaboration and coordination of community activities on the ISP. Inspired by our FOSS experience, we described how the ISP community should operate and what is needed to ensure its sustainability. Thus, ISP providers could learn and adopt best practices from FOSS communities and projects to help them create sustainable communities (Q1).

Giving the feedback we received from the NICT research and scientific community, we are optimistic that this model could be valuable for creating and supporting the environment conducive for communities to collaborate and share their knowledge and experience in the production and consumption of information assets (Q2).

Furthermore, this model is a response to NICT`s policy of integrating, among others, accessibility and socio-cultural aspects into all of the institute`s technologies. Apart from fulfilling this policy objective, this paper also highlighted elements we may consider when developing standards for sustaining communities on ISPs.

However, it may be too early to be over optimistic until our ISP community reaches a critical mass, and the model parameters evaluated.

Albeit, this research may serve as a good starting point for developing standards, as well as getting ICT services providers to start thinking about how to create sustainable communities on the cloud computing information services platforms.

6.1. Future Work

For our future work, we plan to improve community accessibility to the tools layer by creating a structure, similar to ConnectedSpaces described by [24].

Furthermore, this model was presented and discussed with NICT ISP researchers and scientists. Their interest was high and encouraging. Our next plan is to pilot the model on one component of the ISP such as the Knowledge-Language Grid Testbed [19]. Participant observation research method [35] will be used to observe, record, and analyze stakeholders' involvement. A *pre-* and *post-* questionnaire will be administered, and the Technology Acceptance Model [36] will be used to evaluate the perceived usefulness, perceived ease of use, and stakeholders' perception of the model.

REFERENCES

[1] Daniel Nurmi, et al., "The eucalyptus open-source cloud-computing system," in *Proceedings of the 9th IEEE/ACM*

International Symposium on Cluster Computing and the Grid, Washington, USA, pp. 124–131, 2009.

[2] M. Arlitt, et al., "Cloud sustainability dashboard," in *Sustainable Systems and Technology (ISSST), IEEE International Symposium*, p. 1-1, 2010.

[3] Kamal Dahbur, Bassil Mohammad, and Ahmad Bisher Tarakji, "A survey of risks, threats and vulnerabilities in cloud computing," in *Proceedings of the International Conference on Intelligent Semantic Web-Services and Applications*, New York, USA, pp. 12:1–12:6, 2011.

[4] I. Voras, et al., "Evaluating open-source cloud computing solutions," in *MIPRO2011 Proceedings of the 34th International Convention*, pp. 209 –214, 2011.

[5] Vinayak Borkar, Michael J. Carey, and Chen Li, "Inside "big data management": ogres, onions, or parfaits?," in *Proceedings of the 15th International Conference on Extending Database Technology*, New York, pp. 3–14, 20121.

[6] Dan Ionescu, "On-the-cloud computing: Challenges in controlling cloud resources," in, *International Conference on Collaboration Technologies and Systems*, p. 299-300, 2012.

[7] Yi Wei and M.B. Blake, "Service-oriented computing and cloud computing: Challenges and opportunities," *Internet Computing, IEEE*, vol. 14, no. 6, pp. 72 –75, 2010.

[8] Masahiro Tanaka, Yohei Murakami, Koji Zettsu, "Data-intensive Services for Large-scale Archive Access", *Proceedings of the IEEE 9th International Conference on Service Computing*, pp.617-624, 2012

[9] Yohei Murakami, Masahiro Tanaka, Arif Bramantoro, Koji Zettsu, "Data-Centered Service Composition for Information Analysis", *Proceedings of the IEEE 9th International Conference on Service Computing*, pp.602-608, 2012

[10] Sulayman K. Sowe and Antonio Cerone, "Integrating Data from Multiple Repositories to Analyze Patterns of Contribution in FOSS Projects," *Journal Electronic Communications of the EASST*, vol. 33, pp. 442–460, 2010.

[11] Alistair Barros and Uwe Kylau, "Service delivery framework an architectural strategy for next-generation service delivery in business network," in *Proceedings of the Annual SRII Global Conference*, Washington, pp. 47–58, 2011.

[12] Sanjeev Sharma, "BMP Patterns & Practices in Industry," Tech. Rep., ORACLE White Paper, May, 2012.

[13] Luis V. Casaló, Carlos Flavián, Miguel Guinalíu, "Relationship quality, community promotion and brand loyalty in virtual communities: Evidence from free software communities," *Int. J. Inf. Manag.*, vol. 30, no. 4, pp. 357–367, Aug. 2010.

[14] Sulayman K. Sowe, Ioannis Stamelos, and Lefteris Angelis, "Understanding Knowledge Sharing Activities in Free/Open Source Software Projects: An Empirical Study," *Journal of Systems and Software*, vol. 81, no. 3, pp. 431–446, 2008.

[15] Sulayman Sowe, Ioannis Stamelos, and Lefteris Angelis, "Identifying Knowledge Brokers that Yield Software Engineering Knowledge in OSS Projects," *Information and Software Technology*, vol. 48, no. 11, pp. 1025 – 1033, 2006.

[16] Eric S. Raymond, *The Cathedral and the Bazaar*, O'Reilly & Associates, Inc., Sebastopol, CA, USA, 1st edition, 1999.

[17] Kevin Crowston and James Howison, "Assessing the health of open source communities," *Computer*, vol. 39, no. 5, pp. 89–91, 2006.

[18] "Overview of the NICT Universal Communication Research Institute," retrieved Sept. 23, from: http://www.nict.go.jp/en/univ-com/index.html

[19] Y. Murakami, et al., "Service grid federation architecture for heterogeneous domains," in 9th *International Conference on Services Computing (SCC)*, pp. 539 –546, 2012.

[20] Surajit Chaudhuri, "How different is big data?," in *Proceedings of the IEEE 28th International Conference on Data Engineering*, Washington, USA, pp. 5–5, 2012.

[21] Eloy Gonzales and Koji Zettsu, "Association rule mining from large and heterogeneous databases with uncertain data using genetic network programming," in *The 4th International Conference on Advances in Databases, Knowledge, and Data Applications*, pp. 74–80,2012.

[22] Karl Fogel, *How To Run A Successful Free Software Project Producing Open Source Software*, CreateSpace, Paramount, CA, 2009.

[23] Paul B. Laat, "How can contributors to open-source communities be trusted? on the assumption, inference, and substitution of trust," *Ethics and Inf. Technol.*, vol. 12, no. 4, pp. 327–341, 2010.

[24] J. Buford, K. Mahajan, and V. Krishnaswamy, "Federated enterprise and cloud-based collaboration services," in *IEEE 5th International Conference on Internet Multimedia Systems Architecture and Application*, pp. 1–6, 2011.

[25] Rute Sofia, et al., "Moving towards a socially-driven internet architectural design," *SIGCOMM Comput. Commun. Rev.*, vol. 42, no. 3, pp. 39–46, 2012.

[26] A. J. Dinusha, et al., "Identifying prosumer's energy sharing behaviours for forming optimal prosumer-communities," in *Proceedings of the 2011 International Conference on Cloud and Service Computing*, Washington, pp. 199–206, 2011.

[27] Wolfgang G. Stock, "Folksonomies and science communication: A mash-up of professional science databases and web 2.0 services," *Inf. Serv. Use*, vol. 27, no. 3, pp. 97–103, 2007.

[28] Kouji Ohboshi, "Mobile, broadband, ubiquitous, and the information renaissance," in *Proceedings of the 15th international symposium on System Synthesis*, pp. 1–1, 2002.

[29] Yod Samuel, et al., "Prosumers and accessibility: how to ensure a productive interaction," in *Proceedings of the 2009 International Cross-Disciplinary Conference on Web Accessibililty (W4A)*, New York, pp. 50–53, 2009.

[30] Bikram Sengupta and Abhik Roychoudhury, "Engineering multi-tenant software-as-a-service systems," in *Proceedings of the 3rd International Workshop on Principles of Engineering Service-Oriented Systems*, New York, pp. 15–21, 2011.

[31] Debian Maintainer, retrieved on Sept. 23, 2012, from: http://wiki.debian.org/DebianMaintainer .

[32] M. Yuriyama, T. Kushida, and M. Itakura, "A new model of accelerating service innovation with sensor-cloud infrastructure," in *SRII Global Conference (SRII),* pp. 308 – 314, 2011.

[33] G. Goth, "Sprinting toward open source development," *Software, IEEE*, vol. 24, no. 1, pp. 88 –91, 2007.

[34] Amy S. Ward, "How to build a sustainable community," October 2010. Retrieved on Sept. 23, 2012, from: http://www.socialbrite.org/2010/05/10/how-to-build-a-sustainable-community/

[35] Jack D. Douglas, *Investigative Social Research*, Beverly Hills, CA: Sage Publications, 1976.

[36] Fred D. Davis, "Perceived usefulness, perceived ease of use, and user acceptance of information technology," *MIS Quarterly*, vol. 13, no. 3, pp. 319–340, Sept. 1989.

[37] Sulayman K. Sowe, Ioannis G. Stamelos, Ioannis M. Samoladas (Eds.), *Emerging Free and Open Source Software Practices*, IGI Global, Covent Garden, London, 2008.

[38] Borenstein, N.; Blake, J., "Cloud Computing Standards: Where's the Beef?," *Internet Computing, IEEE* , vol.15, no.3, pp.74-78, May-June 2011

[39] ITU-T Focus Group on Cloud Computing G CloudRetrieved on Nov. 21, 2012, from http://www.itu.int/en/ITU-T/focusgroups/cloud/Pages/tor.aspx

[40] Kaisa Mattila and Minna Waljas. Towards user-centered mashups: exploring user needs for composite web services. In *CHI '11 Extended Abstracts on Human Factors in Computing Systems* (CHI EA '11), pp. 1327-1332, 2011.

SECURITY TECHNOLOGIES FOR THE PROTECTION OF CRITICAL INFRASTRUCTURES – ETHICAL RISKS AND SOLUTIONS OFFERED BY STANDARDIZATION

Simone Wurster

Chair of Innovation Economics
Berlin University of Technology, Berlin - Germany
simone.wurster@tu-berlin.de

ABSTRACT

The added value of standards is shown in numerous research articles. Several recent studies also highlight the need for security standards. Security products and services may bear ethical and privacy-related risks which can impede acceptance of new security solutions. Specific privacy standards may help to overcome such problems, but privacy issues of security technologies are not covered by standardization research so far.
This paper deals with the topic from mainly German and European perspectives. Based on a survey in the German security research program, it gives an overview of security technologies, the specific risks they bear and their importance. Three technology-related categories were identified: surveillance solutions for detection from distance, solutions for obtrusive detection and data processing. Relevant risks were described and discussed. Solutions based on standardization were shown. The paper finishes by giving recommendations for new privacy standards.

Keywords— *P*rivacy, standards, public security, critical infrastructures, surveillance technologies, data processing

1. INTRODUCTION

The global intensity and frequency of criminal and terrorist attacks since the turn of the century has shown the vulnerability of democratic societies and the need for protecting so-called critical infrastructures in particular (see [15]). To address these threats, numerous countries like Canada, the United States as well as the European Union and many European member States like Austria, France, Germany and the United Kingdom established security research programs. Several recent studies also highlight the need for security-related standards, e.g. [8] and [10].

Critical infrastructures comprise all physical and information technology facilities, networks, services and assets which, if disrupted or destroyed, would have a serious impact on the health, safety, security or economic well-being of citizens or the effective functioning of the government in a country (see[9]). They include energy installations and networks; communications and information technology; water (dams, storage, treatment and networks); transport (airports, ports, intermodal facilities, railway and mass transit networks and traffic control systems) and the government (e.g. critical services, facilities, information networks, assets and key national sites and monuments) (see [9]).

[17] defines security as 'a system of measures, including their embodiments and their interactions, designed to ward off intentionally destructive activity resulting in injury or material damage'.

The European Commission distinguishes between two kinds of security: a) security of the society (public security) and b) ICT (information and communication technology) security. Defense and space issues are not covered by the European public security concept. Security of the society (category a) includes four dimensions: security of the citizens, security of infrastructures and utilities, border security as well as restoring security and safety in case of crisis. With the exception of cryptography, which is considered a key technology for any security application, information and communication technologies are not covered by this category, which is why a separate category exists (see [10]). An overview of European and international ICT security standards is, for example, given by the ITU-T Study Group 17 – Security (see [14]).

The European Commission's perception of security relates to public security and "includes among others, protection against threats by terrorism, severe and organised crime, natural disasters, pandemics and major technical accidents" [10]. Following the European Commission's emphasis on public security standards, the focus of this paper is on standards in this field.

Privacy is a specific issue in the security field. Definitions are, for example, given by Bok (1982) and Breckenridge (1980).

"(Privacy is) the condition of being protected from unwanted access by others – either physical access, personal information, or attention" (Bok, 1982, quoted by [19]).

"Privacy, in my view, is the rightful claim of the individual to determine the extent to which he wishes to share of himself with others and his control over time, place, and circumstances to communicate to others. It means his right to withdraw or to participate as he sees fit. It is also the individual's right to control dissemination of information about

himself; it is his own personal possession" (Breckenridge, 1980, quoted by [19]).

information. Principle privacy-related rights are defined by the Universal Declaration of Human Rights (UDHR) and in Europe for example by the EU Convention for the Protection of Human Rights and Fundamental Freedoms (ECHR):

- Article 12 UDHR: No one shall be subjected to arbitrary interference with his privacy, family, home, or correspondence, nor to attacks upon his honour and reputation.
- Article 8 ECHR: Everyone has the right to respect for his private and family life, his home and his correspondence. There shall be no interference by a public authority with the exercise of this right except for well-defined circumstances such as national security.

A lot of technologies, products and systems to protect critical infrastructures are currently in development or already available. Privacy goals and security-related goals may contradict each other. Regarding critical infrastructures security has specific importance and fulfilling both goals bears specific challenges. Specific standards may offer solutions.

Standardization is 'the activity of establishing and recording a limited set of solutions to actual or potential matching problems directed at benefits for the party or parties involved balancing their needs and intending and expecting that these solutions will be repeatedly or continuously used during a certain period by a substantial number of the parties for whom they are meant' ([6], p. 13).

[2], [5] and [18] give overviews of the many advantages standardization provides. General advantages include, for example, its contribution to global market access for innovative solutions, economies of scale, cost savings as well as the facilitation of compatibility and interoperability. Standardization also raises the acceptance of innovations among customers and public procurers and facilitates the licensing of patents by referencing them into standards (see [5]). Advantages for enterprises are also based on its shaping of the framework conditions of new and emerging markets and the access of new technologies to the market while research organizations may profit from a facilitated transfer of technology into marketable products and services, of the dissemination of research results and of enhanced recognition and reputation (see [5]).

A specific area in which standardization can raise the acceptance of innovations and enhance reputation in the security field is privacy. Therefore, enterprises and research organizations which develop security technologies may profit from the establishment of appropriate privacy standards.

Although some researchers are closely involved in standardization processes, the vast majority of scientists (not only in the security field) seldom regard standardization as a high priority. As a result, many researchers do not use the special opportunities that standardization can offer them (see [4]).

According to both definitions, privacy has two dimensions: physical freedom of a person and having control over personal

The project InfraNorm addresses specific problems in the transfer of security research results to market implementation and offers solutions based on standardization. It is a joint project between the DIN German Institute for Standardization and the Berlin University of Technology and is funded by the German Federal Ministry of Education and Research. Its goal is to initiate the development of standards for the protection of transportation infrastructure. InfraNorm collaborates with ten associated project consortia, initiated to improve the protection of critical infrastructures such as airports, train stations and ports as well as the protection of railways, bridges and tunnels.

2. LITERATURE REVIEW AND RESEARCH GAP

Investigating security-related standards requires in-depth insight into standard-related and security issues.

[3] provides important innovation economic findings regarding the importance of security standards and related needs. Technological security solutions very often combine soft- and hardware. New security systems must be integrated into existing security infrastructures and require interoperability. Interface standards are particularly critical in the introduction stage of new technology. Furthermore, the implementation of new security systems requires acceptance by the market and the users, which can be facilitated by standards. They also promote the transition from old to new technologies. In the context of security solutions with multiple components interface standards particularly support a variety of offerings (see [3]).

Privacy issues of security technologies are for example investigated by [1], [11], [16], [20] and [21]. The authors show that privacy problems can impede acceptance of new security solutions. According to [1] many security solutions bear ethical risks:

"(In democratic societies, some) surveillance activities are necessary or desirable in principle - for example, to fight terrorism and serious crime, to improve entitlement and access to public services, and to improve healthcare. But unseen, uncontrolled or excessive surveillance activities also pose risks that go much further than just affecting privacy. They can foster a climate of suspicion and undermine trust" ([1], quoting the 28th International Conference on Data Protection and Privacy Commissioners).

Measures to address concerns about critical infrastructure and the physical safety of the population in particular usually have a substantial impact on privacy (see [13]). Specific privacy standards may help to overcome such problems, but there has been no scientific work which investigates standardization related to privacy issues of security technologies so far.

Three kinds of civil security-specific settings can be distinguished:

- private places (e.g. private owned houses or company buildings)

- public places (e.g. public parks, schools etc.) and
- semi-public places like airports, train stations, ports etc. (see [20]).

Many semi-public areas represent critical infrastructures and are used by millions of people every day worldwide. Therefore the security of these areas has great importance. Besides several guidelines on how to ensure public security, specific privacy-related recommendations for the protection of semi-public areas which represent critical infrastructures are missing.

The investigation of ethical and privacy-specific issues needs to consider all groups of security technology and security solutions, for example closed-circuit television (CCTV) and radio-frequency identification (RFID) technologies separately: "It is crucial to clearly distinguish different types of detection technologies (i.e. CCTV, RFID tags, biometrics, etc.) in order to match appropriate data protection solutions to each of them separately" ([1]:4).

3. SURVEY IN THE GERMAN SECURITY RESEARCH PROGRAM

In summer 2011 a survey about security research and standardization was done among the participants of the German framework program "Research for civil security" (see [12]). In order to gain a deeper insight into ethical and privacy-related problems of security technologies and to identify possible solutions, a follow-up study with 23 participants of the German program was done. The participants consisted of 6 people from supplier companies of security-related products and services, 5 from research organizations, 8 from universities, 1 person from an industry association and 3 people representing the end user. Six of ten questions were related to ethical and privacy-specific risks of security technologies:

1. What security-related technologies, products or services bear special ethical or privacy-specific risks in your opinion?
2. Please use up to five of the described technologies, products or services to rank their risk potential.
3. Please name ethical and privacy risks of the top-ranked technologies, products or services.
4. What other ethical and privacy-specific risks are important with regard to other security-related technology, products or services from your point of view?
5. In what way is there a need for standards for better addressing ethical and privacy specific aspects in the development and use of security related products and services from your point of view?
6. Which technologies, products and solutions have specific standardization needs to reduce ethical and privacy risks? Please describe the need in note form.

The completion the questionnaires took place between June and July 2012. The results are presented in the next chapter.

4. ETHICAL AND PRIVACY-SPECIFIC RISKS OF SECURITY SOLUTIONS

Survey data was coded and clustered with the software Atlas.TI. According to Figure 1, related to question 1 four areas with ethical and privacy-specific problems were identified: detection technologies, processing of data, security services and additional topics which only received two mentions. The first two areas address specifically security products and technology-related fields. The figure also shows that the ethical and privacy-specific risks of detection technologies are regarded as most important. Figure 2 describes their nature in more detail.

Besides general topics included in the data processing category, specific emphasis is given to medical data. 'Security services' relates to the services provided by specialized security firms. The category 'additional topics' includes for example the use of social media in criminal investigations which are both not in the focus of this technology-related study. It also includes ethical issues related to emergency call centers. Unfortunately, no specific description of the nature of the relevant ethical risk is given.

According to Figure 2, two specific areas of detection technologies were identified: detection from distance and obtrusive detection. The first cluster includes, for example, body scanners, biometric devices and access control. Technologies which allow detection from distance comprise all kinds of video surveillance solutions including intelligent video surveillance as well as identification technologies, for example to identify license plate numbers.

Based on the next question, the participants were asked to form a ranking of the products and technologies mentioned in the previous question according to their ethics and privacy-specific risk potentials. Figure 3 shows the relation between the number of entries according to question 2 and a weighted score according to the specified order of precedence used by the participants1. Similar to the results of question 1, detection with distance is regarded as bearing the most ethical and privacy-specific risks.

Question 3 and 4 addressed specific ethical and privacy risks. Related to the identified technologies, three groups of problems became apparent: restrictions to freedom, abuse and discrimination. Table 1 provides an overview of the identified ethical risks of the first-ranked technologies, products and services.

[1] Scoring according to the specified rank: rank 1 = 1 - rank 5 = 1/5.

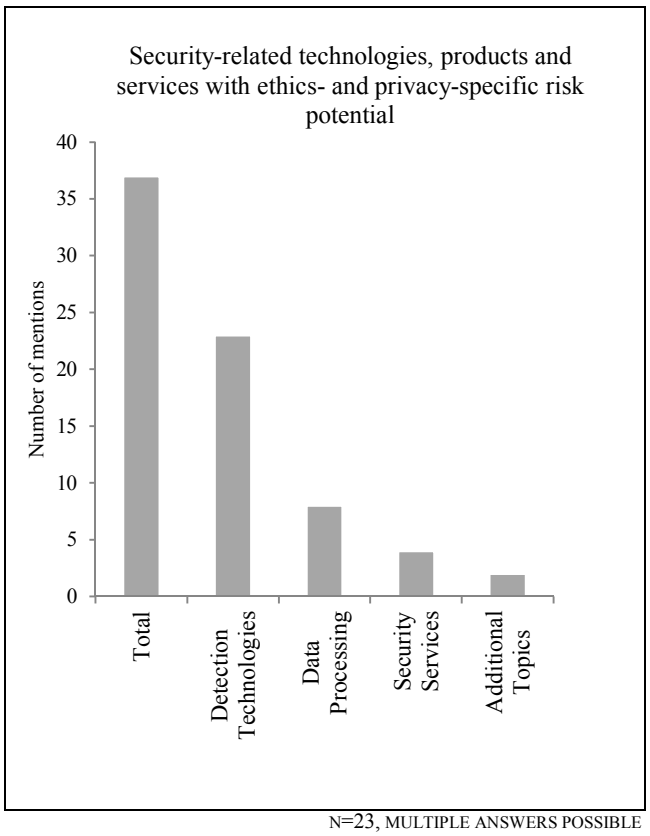


Figure 1. Security solutions with ethical risks

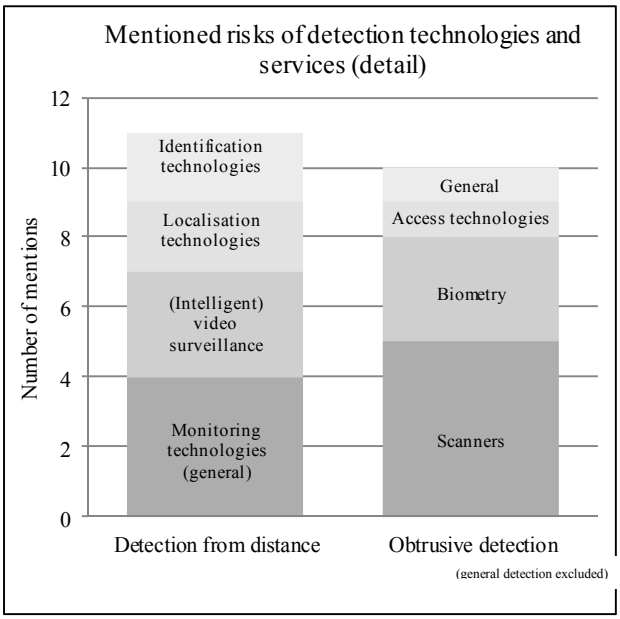


Figure 2. Importance of ethical risks of different detection technologies

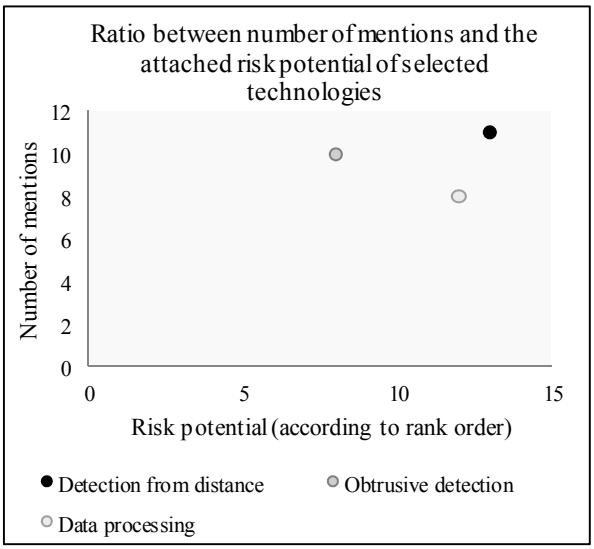


Figure 3. Analysis of the risk potential of the identified technologies

Table 1. Overview of potential ethical risks

Restricted freedom	
- Interference with privacy - Identification of individuals - Data mining / data analysis, profiling	- People tracking - Lack of confidentiality - Lack of legitimacy - No consent - Lack of proportionality
Abuse	Discrimination
- Abuse in general - Voyeurism	- D. in general - Motion-based profiling - Data mining / data analysis, profiling

A few technologies are associated with several risks. Surveillance, for example, bears all three categories of risks. Surveillance data could be used to create personal profiles and thus infringe the privacy rights of those affected. Depending on the type of data, it can also lead to discrimination.

Restricted freedom

The risk to reduce freedom by pursuing security objectives is related to both, detection from distance and obtrusive detection as well. Particularly surveillance technologies could be used for the identification of persons and the creation of profiles by aggregating data or their interfacing with other information without the consent of the observed or particular suspicion.

Abuse

Regarding the misuse of data, the risk of an unauthorized disclosure of confidential data is regarded as having specific importance. Therefore, appropriate data protection needs to be guaranteed. In addition, a lack of rules for data access can bear the danger of an unauthorized use of the collected data. Misuse of data is often accompanied by a violation of personal rights and is sometimes based on a lack of confi-

dentiality. Besides data breach in general, voyeurism is also part of this category.

Discrimination

Most of all, the risk of discrimination was related to the creation of profiles.

Additional aspects mentioned by the participants are a lack of controls and transparency, function creep and misidentification.

Question 5 and 6 focused on the need for standards for better addressing ethical and privacy specific aspects in the development and use of security related products and services. Most responses on question 5 included general recommendations independent of specific technical areas. They include

- The integration of privacy topics into standards in general
- Certifications for ethic-friendly security products and
- Better information of the public regarding security measures and the use of detection technologies and formulation of their rights.

Two additional suggestions were related to test criteria and standards for the use of medical data and ethical aspects in the field of protection and rescuing.

Instead of giving a specific answer, several participants submitted general comments including statements that describe a strong need for ethical standards for security technologies as well as hints that aspects exist which cannot be standardized.

Question 6 addressed specific technologies, products and solutions and related standardization needs to reduce ethical risks. Based on the answers, six technology fields were identified: Security services, Data storage, Video surveillance, Biometrics, Access control and Sensors.

Suggestions concerning security services include for example, quality standards for training und qualification. Recommendations regarding data storage require the implementation of neutral supervision, a specification of the storage period as well as a specification of the kind of data. Answers addressing video surveillance are quite similar and refer mainly to the storage period. Suggestions in the biometric field were related to the matching of biometric records. A domain-specific legal basis which prescribes the conditions and limitations of the procedure is regarded as necessary. Sensor-specific suggestions refer to an ethic standard for this field in general. In the context of access control a wish to make use of verification procedures instead of identification techniques was expressed.

5. MEASURES TO REDUCE ETHICAL AND PRIVACY-SPECIFIC RISKS BY STANDARDIZATION

Database and document analyses were done to compare the needs with existing standards and other relevant documents. As mentioned at the beginning, international and regional, for example European, regulations exist. An important European document is the Directive 95/46/EC on the protection of individuals with regard to the processing of personal data and on the free movement of such data. Current activities in Europe include in particular common work to create a new General Data Protection Regulation which will displace Directive 95/46/EC. The draft was released in January 2012. Its finalization is planned for the end of 2013 while its implementation is expected to be finished in 2015 or later. The regulation does not include specific technology-related passages. More technology-specific documents regarding privacy are, for example, provided by the European standardization organization CEN though most documents are CEN Workshop Agreements (CWA)2. The documents include, for example3:

- CWA 16113 Personal Data Protection Good Practices
- CWA 15499 Personal Data Protection Audit Framework (EU Directive EC 95/46): Part I and II
- CWA 15292 Standard form contract to assist compliance with obligations imposed by article 17 of the Data Protection Directive 95/46/EC

as well as CWA 15262 which describes data protection auditing practices and CWA 15263 which shows the need for Privacy-Enhancing Technologies (PET).

Additional technical specifications (TS) and reports (TR) in the privacy field include three standards from the European standardization organization ETSI: *ETSI TS 102 656, ETSI TR 102 661 and ETSI TS 102 657*. The documents address the field of telecommunication. Due to specific privacy needs and related legal conditions, use of these standards is not compatible with the current privacy-specific requirements of all European member states.

An important role in the current development of privacy standards and specifications is played by the ISO/IEC Joint Technical Committee (JTC) 1/Sub Committee (SC) 27/Work Group (WG) 5 Identity Management and Privacy Technologies. It cooperates with many organizations and committees including the ITU. Currently the work group is developing several standards in the areas of privacy impact assessment and privacy information management systems. Further projects are being carried out in the area of privacy architecture. They relate, for example, to privacy policies and privacy enhancing technologies (PET) and services. Projects by the work group include *ISO/IEC 29101 Information technology -- Security techniques -- Privacy architecture framework* which is currently an ISO/IEC Committee Draft (CD), *ISO/IEC 29115 Information technology -- Security techniques -- Entity authentication assurance framework* which is currently an ISO/IEC Draft Interna-

[2] The adoption of CWAs is voluntary in the European member states.

[3] An extended overview of relevant standards and directives is given by the InfraNorm standardization manual (forthcoming).

tional Standard (DIS) as well as *ISO/IEC 29100 Information technology -- Security techniques -- Privacy framework* which is already available and includes eleven privacy principles (see Table 2).

ISO/IEC DIS 29115 is a joint project of ISO/IEC and ITU-T and is also called ITU-T X.1254.

CWA 16113 Personal Data Protection Good Practices and *ISO/IEC 15944-8 Information technology -- Business Operational View -- Part 8: Identification of privacy protection requirements as external constraints on business transactions* pay specific attention to privacy in the context of public security, for example to the privacy principle "Individual participation and access (to data of the individual)":

"The Directive [95/46/EG] set out a small number of circumstances in which their right to see personal records can be limited. This is necessary in order to strike a balance between the rights of the individual and some important needs of civil society, on the other hand." (CWA 16113)

Therefore, the CWA includes "the processing is necessary in order to fulfill a task in the public interest" in its list of legitimate reasons for processing personal data. ISO/IEC 15944-8 describes exceptions as follows:

"Privacy protection requirements of jurisdictional domains may contain exceptions (…). The most common exceptions are those relating to national sovereignty and security, law enforcement, public safety and health. Exceptions of this nature often require access to personal information about a particular individual and the tracing of any other personal information pertaining to that individual (…)".

Therefore, ISO/IEC 15944-8 includes a specific rule:

"Where exceptions to the application of privacy protection principle exist, they shall be: 1) limited and proportional to meeting the objectives to which these exceptions relate, and 2) a) made known to the public; or b) in accordance with law."

Table 2. Privacy principles of ISO/IEC 29100

1. Consent and choice	6. Accuracy and Quality
2. Purpose legitimacy & specification	7. Openness, transparency and notice
3. Collection limitation	8. Individual participation and access
4. Data minimization	9. Accountability
5. Use, retention and disclose limitation	10. Information security
	11. Privacy compliance

Specific guidelines for the permanent protection of semi-public areas which represent critical infrastructures (e.g. ports and airports) are not covered by this standard and managers of these infrastructures express need for action.

Additionally privacy-specific activities in specific technology fields exist, for example, related to RFIDs. ISO/IEC JTC1/SC31 develops a global security system for data that is transmitted by RFID tags and the report *ETSI TR 187 020* includes a gap analysis of additional privacy issues in the context of RFID. The report shows also 34 projects to be completed by the end of 2013.

Regarding **biometrics** there are five documents which have specific importance: *ISO/IEC 19784-2, ISO/IEC 19785-1, ISO/IEC 19792, ISO/IEC 24745 and ISO/IEC TR 24714-1. ISO/IEC TR 24714-1*, for example, defines 14 privacy guidelines. Although the guidelines are state-of-the art, it is a technical report only. An additional problem is that no specific applications for the protection of critical infrastructures, public and semi-public areas and the specific use of data in these contexts exist.

Documents related to **data storage** include the European Standard *EN 15713* as well as the ISO report *ISO/TR 15801*. With regard to specific application fields, for example the report *ISO/TS 21547* exists. Recommendations for data storage which specify a specific period are not available so far. An investigation on **sensors-specific** standards showed that ethical aspects are not represented appropriately yet.

Regarding **video surveillance**, the *CWA 16113* shows a few principles "to strike a balance between the rights of the individual and some important needs of civil society, on the other hand". Specific aspects related to video surveillance are also included in the *ISO 22311 Societal security - Video-surveillance - Export interoperability*. It calls for

- monitoring access to the data
- a mandatory storage time and an appropriate deletion of data after a relevant period
- training of staff in dealing with sensitive data.

ISO DIS 22311 includes the comment that privacy-specific aspects should be elaborated in more detail as soon as possible. It only describes the need to define a storage period but does not define the length of the period itself.

[11] found out that 40% of the European Society think that CCTV invades privacy. The participants of the InfraNorm survey mentioned specific risks regarding intelligent video surveillance. It is a new research field with specific privacy issues. Intelligent video surveillance is based on a combination of video elements, the application of data analysis methods and data storage. Compared with traditional forms of video surveillance a specification of the monitoring is indicative: While conventional systems record all events during a monitoring period, intelligent video surveillance systems only document detected events that deviate from the "act normal." Resulting problems include in particular risks of abuse and discrimination risks and possible intimidation effects (see [20]). Specific aspects related to these technologies are not included in the current version of ISO 22311. Therefore, additional work is needed.

The specific recommendation regarding privacy issues of **access control** need to be investigated in more detail. Private **security services** are no specific topic of the development of security technologies. Fundamental ethical issues regarding airport security services are for example in Europe covered by the standard *EN 16082*.

The need for specific regulations is for example stressed by [20] related to intelligent video surveillance.

As mentioned in the previous chapter, several participants in the InfraNorm survey described a need for appropriate certification schemes to show the fulfillment of specific

privacy-related requirements. In the IT-field, EuroPriSe, the European Privacy Seal was developed. It certifies that an IT product or IT-based service is compliant with European regulations on privacy and data protection. Although EuroPriSe has not strongly penetrated Europe to date (see [7]), it shows that the development of privacy-related certificates is not impossible. Similar projects, particularly in the fields of video surveillance, physical privacy and obtrusive detection are desirable.

6. FINAL REMARKS

In the preceding chapters, needs for different privacy standards in the security context were expressed. Three technology-related clusters with privacy risks were identified: surveillance solutions for detection from distance, solutions for obtrusive detection and data processing. According to Figure 2 the risk potential of detection from distance including video surveillance was top-ranked. Additionally, the ISO 22311 describes the need to extend privacy-specific rules for video surveillance as soon as possible. The specific needs to define privacy rules for contexts of public and semi-public security as well as the protection of critical infrastructures were also described. In summary, six new working items for new standards were suggested:

- A privacy standard for the protection of (semi-public) critical infrastructure, particularly airports and ports
- General definition of the kind of data stored for security reasons and of specific storage periods when there is no specific suspicion
- Matching of data
- Use of biometric data in the context of public security
- Ethical standards for sensors
- General requirements for the processing of video data, the storage period and the deletion when there is no specific suspicion.

Besides the identified topics for new standards, the need for a regulative document which covers ethical and privacy-related aspects of intelligent video surveillance which may be followed by specific standards was described.

Additionally, as shown in chapter 0, the new importance of privacy topics in the civil security field has caused the development of merely CWAs, specifications or technical reports in many areas in Europe. Several current CWAs relate to Directive 95/46 EC. The General Data Protection Regulation may require the development of new related standards. The development of formal standards based on selected CWAs and technical specifications and reports from the European standardization organizations CEN, CENELEC and ETSI is recommended. In the international arena often no similar document exists. Therefore, it is also recommended to establish international ISO, IEC or ITU standards based on the European documents described earlier.

Volunteers are needed in the different countries to start new standardization projects in order to realize these goals.

Standardization alone does not guarantee the realization of specific privacy-related requirements. As mentioned by several participants in the InfraNorm survey, appropriate certification schemes and procedures are necessary to ensure the implementation of the desired levels of privacy.

ACKNOWLEDGMENT

The author would like to thank the German Federal Ministry of Education and Research (BMBF) for the financial support.

REFERENCES

[1] Article 29 Data Protection Working Party (2007). Opinion 1/2007 on the Green Paper on Detection Technologies in the Work of Law. Enforcement, Customs and other Security Authorities http://www.dataprotection.ro/servlet/ViewDocument?id=227

[2] Blind, K. (2004). The Economics of Standards: Theory, Evidence, Policy. Cheltenham 2004.

[3] Blind, K. (2008). Standardization and Standards in Security Research and Emerging Security Markets. Fraunhofer Symposium 'Future Security', 3rd Security Research Conference Karlsruhe, 2008, 63-72.

[4] Blind, K., Gauch, S. (2007). Standardization benefits researchers – Standards ought to be developed in parallel to the research processes. In: Wissenschaftsmanagement, Special 2/2007 (Engl. Version), 16-17.

[5] CEN/CENELEC WG STAIR (2011). An Integrated Approach for Standardization, Innovation and Research. ftp://ftp.cencenelec.eu/PUB/ /Brochures/STAIR.pdf

[6] de Vries, H. J. (1999). Standards for the Nation. Analysis of National Standardization Organisations. Bosten, Dordrecht, London 1999.

[7] ECORYS (2011). Security Regulation, Conformity Assessment & Certification. Final Report. http://ec.europa.eu/enterprise /policies/security/files/doc/secerca_final_report_volume__1_main_ report_en.pdf

[8] ESRIF (2009). ESRIF Final Report. http://ec.europa.eu/ enterprise/policies/security/ files/esrif_final_report_en.pdf.

[9] European Commission (2004). Critical Infrastructure Protection in the fight against terrorism (COM/2004/0702). http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:52004DC0702:EN:HTML

[10] European Commission (2011). Programming Mandate Addressed to CEN, CENELEC and ETSI to Establish Security Standards. ftp://ftp.cencenelec.eu/CENELEC/EuropeanMandates/M_487.pdf

[11] Hempel, L., Toepfer, E. (2004). CCTV in Europe. Final Report. Urbaneye Working Paper No. 13 (August 2004).

[12] InfraNorm (2011). Bedeutung von Sicherheitsnormen, -standards und -spezifikationen. http://www.inno.tu-berlin.de/fileadmin/a38335100/PDF_Dateien/Publikationen/Infranorm_Studie_Sec_Normen.pdf

[13] ICO (2010) Privacy Impact Assessment Handbook 2.0. http://www.ico.gov.uk/upload/ documents/ pia_handbook_html_v2.

[14] ITU (2011). ICT Security Standards Roadmap. http://www.itu.int/ITU-T/studygroups/com17/ict/

[15] Kuechle, H. (2009). Bedrohungen und Schutz der kritischen Infrastruktur an Haefen, Flughaefen und Bahnhoefen. ZFAS (2009) 2:14–23.

[16] PRISE (2008). Legal Evaluation Report. http://prise.oeaw.ac.at/docs/PRISE_D3.2_Legal_Evaluation_Report.pdf

[17] Sinay, J. (2011). Security Research and Safety Aspects in Slovaika. In: Thoma, K. [ed.] (2010). European Perspectives on Security Research. Berlin Heidelberg 2011, 81-90.

[18] Swann, P. (2010). The economics of standardization: an update. Report for the UK Department of Business, Innovation and Skills (BIS). Complete Draft. Version 2.2, 27 May 2010.

[19] Wilkins, L., Christians, C. G. (2008). The handbook of mass media ethics. New York, 2009.

[20] Wuerttemberger, T. (2012). Rechtswissenschaftliche Begleitforschung zur intelligenten Videoueberwachung. BMBF-Innovationsforum „Zivile Sicherheit". http://www.bmbf.de/pubRD/B1-I_Wuerttemberger_Redemanuskript.pdf

[21] Wright, D., de Hert, P. [eds.] (2011). Privacy Impact Assessment. Law, Governance and Technology Series, Vol. 6. Dordrecht 2011.

SESSION 2

FUTURE COMMUNICATION SERVICES TO SUSTAIN COMMUNITIES

VISIBLE LIGHT COMMUNICATION USING SUSTAINABLE LED LIGHTS

Shinichiro Haruyama

Graduate School of System Design and Management,
Keio University, Japan

ABSTRACT

LED lights are becoming widely used for homes and offices for their luminous efficacy improvement. Visible light communication (VLC) is a new way of wireless communication using visible light. Typical transmitters used for visible light communication are visible light LEDs and receivers are photodiodes and image sensors. We present new applications which will be made possible by visible light communication technology. Location-based services are considered to be especially suitable for visible light communication applications.

Keywords— visible light communication, led, image sensor, photo diode, location-based service

1. INTRODUCTION

White LEDs have recently been used as efficient light sources replacing incandescent light bulbs and fluorescent lamps. Figure 1 shows the luminous efficacy improvement curves for LED lamps and luminaires [1]. Currently the luminous efficacy of LED lamps and luminaires is around 100 lm/W (lumens per Watt), and expected to reach 200 lm/W around 2025, which is much higher than incandescent lamps (around 20 lm/W) and fluorescent lights (around 100 lm/W). LED lamps do not only have high luminous efficacy, but also long sustainability. LED lamps typically have a lifetime of 40,000 hours, which is 40 times longer than incandescent lamps.

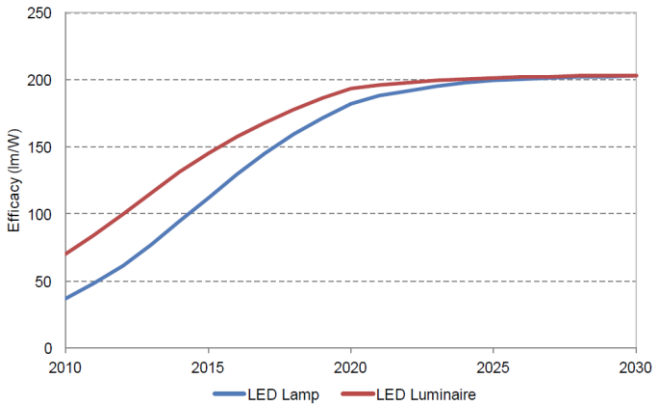


Figure 1. LED Luminous Efficacy Improvement

As the LED light technology improves, the price of LED light is falling rapidly as shown in Figure 2 [2]. The price of a 60 Watt LED light is expected to break US $10 in 2014 and US $5 in 2020.

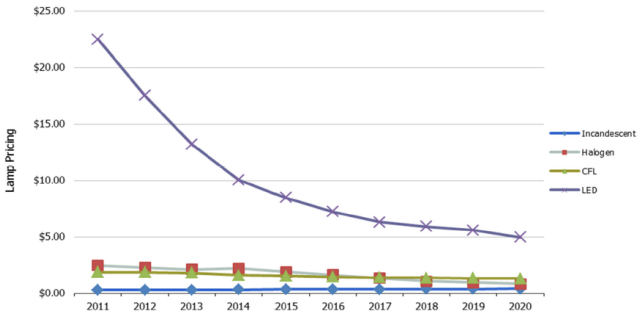


Figure 2. 60 Watt LED light price Trend

Thanks to the luminous efficacy improvement and long sustainability along with the lowering cost, LED lights are gaining a larger share every year. Its share will become 64 percent of the global lighting product market as shown in Figure 3 [3].

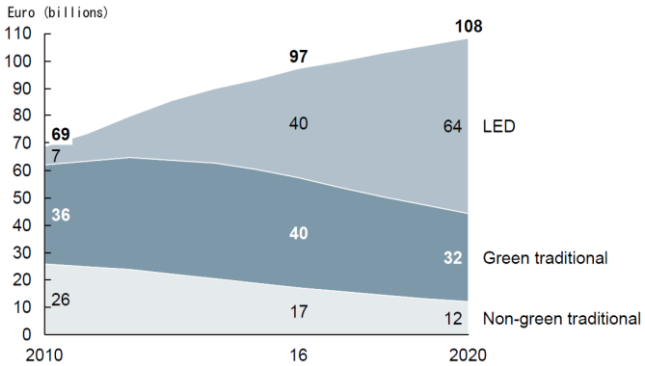


Figure 3. Global Lighting Product Market Trend in the World

("Green" is defined in line with typical energy efficiency standards, e.g., Energy Star for CFL lightbulbs.)

Governments in many countries are starting to ban inefficient lighting technologies such as incandescent lamps as shown in Figure 4 [3]. Countries such as USA, EU, Japan, China, Russia, and Brazil started to ban 100 Watt

incandescent lamps by the end of 2012, and most will ban all incandescent lamps by 2016.

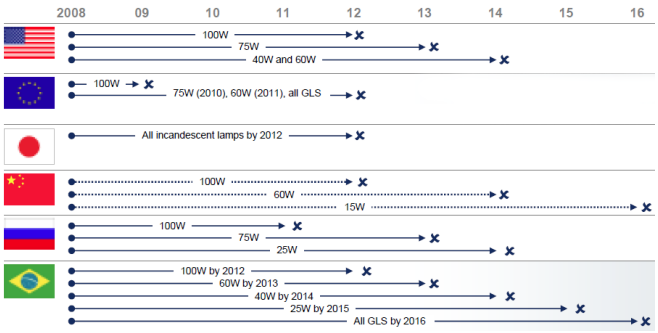


Figure 4. Incandescent Ban Plan in Residential Lighting in the World

Using increasingly popular LED lights, the visible light communication using LEDs is expected as a means for ubiquitous communication.

Visible light communication (VLC) has following advantages over other competing radio communication technologies such as WiFi and cellular phone wireless communication [4]: Visible light spectrum is available for communication because the frequency above 3THz is not currently regulated by the Radio Regulation Law. Visible light does not penetrate thick materials such as walls and partitions, which can be a security advantage. Visible light usually poses no health hazards to human body and eyes. Visible light also has following advantages over infrared communication technologies: Visible light can be literally visible so that human notices where the data is transmitted from. In addition, since LED lighting has recently become part of a building infrastructure, making visible light communication infrastructure is fairly easy by adding communication function to LED lighting.

Figure 5 shows a representative use of visible light communication, where an LED light is used as a data transmitter and a cellular phone with visible light sensor is used as a data receiver. The application of this system is indoor location service where a user uses a cellular phone with a photo diode, which detects signals from an LED light. This application is especially useful indoor because GPS receivers do not work well indoors even though they work well outdoors.

Figure 5. Visible Light Communication using LED Light: NEC, Matsushita Electric Works, Ltd, Keio University, CEATEC demonstration, Japan, 2004

The LED backlight of an LCD display can also be used as a data transmitter as shown in Figure 6. This is a visible light communication system made by NEC Corporation and Fuji Television in 2007, whose transmitter is an LED backlight of an LCD display and the receiver is a PIN photo diode attached to a PDA. While the regular image content is being displayed on the LCD screen, the LEDs in the backlight are turned on/off at a high speed to transmit text data. The transmitted data is received by a PDA device so that the text is displayed on the PDA. This system allows the transmission of information to hearing-impaired people or sight-impaired people.

Figure 6. Visible Light Communication to send Information using LED Backlight of LCD Display

The application of visible light communication using LED backlight panels as transmitter and PIN photo diode as receiver is shown in Figure 7. This prototype of digital signage was made by Visible Light Communications Consortium (VLCC) in Japan in September 2009. In this application, backlight LEDs send advertisement information, which is received by a user's terminal using a PIN photo diode. VLCC has been working with JEITA (Japan Electronics and Information Technology Industries Association) to define the standard of visible light ID system which can be used for applications such as location-based services and digital signage.

Figure 7. Visible Light Communication to send Advertisement Information using LED Backlight

When an LED light is used for illumination, its brightness has to be controlled. [5] discusses about the dimming control of LED lights as well as visible light communication for IEEE 802.15.7 standard, which was IEEE's first Wireless Personal Area Network (WPAN) standard for visible light communication. The IEEE 802.15.7 defines PHY and MAC layer for both bi-directional communication mode and broadcasting mode.

2. PROPERTIES OF VISIBLE LIGHT COMMUNICATION

Visible light communication has properties that are both advantageous and disadvantageous compared to radio-wave wireless communication. Its disadvantages are communication distance and data rate. The communication distance using visible light communication is typically between 1 to 100 meters. This distance is short compared to radio-wave communication, due to the fact that visible light communication is basically line-of-sight communication, which means that communication is interrupted when there is an object between a transmitter and a receiver. There is another disadvantage of visible light communication, which is data rate. Its data rate is typically between kilobits per second to 10 megabits per second, although there have been active researches going on to reach the speed of gigabits per second [6]. The bottle neck of the data rate is caused by the performance of either white LEDs or receiving photo sensors. The above disadvantageous properties of visible light communication may limit its use for many applications. However, these seemingly disadvantageous properties are indeed useful for some applications by taking advantage of line-of-sight property. We believe that those useful applications include location-based services and new graphical user interfaces that combine visual imagery with visible light communication. In this paper, we will explain some useful applications using the advantageous properties of visible light communication.

3. LOCATION-BASED SERVICES USING PHOTODIODE AS RECEIVER

We believe that the applications of visible light communication to location-based services and new graphical user interfaces that combine visual imagery with visible light communication have potential widespread use. For these applications, users are able to know the information associated with a transmitter. If a transmitter is attached to a building or a fixed place, location information will be obtained.

Indoor navigation is convenient for everyone, and it is especially indispensable for the visually impaired. We proposed such a navigation system for the visually impaired as shown in Figure 8 [7]. LED lights emit visible light with location data and a smartphone with a visible light receiver receives the data. The smartphone calculates the optimal path to a designation and speaks to the visually impaired through a headphone.

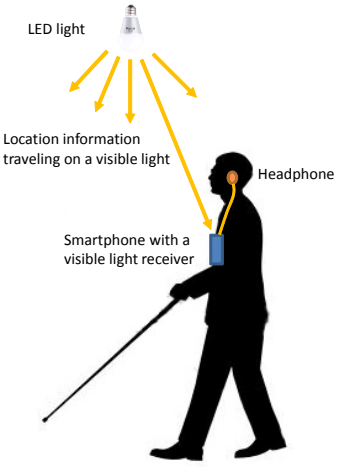


Figure 8. Indoor Navigation System for the Visually Impaired using Visible Light Communication

We made its navigation prototype and tested it for the visually impaired in 2012 as shown in Figure 9. We found that the prototype was able to navigate the visually impaired users fairly well with speech guidance.

Figure 9. Indoor Navigation Prototype for the Visually Impaired using Visible Light Communication

A prototype for a supermarket made by Nakagawa Laboratory in Tokyo is shown in Figure 10. An LED light sends location information and a shopping cart with a photodiode receives the location information.

LED Light sending location data

Shopping cart with photodiode near the wheels

Figure 10. Customer flow analysis system for a supermarket

The data of the path of a shopping cart can be recoded in a memory installed in the cart for one week with a battery. After all the data of all the carts are gathered, a statistical analysis is performed, and the example of the analysis is shown in Figure 11.

Figure 11. Example of customer flow analysis at a supermarket

The thick lines at a particular point in the supermarket indicate that many customers with carts walked through that point. Using this system, a supermarket manager is able to rearrange the goods to sell or improve the floor plan.

4. LOCATION-BASED SERVICES USING IMAGE SENSOR AS RECEIVER

Another interesting device that can be a receiver of visible light communication is an image sensor. An image sensor is able to do simultaneous image acquisition and data reception. Figure 12 shows the concept of visible light communication using image sensor as receiver. An image sensor continuously takes images of a scene with an LED light whose light intensity is modulated and a receiver detects the optical intensity at a pixel where the LED light is focused on.

Image sensors used for digital cameras or video cameras usually have frame rate of tens of frames per second. If a visible light signal from a visible light LED is received at a pixel of such an image sensor, the data rate is on the order of only several bits per second. However, using a high-speed image sensor whose frame rate is thousands of frames per second, it is possible to achieve data rate on the order of kilobits per second.

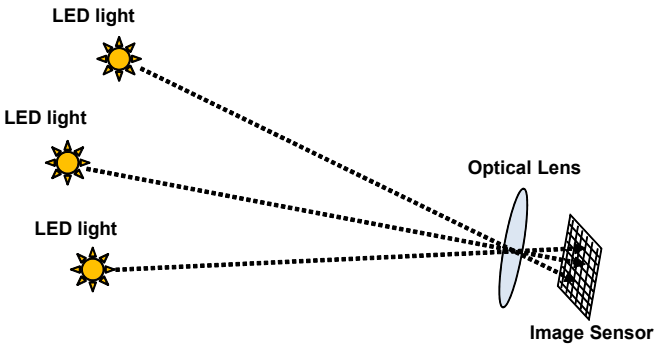


Figure 12. Concept of Visible Light Communication using Image Sensors as Receivers

The advantage of using an image sensor as receiver over a single photo diode is that even if there is a strong interfering

light along with a desired signal, the interfering light will be focused onto a pixel which is different from a pixel onto which a desired signal is focused. This implies that image sensor reception is much more robust against interference than single photo diode reception.

A typical example of visible light communication using image sensors is shown in Figure 13 through 16. In Figure 13, a combination of digital camera and visible light communication is shown. LED transmitters are attached to users. The data associated with a user is sent from the LED transmitter and an image sensor detects not only its direction of a transmitter in an image, but also its received data contents. The monitor displays its contents at a location in an image where the data is sent from.

Figure 13. Application Example of Visible Light Communication using Image Sensors as Receivers
(Photo: Courtesy of Mr. N. Iizuka, Casio Computer Co., Ltd., Demonstration of Image Sensor Receiver, 2008)

Another application of visible light communication using image sensors is accurate position detection. Two examples of accurate position detection are shown: one in Figure 14 and the other in Figures 15 and 16.

In the first example in Figure 14, a photogrammetric method is used to detect the locations of LEDs attached to a water tank and the ground [8], [9]. The accuracy of position using photogrammetric method and visible light communication was about several millimeters at a distance of about 50 meters away from an image sensor. This accuracy of position is comparable to that of a typical surveying device called a total station. This system has another advantage over a total station, which is continuous monitoring of positions over time. The positions of LEDs attached to a water tank in Figure 14 were monitored for 24 hours. The roof of the water tank expands in the daytime and shrinks at night due to the heat from the sunshine. The visible light communication photogrammetric system was able to detect the position displacement of several millimeters with an accuracy of a millimeter at the distance of 40 meters.

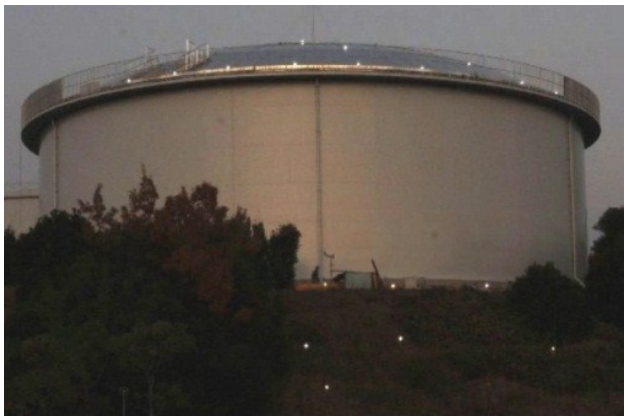


Figure 14. Survey Measurement using Image Sensors as Receivers

In the second example in Figures 15 and 16, a mobile robot detects its position by receiving location data from LED lights using two image sensors [10].

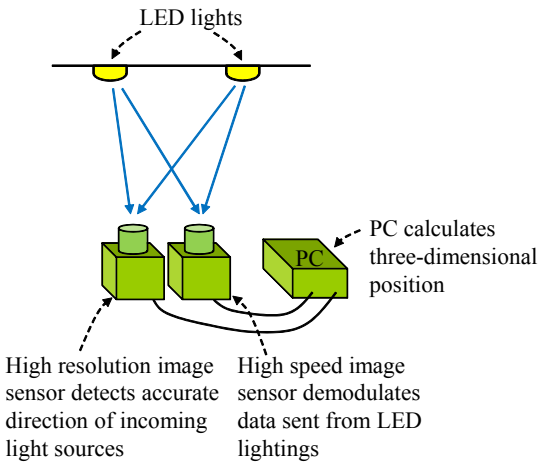


Figure 15. Robot Control System using Visible Light Communication

Figure 15 shows a robot control system using visible light communication. LED lights on the ceiling send location data. A robot has two image sensors: one image sensor obtains high-resolution image to detect an accurate direction of incoming light, and the other image sensor obtains high frame rate images in order to demodulate the incoming data on a visible light.

Figure 16. Robot Control Prototype with Centimeter Accuracy

We made a robot control prototype as shown in Figure 16 and found that a robot was able to detect its position with an accuracy of centimeter, which was good enough to control its motion.

5. CONCLUSION

We showed advantages and disadvantages of visible light communication and explained the effectiveness of location-based services for visible light communication by showing some examples. It is expected that visible light communication will be widely used as LED light market expands worldwide.

REFERENCES

[1] US Department of Energy, Energy Efficiency & Renewable Energy, Building Technologies Program, "Energy Savings Potential of Solid-State Lighting in General Illumination Applications", January 2012

[2] Philip Smallwood, "Global 60W Replacement Lamp Market", LED Lighting Evolution Conference, Boston, Massachusetts, USA, June 2012

[3] McKinsey & Company, "Lighting the way: Perspectives on the global lighting market", 2nd Edition, August 2012

[4] Kavehrad, M., " Sustainable energy-efficient wireless applications using light", IEEE Communications Magazine, Volume 48, Issue 12, pp. 66 - 73, December 2010

[5] Rajagopal, S., Roberts, R.D., Sang-Kyu Lim, "IEEE 802.15.7 visible light communication: modulation schemes and dimming support", IEEE Communications Magazine, Volume 50, Issue 3, pp. 72 - 82, March 2012

[6] Christoph Kottke, Jonas Hilt, Kai Habel, Jelena Vucic, and Klaus-Dieter Langer, "1.25 Gbit/s Visible Light WDM Link based on DMT Modulation of a Single RGB LED Luminary", Proc. European Conference on Optical Communications (ECOC 2012), Amsterdam, The Netherlands, September 2012

[7] Madoka Nakajima, Shinichiro Haruyama, "Indoor navigation system for visually impaired people using visible light communication and compensated geomagnetic sensing", 2012 1st IEEE International Conference on Communications in China (ICCC 2012), August 2012

[8] Hideaki Uchiyama, Masaki Yoshino, Hideo Saito, Masao Nakagawa, Shinichiro Haruyama, Takao Kakehashi, Naoki Nagamoto, Proc. 34th Annual Conference of IEEE Industrial Electronics Society (IECON), pp.1771-1776, Florida, USA , November 2008

[9] H. Mikami et al., Reports of Technical Research and Development of Sumitomo Mitsui Construction Co., Ltd., No.9, pp. 79-84, 2011

[10] Toshiya Tanaka, Shinichiro Haruyama, "New Position Detection Method using Image Sensor and Visible Light LEDs", Proc. IEEE Second International Conference on Machine Vision (ICMV), pp. 150 – 153, December 2009

SELECTING THE BEST COMMUNICATION SERVICE IN FUTURE NETWORK ARCHITECTURES

Rahamatullah Khondoker *, *Paul Mueller*

Integrated Communication Systems
University of Kaiserslautern
Kaiserslautern, Germany
{khondoker, pmueller}@informatik.uni-kl.de

Kpatcha Bayarou

Fraunhofer Institute for Secure
Information Technology (FhG-SIT)
Darmstadt, Germany
kpatcha.bayarou@sit.fraunhofer.de

ABSTRACT

As the number of future network architectural approaches increases, the possibility of offering many similar services with different qualities of service is increasing. Therefore, it will be required to select a suitable, or the best, service from the set of alternative services. This paper proposes a matching process and an adapted analytic hierarchy process to accomplish this task. The matching process is used to determine if a service is suitable. When more than one suitable service is available, the adapted analytic hierarchy process is used to select the best service.

Keywords— Future Internet, NGN, service-orientation, service description, network architectures, service selection

1. INTRODUCTION

In todays Internet, protocols are tightly coupled with the application, which results in difficulties in automatically switching between the functionalities based on the application requirements. Traditionally, an email application uses TCP, a Voice over IP (VoIP) application uses UDP, some video streaming applications use SCTP. However, a video application cannot just switch between UDP and SCTP based on its variety of demands.

For introducing flexibility in network architectures and enabling innovations, several projects like GENI, FIND, G - Lab, PL - Lab, AKARI, have been funded in USA, Europe and Asia. The results of these projects are a set of future network architectures like Autonomic Network Architecture (ANA) [1], Netlet-based Node Architecture (NENA) [2], eXpressive Internetwork Architecture (XIA) [3], Service-Oriented Network Architectures (SONATE) [4] and Recursive InterNetwork Architecture (RINA) [5].

Some of these approaches are based on communication services. Here we consider only communication services not web services. A communication service can represent a fine-grained functionality like an algorithm for forward error correction (e.g., hamming code) or compression (e.g., Huffman tree) or it can even represent a coarse-grained functionality like the functionality of the TCP/IP network stack or an access technology like WiFi.

Most of future network architectural approaches need to use a suitable service, or to select the best service, if there more than one suitable service is available. Selection of a suitable service can be done by matching the description of the offered services with the application requirements. This match can result in a several suitable services. Now, the question is, which suitable service should be selected and used? The answer is that we should select the best one, as we do in our day to day life.

Selecting the best service using a single selection criterion is trivial. For example, if there are two communication services where one offers 100ms end-to-end delay and another offers 200ms, then we should obviously select the one with the lowest delay.

However, communication services have multiple selection criteria such as delay, throughput, loss ratio, jitter and cost. That is why, selecting the best communication service is a Multi-Criteria Decision Making problem (MCDM). For solving such a problem, several Multi-Criteria Decision Analysis (MCDA) approaches are used in managerial science like Multiple Attribute Utility Theory (MAUT), Analytic Hierarchy Process (AHP), ELECTREIII and Evamix [6].

We used AHP to select the best service for two reasons, firstly, it supports relative prioritization and, secondly, there is a way to check the consistency of the evaluation measures.

The main requirement for using AHP is to assign pairwise priority both for the requirements and for the offers. However, as offerings are decoupled from the application requirements, a mapping mechanism is required from the measured values of the offerings to the pairwise priority assignment scale. We use a mapping mechanism based on monotonic interpolation and extrapolation.

The outline of the paper is as follows. We present a service selection model in section 2. The components of communication service selection using the analytic hierarchy process are discussed in section 4. In a service-oriented network architecture, offerings are decoupled from the application, for

*Main corresponding author. He is a PhD student at the department of computer science in the University of Kaiserslautern. Moreover, he is affiliated with the Fraunhofer Institute for Secure Information Technology located in Darmstadt, Germany

this reason a mapping mechanism is necessary to map from the measured value of the offers to the pairwise prioritization scale. We propose a mapping mechanism using monotonic interpolation and extrapolation in section 4.3.2. We implemented and evaluated the selection process using a maximum of six selection criteria and six services which is discussed in section 5. After that, related work for future network architectural approaches and service selection is presented in section 6. Section 7 concludes the paper.

2. SERVICE SELECTION MODEL

A model for fine-grained service selection and composition is shown in figure 1. The main aim of the process is to create a protocol graph (i.e., a network stack) for a network connection. To achieve this goal, it takes the requirements from the application, constraints from the network, policies from the network or system administrator, and the offered services from the network. Considering all of these inputs, it composes the protocol graph of building blocks (the implementation of a protocol or a mechanism). Automatic selection of a suitable, or the best, fine-grained functionality is required during the composition process.

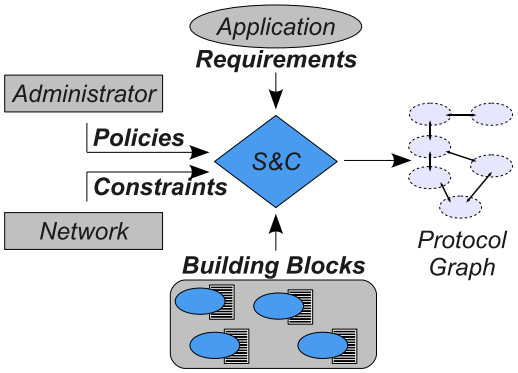


Figure 1. A model for fine-grained service selection and composition

A model for coarse-grained service selection is shown in figure 2. The three main entities in this model are the service consumer, the service provider and the service broker. The service broker selects a suitable, or the best service, from the services offered by the different service providers by considering the requirements specified by (or chosen from the predefined specification) an application developer through an application programming interface (API). Service providers like SONATE and NENA frameworks can be categorized based on their composition approaches. Services can be offered by conventional providers like TCP / IP, UDP / IP and SCTP / IP. Services can be composed during design time, deployment time, partial runtime and runtime. In compound approaches, services are composed during design time, potentially assisted by software. In this approach, the selection of an appropriate compound service is done during runtime by a service broker. The template approach is an example of partial runtime composition, where the placement of functionalities is done during design time and a suitable, or the best, mechanism is chosen during runtime. Services can also be provided by a dynamic selection and composition provider where the selection and composition of the protocol graph is done during runtime.

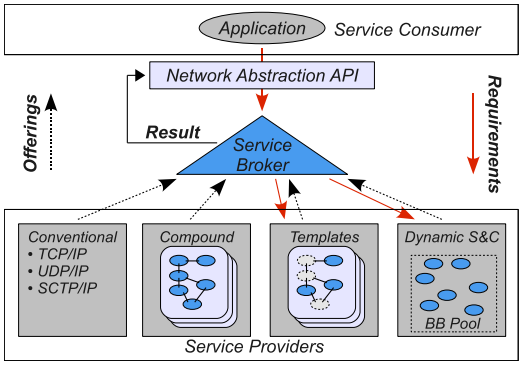


Figure 2. A model for coarse-grained service selection in a service-oriented network architecture

Partial runtime and dynamic selection and composition providers cannot register their services to the broker until they get the application's requirements and perform their composition. Other providers can register their service to the broker beforehand.

The service broker returns a suitable, or the best, service to the application through the API.

3. TERMINOLOGY: CRITERIA FOR SERVICE SELECTION

The criteria that are used to select a suitable, or the best, communication service are specified by the field expert. The assumption here is that, an experienced VoIP application developer knows the criteria that should be considered for his application. Even though functional criteria are also considered in service selection, we considered here only the following quality of service criteria:

1. Delay: Delay is defined as the elapsed time to transfer a packet from the sender's application to the destination receiver's application across the network. Delay is measured by seconds or fraction of seconds. [7]

2. Jitter: Variation in delay of packets arriving in the destination.

3. Energy Consumption: The power that is required to process a packet is called energy consumption. Energy consumption is usually measured in Joules (J).

4. Data Length: The length of a packet consisting of a payload of data and a header is called data length (sometimes called packet length).

5. Loss Rate: When a transmitted packet does not successfully arrive at its destination, it is called a lost packet. The loss rate is the ratio of the number of lost packets and the total number of sent packets. Loss rate can also be called Packet Loss Ratio (PLR). [7]

6. Throughput: The average rate of successful data delivery over a wired or wireless communication is called throughput. Throughput is measured in bits per second, in short, bits/sec or bps. [7]

4. COMPONENTS FOR SERVICE SELECTION

Service selection is a process to select a suitable service, or the best service if more than one suitable service is available. The components of our service selection approach are:

1. Description of application requirements and network offerings
2. Matching process
3. Analytic Hierarchy Process
4. Network abstraction API

4.1. Description of application requirements and network offerings

Service selection requires the description of application requirements, network and administrator constraints, and network offerings. This requirement can be fulfilled by the description language for communication services of future network architectures [8]. All of these requirements, constraints and offerings can be described by using the construct $\{effect\ operator\ attribute\}$.

An effect is a single outcome of an execution of algorithm or protocols, sometimes called building blocks. Effects can be functional and non-functional. Functional effects are the effects which are required for proper functioning of a building block. For example, the effect $LossRatio$ can be used by the retransmission building block to know how many packets to retransmit. Non-functional effects, on the other hand, are the effects which might not be necessary for functioning. For example, the processing time of a building block can be seen as an example of such an effect.

An attribute is the value of an effect. For example, 0% can be seen as an attribute of the effect packet loss.

An operator connects an effect to an attribute. The packet loss offering of a retransmission building block can be written as $\{LossRatio = 0\%\}$.

This simple construct can be used to express the requirements of an application. For example, the error correction demand of an email application can be expressed as $\{ErrorCorrection = True\}$.

The usage of an effect in the description is mandatory. But, the usage of an operator and an attribute is optional. For example, the error correction demand can be described as $\{ErrorCorrection\}$ by omitting an operator and an attribute.

This construct allows the description of the network offerings. For example, the packet loss offering of a forward error correction algorithm can be expressed as $\{LossRatio = 0\%\}, \{Delay = low\}, \{Bandwidth = high\}$.

A network or administrator constraints can be expressed by using the construct. For example, for using a certain network, authentication must be performed $\{Authentication = True\}$.

This construct supports to describe both fine-grained and coarse-grained functionality in a similar way. For example, the ProcessingTime of a single building block or a protocol graph can be expressed by using the same construct.

4.2. Matching process

Suitable services are chosen by matching the offered effects with the required effects. For example, an application can support the maximum end-to-end delay of 100 ms which is expressed by $\{end-to-endDelay <= 100ms\}$ whereas a protocol graph offers $\{end-to-endDelay = 80ms\}$. The broker can select the protocol graph as a suitable service.

For matching application requirements with the network offerings, each effect must be uniquely identified. This necessitates developing a taxonomy of effects to describe communication service illustrated in the ITU-T paper [9]. This taxonomy facilitates an application developer to specify effects either in a generic manner or in a specific way. For example, an application developer can ask for the Security effect in general, $\{Security = True\}$, or it can ask for the data origin authentication effect, $\{Data-Origin-Authentication = True\}$, to be more precise.

As the values of the offered effects are measured or precalculated values, mostly, they contain the operator is equal to $(=)$.

But, the required effects might contain other operators like less than $(<)$, less than or equal to $(<=)$, greater than $(>)$ and greater than or equal to $(>=)$.

An application might also express its requirements as an interval. For example, a video streaming application might express its packet loss requirement as $\{LossRatio <= 3\%\}$.

The application can work when the packet loss is between 0% and 3%.

During the selection process of fine-grained or coarse-grained functionalities, several of them can be determined as suitable services when they match the requirements from the application. In that case, the best service should be selected and used. We adapted Analytic Hierarchy Process (AHP) for doing this task.

4.3. Analytic Hierarchy Process (AHP) for service selection

Selecting the best service using a single selection criterion is trivial. For example, if there are two communication services

where one offers 100ms end-to-end delay and another offers 200ms, then we should obviously select the one with less delay.

However, communication services have multiple selection criteria such as delay, throughput, loss ratio, jitter and cost. That is why, selecting the best communication service is a Multi-Criteria Decision Making problem (MCDM). For solving such a problem, several Multi-Criteria Decision Analysis (MCDA) approaches are used in managerial science like Analytic Hierarchy Process (AHP) [10], ELECTREIII [11], Evamix [12], Multiple Attribute Utility Theory (MAUT) [13], Multi - Objective - Programming (MOP), Goal Programming (GP) [14], NAIADE [15] and Regime [16].

We used AHP to select the best service for two reasons, firstly, it uses an absolute scale to derive priorities that also belong to the relative absolute scale (like probabilities) that can be combined like the real number system. secondly, there is a way to check the consistency of the evaluation measures.

4.3.1. Adaptation of Analytic Hierarchy Process (AHP) for service selection

The Analytic Hierarchy Process (AHP) needs to be adapted for selecting the best communication service automatically.

AHP is a process designed for assisting human decision making which is used in many application areas like social, personal, education, manufacturing, political, engineering, industry and government [17]. Basically, AHP is used for determining priorities of different alternatives. The details of the AHP process is beyond the scope of this text.

To use AHP in communication service selection, the following steps are performed

1. Define the goal and the selection criteria for achieving the goal
2. Priority assignment of the selection criteria as an application requirement
3. Priority assignment of the criteria for the offered services

The first step is to define the goal, which is to select the best communication service, and the selection criteria to achieve that goal. The selection criteria are actually a set of required effects. Examples of selection criteria are delay, throughput, loss rate, jitter, MTU and cost. Both functional and non-functional criteria can be selected.

After determining the selection criteria, the next step is to assign pairwise priority between the selection criteria. One of the reasons of pairwise priority assignment is that it is easier for a person to take two criteria and to assign priority one over the other. It is initially difficult for a new application developer to assign pairwise priority. But, the efficiency of the priority assignment process can be improved with the experience of the application developer.

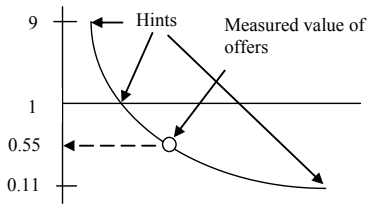


Figure 3. Mapping mechanism

The third step of the process is to assign pairwise priority between the offered services based on those selection criteria. However, as pairwise priority assignment is a time-consuming task, and as offerings are decoupled from the application, the pairwise priority assignment of the offered services based on those selection criteria needs to automated.

This requires a mapping mechanism to map the measured/ calculated values of the offered services to the pairwise priority assignment scale which will be discussed in the next section.

The priority vector coming from the application side is then multiplied by the priority vector from the offering side. The result is then called the overall priority vector. The service with the highest priority value in the overall priority vector is the best service.

4.3.2. Automated priority assignment for the offerings

Different communication services can have different effects. The value (or attribute) of these effects can be assigned beforehand based on benchmarks or can be obtained dynamically by using sensing software. Whichever way the attributes are obtained, the offered effects need to be automatically prioritized as the offerings are decoupled/hidden from the application. Therefore, an automatic mapping mechanism from measured values to the priority scale (1, 9) is required.

The mapping should have certain properties. First, the mapping must be generic, i.e. not specific to effects or units of measured values. Second, the mapping must be monotonic.

An approach for mapping has been proposed which uses a monotonic interpolation/extrapolation scheme [9] as shown in figure 3. In this case, the application requirements provide value points for interpolation/ extrapolation (must be monotonic) of measured values to the priority scale. A monotonic interpolation/extrapolation of these points is used to define a mapping. In addition, the specific measured values of the offerings are then mapped to these priorities. Assuming that $f()$ is a function used to define a mapping. As an example, considering interpolation, the requirements must contain at least the following two points

- $x_0, where f(x_0) = 1$
- $x_n, where f(x_n) = 9$

If there are measurement values, y, not within the interval $[x_0, x_n]$, we can extrapolate

- $if y < x_0, then f(y) = 1$
- $if y > x_n, then f(y) = 9$

To use inter-/extrapolation, an application developer must specify two points but can have as many parameters as he wants to be more precise.

The aforementioned mapping mechanism is used to assign a priority of one service over another for every selection criteria (effect).

4.4. Network abstraction API

An application programming interface (API) is required to send the application requirements to the broker and to return a suitable or the best service to the application. Affiliated with the SIG FUNCOMP, a special interest group for functional composition of the German-Lab project, we created an interface titled *GAPI: A G-Lab Application-to-Network Interface* which can be used for this purpose [18].

5. IMPLEMENTATION

The aforementioned service selection process has been implemented using the Java programming language version 1.6. The requirements and offerings are assigned statically in variables. No database is used to store those values. A separate method has been implemented to map the offered values to priorities as shown in the figure 4.

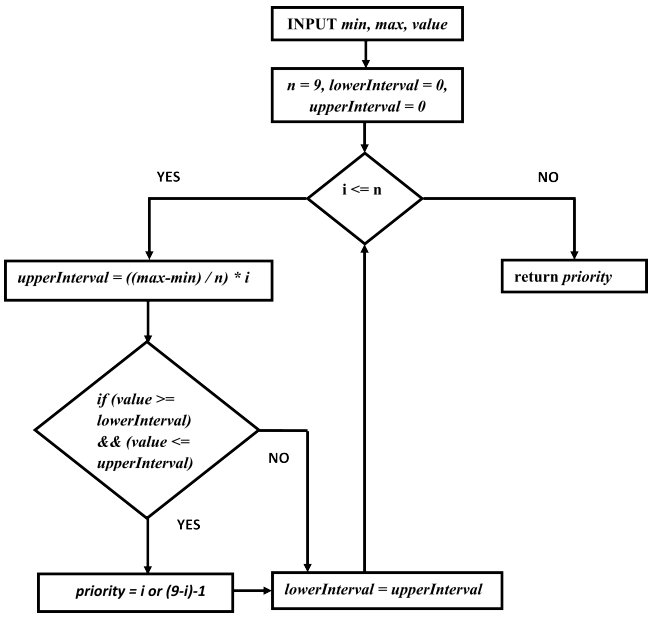


Figure 4. Priority assignment algorithm

5.1. Service selection time

To measure selection time, six selection criteria have been chosen and pairwise prioritized as shown in figure 5. The values of 0.11, 1 and 9 means that the lowest, equal and the highest priorities respectively. The measured values of the offerings is shown in the figure 6. For this experiment, CentOS is used on a Pentium(R) Dual-Core CPU E5300 with 2.6 GHz speed and 6 GB RAM.

Delay	Throughput	Jitter	Loss Rate	Energy Consumption	Data Length
1	5	9	5	9	5
0.2	1	1	2	1	1
0.11	1	1	1	1	1
0.2	0.5	1	1	2	5
0.11	1	1	0.5	1	1
0.2	1	1	0.2	1	1

Figure 5. Pairwise priorities of the six selection criteria

Effects	S1	S2	S3	S4	S5	S6
Delay	10	50	250	200	220	260
Throughput	1	2	10	12	15	32
Jitter	1	2	10	15	17	20
Loss Rate	2	4	5	8	10	15
Energy Consumption	10	50	40	70	120	100
Data Length	1500	500	600	1400	700	1000

Figure 6. Measured values of the offered services

Beginning with the two selection criteria and two services, both selection criteria and the offered services have been incremented by 1 until 6 and the service selection and mapping times have been measured. We found that the mapping time is linearly increased with 23 micro seconds is required for mapping the 6 services using the six selection criteria as shown in the figure 7. Selection time is exponentially increased and requires 0.48 ms to select the best service among the six offered services using the 6 selection criteria as shown in figure 8.

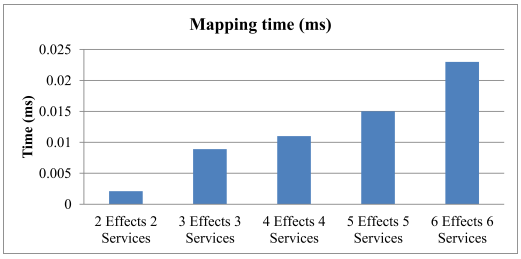


Figure 7. Mapping time

Mapping can be done during runtime or beforehand, when the measured values are already available. In that case, only selection time is considered.

5.2. Benefits and future work

This selection approach has several advantages; first, pairwise prioritization of requirements as an input, second, con-

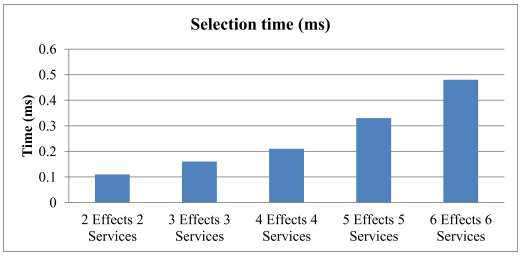


Figure 8. Selection time

sistency checking, third, benefits of relative prioritization over linear prioritization.

It is easy for people to compare two objects by using their properties. For example, a recruitment manager needs to select the best candidate for the job. One candidate has an excellent education but no working experience and another person has a good education but has 2 years of experience. The manager will take these two selection criteria of the candidate (education, working experience) and can easily identify which is more important to him. If working experience is more important to him, he will select the second candidate and otherwise he will select the first one.

When the number of selection criteria increases, the consistency of the pairwise priority assignment needs to be checked. As discussed earlier, the analytic hierarchy process provides a way to check consistency.

AHP uses relative prioritization rather than linear prioritization which is used in MAUT. In linear prioritization, the priority value of the requirement is assigned linearly like $delay > throughput > loss$ which means that a service with the lowest delay should be selected at first. If two services have the same delay, then the service with the highest throughput is selected. In relative prioritization, the selection criteria is pairwise prioritized. That means, a service is selected based on all of the considered criteria not only a single criteria like in a linear prioritization technique.

Currently, the selection time is calculated by considering at most six effects and services, as today the number of networking services and selection criteria is limited. However, in the future, evaluation can be done by increasing the number of services and criteria.

6. RELATED WORK

The work related to layerless future network architectures is presented at first. Then, the work for service selection is presented.

In the early 1990s, a small group of network researchers concentrated on dynamic micro-protocol composition, meaning that they decomposed the functionality of existing protocol stacks into a set of micro-protocols, and then composed those micro-protocols dynamically based on incoming requests from an application. Some of those works are Dynamic Configuration of Protocols (DaCaPo) [19] and Function Based Communication Subsystem (FCSS) [20]. In [21] the authors point to a drawback of the above approaches and ask for a generic description so that new deployments can be facilitated and implementation customization can be kept to a minimum. [22] focused on networking protocols rather than the functionality, services or roles provided by those protocols which were focused on by [23] and [24].

Some recently completed and ongoing projects are working on Network Functional Composition. Those projects are Automatic Network Architecture (ANA) [1], NetServ [25], Recursive InterNetwork Architecture (RINA) [5], eXpressive Internet Architecture (XIA) [3], Forwarding on Gates (FoG) [26], Net-Silo [24], 4WARD [27], Self-Net (Self-Management if Cognitive Future Internet Elements) [28] and the Recursive Network Architecture (RNA) [29]. Descriptions of some of the aforementioned projects has been comprised in a state-of-the-art paper [30].

A template-based approach is similar in concept to the NENA approach. In the NENA approach, netlets (i.e., a network stack) for each domain are composed during design time by network engineers assisted by software. Selection of an appropriate netlet is done during runtime by using MAUT [31]. However, selection of appropriate mechanisms (i.e., building blocks) is not done in the NENA approach. In the template-based approach, not only are appropriate templates selected at runtime but also appropriate mechanisms are selected.

As selecting communication services to make a protocol stack automatically is a new field, few related works have been found. The mentionable one is a MCDA approach, MAUT which is used in NENA to select the best composed protocol stack during runtime. However, MAUT has no integrated mechanism to check consistency of the given priorities. That is why, an external mechanism is required for doing this task which is not available.

Most of the approaches right now use static selection of functionality during design time. Some of those approaches are ANA, RINA, XIA and FoG.

7. CONCLUSION

Driven by Future Internet projects like GENI and FIND, worldwide research of future network architectures results in several architectural approaches like NENA, XIA, SONATE, RINA, and ANA, to name a few. Even though the same service with different qualities of attributes can be offered by the same architecture, the probability of having such a case can be even higher when there are many architectural approaches and virtualization techniques.

Therefore, selecting a suitable, or the best service, based on application requirements is essential. A suitable service can be selected just by matching the description of the offered services with the requirements. Selection of the best service is required.

Selecting the best service using a single criterion is trivial. For example, considering a single selection criterion delay,

Table 1. The requirements matrix (CR = 6.23%)

Effects	Delay	Throughput	Jitter	Priority
Delay	1	5	9	0.7651
Throughput	0.2	1	1	0.1288
Jitter	0.11	1	1	0.1062

the best service is the one with the lowest delay. However, communication services have multiple selection criteria. That is why, selecting the best service is a multi-criteria decision making problem.

For solving such a problem, different multi-criteria decision analysis methods exist in management science. For example, MAUT, AHP, Evamix, Regime, ELECTRE III, NAIADE and MOP/GP. We chose the Analytic Hierarchy Process (AHP) for communication service selection as it supports relative prioritization and checks consistency.

However, the process is required to be adapted for communication service selection. In a service oriented network architecture, offerings are decoupled from the application. That is why the measured or estimated values of the offered services need to be mapped based on the hints coming from the application. This is done by the proposed mapping mechanism.

We implemented the process of selecting the best service in the Java programming language and evaluated using at most six selection criteria (effects) and offered services. The result shows that 0.503 milliseconds (selection time (.48 ms) + mapping time (.023 ms)) is required to select the best service between six offered services using six selection criteria.

To conclude, applications use networks differently, and therefore have different network requirements. At the same time, networking capabilities and protocols make advances. This paper shows how applications can make use of advancing network capabilities by specifying requirements and using a selection process to choose the best available communication service.

Describing application requirements and communication services supports the parallel development of both applications and communication services, which leads to the evolution of the Internet. As soon as new protocols or networks emerge that fulfill the application requirements, they can be automatically selected by using the service selection process.

8. APPENDIX 1: BEST SERVICE SELECTION: AN EXAMPLE

The goal is to select the best service among the three services: S1, S2 and S3. For achieving this goal using our approach, we choose three selection criteria: Delay, Throughput and Jitter and pairwise-prioritized them as shown in Table 1. As it is seen in the table, delay is given strongly more important than (5) throughput and absolutely more important (9) than Jitter. To make the matrix consistent, throughput and jitter are assigned strongly less important than (0.2) and absolutely less important than Delay (0.11) respectively.

Table 2. Measured/Estimated values of Services

Services	Delay (ms)	Throughput (Mbps)	Jitter (ms)
S1	10	1	1
S2	50	2	2
S3	250	10	10

Table 3. Overall priority vector computation

Req. vector	0.7651	0.1288	0.1062	Priority
	Delay	Throughput	Jitter	
S1	0.5869	0.0740	0.5790	0.52
S2	0.3583	0.1176	0.3685	0.33
S3	0.0549	0.8084	0.0524	0.15

Assuming that the services S1, S2 and S3 offer the values of Delay, Throughput and Jitter according to table 2.

These values are mapped to the scale of (1, 9) using the the mapping algorithm depicted in figure 4. The requirement matrix is consistent as its consistency ratio is less than 10%. The overall priority is then obtained by multiplying the priority vector of the requirement matrix with the offered matrix. The service with the highest value in the overall priority vector is chosen as the best service, which is S1, as shown in table 3.

9. ACKNOWLEDGMENTS

We acknowledge our colleagues for their valuable suggestions. Thanks to Prof.Thomas L. Saaty for providing helpful feedback regarding AHP.

REFERENCES

[1] "ANA," http://www.ana-project.org/, Online; accessed 25-April-2011.

[2] L. Voelker, D. Martin, I. E. Khayat, C. Werle, and M. Zitterbart, "A node architecture for 1000 future networks," in *International Workshop on the Network of the Future 2009, Dresden, Germany*, 2009.

[3] "XIA," http://www.cs.cmu.edu/ xia/technical/technical-approach.html, Online; accessed 3-Feb-2012.

[4] Paul Müller and Bernd Reuther, "Future internet architecture - a service oriented approach (future internet architecture - ein serviceorientierter ansatz)," *it - Information Technology*, vol. 50, no. 6, pp. 383–389, 2008.

[5] "RINA," http://csr.bu.edu/rina/, Online; accessed 03-February-2012.

[6] Andrea de Montis, Pasquale De Toro, Bert Droste-Franke, Ines Omann, and Sigrid Stagl, "Assessing the quality of different mcda methods," 2000.

[7] Rahamatullah Khondoker, "Implementation and evaluation of regular channel access for cooperative spectrum sharing between coexisting wlan systems," *Master Thesis*, 2009.

[8] Rahamatullah Khondoker, Eric MSP Veith, and Paul Mueller, "A description language for communication services of future network architectures," *In Proceedings of the 2011 International Conference on the Network of the Future*, pp. 69 – 76, 2011.

[9] Rahamatullah Khondoker, Bernd Reuther, Dennis Schwerdel, Abbas Siddiqui, and Paul Mueller, "Describing and selecting communication services in a service oriented network architecture," in *the proceedings of the 2011 ITU-T Kleidoscope event, Beyond the Internet? Innovations for future networks and services, Pune, India*, December 2010.

[10] T. L. Saaty, "The analytic hierarchy process," *McGraw-Hill, New York*, 1980.

[11] B. Roy, "Multiple criteria methodology for decision aiding," *Parigi: Economica*, 1985.

[12] H. Voogd, "Multi-criteria analysis with mixed qualitative-quantitative data," *Delft University of Technology, Department of Urban and Regional Planning*, 1981.

[13] R. L. Kenny and H. Raiffa, "Decisions with mulitple objectives: Preferences and value trade-offs," *John Wiley and Sons, New York*, 1976.

[14] Matthias Ehrgott and Xavier Gandibleux, "Multiple Criteria Optimization: State of the Art Annotated Bibliographic Surveys," *Kluwer Academic Publishers*, 2003.

[15] G. Munda, "Multi Criteria Evaluation in a Fuzzy Environment - Theory and Applications in Ecological Economics," *Hidelberg: Physika Verlag*, 1995.

[16] E. Hinloopen, "De regime methode, ma thesis," *Interfaculty Actuariat and Econometrics, Free University Amsterdam*, 1985.

[17] Thomas L. Saaty, "Decision making with the analytic hierarchy process," *Int. J. Services Sciences*, vol. 1, no. 1, pp. 83–98, 2008.

[18] Florian Liers, Thomas Volkert, Denis Martin, Helge Backhaus, Hans Wippel, Eric MSP Veith, Abbas Ali Siddiqui, and Rahamatullah Khondoker, "GAPI: A G-Lab Application-to-Network Interface," *In Proceedings of the 11th Würzburg Workshop on IP: Joint ITG and Euro-NF Workshop on "Visions of Future Generation Networks" (EuroView 2011)*, 2011.

[19] M. Vogt, Th. Plagemann, B. Plattner, and Th. Walter, "Eine laufzeitumgebung fuer da capo," *GI/ITG-Arbeitstreffen Verteilte Multimedia-Systeme*, 1993.

[20] B. Stiller, "FuKSS: Ein funktionsbasiertes kommunikationssubsystem zur flexiblen konfiguration von kommunikationsprotokollen," *GI/ITG-Fachgruppe Kommunikation und Verteilte Systeme*, 1994.

[21] Birgit Geppert and Frank Roessler, "Generic Engineering of Communication Protocols - Current Experience and Future Issues," in *ICFEM '97: Proceedings of the 1st International Conference on Formal Engineering Methods*, Washington, DC, USA, 1997, p. 70, IEEE Computer Society.

[22] M. Vogt, Th. Plagemann, B. Plattner, and Th. Walter, "A Run-time Environment for Da CaPo," in *Proceedings of INET93 International Networking Conference of the Internet Society*, 1993.

[23] Robert Braden, Ted Faber, and Mark Handley, "From Protocol Stack to Protocol Heap: Role-Based Architecture," *SIGCOMM Computer Communication Review*, vol. 33, no. 1, pp. 17–22, 2003.

[24] R. Dutta, G.N. Rouskas, I. Baldine, A. Bragg, and D. Stevenson, "The SILO Architecture for Services Integration, controL, and Optimization for the Future Internet," in *Communications, 2007. ICC '07. IEEE International Conference on*, June 2007, pp. 1899–1904.

[25] Suman Ramkumar Srinivasan, Jae Woo Lee, Eric Liu, Michael Kester, Henning Schulzrinne, Volker Hilt, Srini Seetharaman, and Ashiq Khan, "NetServ: Dynamically Deploying In-Network Services," in *ReArch '09: Proceedings of the 2009 workshop on Re-architecting the internet, Rome, Italy*, 2009.

[26] Florian Liers, Thomas Volkert, and Andreas Mitschele-Thiel, "Forwarding on Gates: A clean-slate Future Internet Approach within the G-Lab project," in *EuroView 2009*, Würzburg, Germany, 2009.

[27] "4WARD EU Project," http://www.4ward-project.eu/, Online; accessed 25-April-2011.

[28] "Selfnet project," http://www.ict-selfnet.eu/, Online; accessed 10-July-2007.

[29] Joseph D. Touch, Yu-Shun Wang, and Venkata Pingali, "A Recursive Network Architecture," http://www.isi.edu/touch/pubs/isi-tr-2006-626/, 2006, Online, accessed 16-Nov-2012.

[30] Christian Henke, Abbas Siddiqui, and Rahamatullah Khondoker, "Network functional composition: State of the art," *2010 Australasian Telecommunication Networks and Application Conference, Auckland, New Zealand*, pp. 43 – 48, 2010.

[31] Lars Voelker, Denis Martin, Christoph Werle, Martina Zitterbart, and Ibtissam El Khayat, "Selecting concurrent network architectures at runtime," in *IEEE International Conference on Communications (ICC 2009)*, June 2009.

USING THE RFID TECHNOLOGY TO CREATE A LOW-COST COMMUNICATION CHANNEL FOR DATA EXCHANGE

Ivan Farris, Antonio Iera, and Silverio C. Spinella

A.R.T.S. Lab, University of Reggio Calabria, Italy
Emails: ivan.farris.524@studenti.unirc.it, antonio.iera@unirc.it, silverio.spinella@unirc.it

ABSTRACT

This paper proposes a methodology to use the RFID technology (more specifically the RFID tags) as a novel "communication channel", to support data exchanges in high pervasive environments, analogously to more traditional short-range communication technologies (WiFi, ZigBee, Bluetooth). To this aim, the further research issue of creating so called RANs (RFID-Area Networks - in analogy with LANs, Local Area Networks, PANs, Personal Area Networks, etc.) is addressed. These are made up of groups of RFID readers into which the functionality for exchanging data over the introduced "RFID virtual channel" within the generic RAN, in either a broadcast or a unicast modality, is embedded. From initial studies on its functional behavior, it emerges that the proposed method may actually allow to exploit a further (currently largely wasted although available "at no cost") channel in future scenarios populated by tagged everyday-life objects.

Keywords— RFID, Communication Channel, RFID Area Network, Pervasive, IoT

1. INTRODUCTION

The ubiquitous (or pervasive) computing vision, which will characterize the human future way of life, is achieved only through the evolution of several technologies, systems, and networks to be used in synergy. Sensors and actuators, M2M and embedded systems, as well as any kind of ubiquitous wireless communication and networking solution are key enablers of this vision.

The RFID technology certainly is part of this evolutionary process, with a leading role. This is due not only to its traditional role of RF technology aimed at tracing goods and people but also to "unconventional" uses that may result from it (e.g., indoor localization [1][2], environmental sensing [3], energy savings [4], etc). The great advantage that the cited technology will have in the future is undoubtedly its extreme pervasiveness in different kinds of environments, ranging from home to work/industrial environments. Not by chance RFIDs were the objects that gave rise to the concept of Internet of Things [5][6], later extended to different smart objects.

Starting from this, much research is focusing on the development of different solutions for "smart" and "sustainable" communities that may benefit from the pervasive presence of *RFID ecosystems* in everyday life. The idea (in a sense, visionary) that we propose in this paper starts from two assumptions:

- there cannot be real RFID ecosystems until the technology, thought to tracking and trace goods and persons, is not used also as a communication means to exchange data of different types (other than the mere identification of an object);
- at present, the RFID tags associated with objects, and those of new generations currently under study by manufacturers and standardization bodies, appear certainly oversized, in terms of memory resources, for mere traceability and identification purposes. Thus, the unused resources must be exploited in the view of the future sustainable technological ecosystems.

In light of the above, we think that the future massive presence of RFID tags could represent a further chance to exploit a new "virtual" communication channel represented by the RFID tags themselves. This channel is already available but not fully exploited yet. What we are thinking about is to find a way to enable low-cost mobile RFID readers to use the residual memory of the RFID tags in the environment to exchange data of a different nature.

This implies to perform a feasibility study on an unconventional usage of standard RFID readers and tags and to come to the definition of a novel communication paradigm enabling the constitution of piconets (called RFID Area Networks, RANs) of devices exchanging data over a virtual RFID channel.

As it will be better discussed in the following, the applications of such a paradigm are manifold. In RFID ecosystems, populated by sensorized RFID tags, the sensed data could be propagated from reader to reader without resorting to the use of different communication technologies. Use-cases, such as home automation for a better life, Machine to Machine for sustainable industrial automation, eHealth applications for pervasive and constant monitoring of the people health, would benefit from this low cost and currently unused additional channel.

At this early stage, we will describe a possible communication paradigm, describe how to implement it, and evaluate its feasibility by also presenting some early estimations of the achievable theoretical performance.

2. BACKGROUND

The proposal described in this paper relies on the massive presence of the RFID technology in the future everyday life. Several studies confirm that this is not just a guess. For example, in [7] it is predicted that by the next few years hundreds of billions of RFID-tagged objects will be available at approximately five cents per tag. Furthermore, a numerous family of novel devices consisting in sensorized RFID platforms are gaining ground in the market. The most relevant example of small devices fully compliant to the EPC standard but smarter than mere RFID tags are the WISPs [3] (Wireless Identification and Sensing Platform), which are sensing and computing devices that are powered and read by off the shelf UHF RFID readers.

The use of the RFID technology in pervasive environments for pervasive computing applications has been widely addressed in the literature. A relevant example is the RFID Ecosystem envisaged in [8], which creates a microcosm for the Internet of Things. In this study, the authors highlight that the incredible amount of information captured by a trillion RFID tags will have a tremendous impact on our lives. With this in mind, their attention is mainly on applications strongly relying on the data captured by RFID systems. We think that some of the interesting envisaged applications could benefit from the paradigm we propose in this paper. As an example, it could be effectively used in RFID ecosystems by low-cost RFID mobile readers to gather information from the surrounding environment and exchange it.

Several networked RFID infrastructures have also been deployed. Undoubtedly, the EPCglobal standard is, currently, the most relevant. It proposes a global system for EPC compliant RFID devices and tags targeted to the world-wide traceability of goods, without neglecting the future integration with sensors and actuators [9] in the view of the Internet of Things. EPCglobal addresses functions at different protocol levels, although it does not analyze the possibility of exchanging data through a virtual RFID transmission medium, like our proposal does. Logically, our choice is to maintain a *full compatibility with the EPC standard* and thus being complementary to the EPCglobal platforms.

3. REFERENCE SCENARIOS AND APPLICATIONS

Before proceeding to the description of how a RAN can be built by relying on the currently available RFID standard technology, let us briefly investigate case studies where applications, services, and systems could benefit from the exploitation of a "virtual RFID channel". Most of them emerge in realistic every-day-life scenarios and definitely may contribute to deploy the concept of "enhancing the quality of human life". In the following, we just give a list of possible application without being exhaustive.

A first category of applications to consider *is positioning and location tracking through the RFID technology*. Among others, several solutions have been proposed in the literature, which are "reader oriented". These techniques aim at locating mobile readers, which interact with RFID tags scattered in the surrounding environment. What we propose can enable the exchange of information among close RFID mobile readers through the virtual RFID channel to the purpose of cooperating and thus attaining a more precise positioning. In other words, we are referring to a cooperative RFID location mechanism entirely implemented by a single technology and allowing for using simpler and cheaper (although more precise) mobile RFID readers for location. Interesting use-cases can be envisaged (i) either in the area of robot swarms (i.e. group of robots that use RFID positioning [10] and that may enhance their relevant positioning by cooperating with other robots) used for safety and security applications, as well as for disaster recovery applications or (ii) in the case in which less complex and low-cost mobile RFID readers (used for example to continuously and precisely trace the movements of elderly people or patients at home) receive position enhancement information from more sophisticated RFID readers in the same area and equipped with more effective positioning techniques (GPS, WLAN based, etc.).

Cooperative data exchange among cellular terminals is also a possible application area. This paradigm (data downloaded from the cellular network and shared over the short links) has been widely investigated in the literature [11]. The RFID channel can be a costless channel, additional to Bluetooth, to exchange low-bit rate data (such as signaling information, etc.).

Distributed sensing is another interesting application to consider. One can think of environments in which sensorized RFID tags are scattered (such as WISPs) and coexist with the standard RFID tags associated to every-day-life objects. If low-cost miniaturized RFID (single technology) readers are also distributed in the environment or embedded in objects, then they can gather information from the sensorized RFID tags and exchange it (in a multi-hop fashion) to reach a high complexity RFID reader (with communication capabilities). This latter may act as a "sink" and as an "anchor" to the Internet, which receives sensing information and relays it towards the external world. Again, very interesting is the use of the single RFID technology. Different use-cases can be envisaged both in the area of the home automation and in the area of assisted living, telemedicine, and eHealth (by using RFID sensors like in [12], for example). Further interesting applications are easily conceivable in the area of the energetic consumption control.

Distributed search engines for things in the environment can also be implemented by utilizing a few fixed multi-technology RFID readers and several low cost, mobile, single-technology (RFID only) readers distributed in the environment.

4. THE PROPOSED PARADIGM

The reference scenario for our proposal is sketched in figure 1. A *Master Reader* (*MR*) creates and coordinates one or more *RFID Area Networks* (*RANs*), with a multiplicity of RFID *Client Readers* (*CR*) and tags associated; each RAN is conceptually equivalent to a Bluetooth Wireless Personal Area Network (WPAN) or a

IEEE 802.11 Wireless Local Area Network (WLAN). Client Reader and RFID tag can be associated at the same time with different RANs, either created by the same Master Reader or by different Master Readers. The data exchange between Master Reader and Client Readers belonging to the same RAN is performed through the use of *passive* RFID tags (already present in the environment for tracing purposes) of type EPCglobal Class1 Gen2 equipped with the User Memory Bank.

The MRs are associated to an additional device, the *Master Control* (*MC*), to which they communicate through standard communication technologies (IEEE 802.11, IEEE 802.3, Bluetooth, etc.). Task of the MC is just to release to each MR a univocal 5 bit address, the ID-Master, to the purpose of optimizing the resource usage and the procedures of data memorization into the RFID tags. The MC implements a NAT (Network Address Translation) protocol to map IP addresses onto ID-Master addresses.

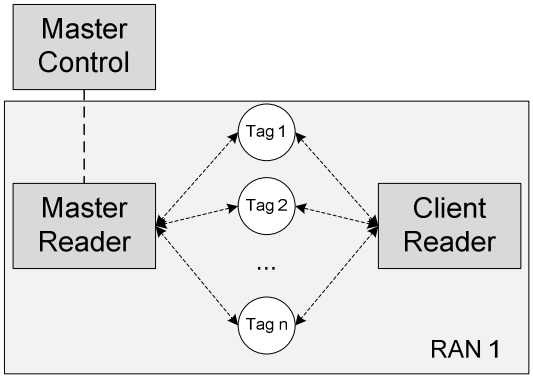


Figure 1. Basic reference scenario

Suitable Discovery Tables are implemented in both the MR and CR to dynamically identify the RFID tags that each MR can utilize as a communication channel to transfer data towards a CR and vice-versa. Logically, to implement the whole set of procedures that our proposal foresees, the RFID *User Memory Bank* must be properly structured. In compliance to the ISO/IEC 15962 standard, which already rules the coding and decoding procedure for the User Memory Bank, a novel *Application Family Identifier* (*AFI*) - described as "Communication Channel", along the lines of ISO/IEC 15961 – and a novel organization of the user memory are proposed.

Through these novel data structures, the MRs and the CRs implement the following functioning phases: *(i) Addressing Phase*, *(ii) Communication Phase* and *(iii) Control Phase*. These will be better addressed in the following subsections.

4.1 Proposed RFID tag memory structure

Before describing the system behavior in any phase, it is worth giving some details on the introduced AFI, the proposed tag memory organization, and the data structure of the Discovery Table. Anything proposed is completely compliant to the EPCglobal Class 1 Gen 2 standard.

EPC Bank

The following values of this memory block are considered: *UMI*, *Toggle* and *AFI*. The UMI (User Memory Indicator) is a bit that establish whether the User Bank is present or not. Specifically, if the bit is "0" and the *Data Format* field of the User Bank is "00000", then the User Bank is present but unused; bit equal to "0" and Data Format not present mean that also the User Bank is not present; if the bit is set to "1", then the User Bank is present and initialized. The Toggle field, 1 bit, distinguish the application type: EPCglobal (bit "0") or ISO (bit "1"). If the Toggle bit is set to "1", then the EPC Bank has an AFI defined according to ISO 15961. The method proposed in this paper requires the *definition of a new AFI*, that we assume to label "*Communication Channel*".

TID Bank

For this memory block the value *User Memory Size* (UMS) is considered, i.e. the number of 16-bit words present in the User Bank of a specific tag.

Table 1. Structure of the UMR

AI	Name	Data Title	Format Binary	#CPO	
0	Cluster Map	CM	1*n bit		
1	Priority Level	PL	2*n bit		
2	ID Reader Source	IDRS	5 bit		
3	ID Reader Destination	IDRD	5 bit		
4	ID RAN	IDRAN	2 bit		
5	Count Success	CS	4 bit		
6	Count Insuccess	CI	4 bit		
7	Sequence Number	SN	5 bit		
8	Reader Address Lease	RAL	5 bit		CPO
9	ID Master	IDM	5 bit		
10	ID RAN Lease	IDRANL	2 bit		
11	Reservation Bits	RB	2 bit		
12	Check Bits	CB	2 bit		
13	Payload	PLD	76 bit	1	
14	ID Reader Source	IDRS	5 bit		
15	ID Reader Destination	IDRD	5 bit		
16	ID RAN	IDRAN	2 bit		
17	Count Success	CS	4 bit		
18	Count Insuccess	CI	4 bit		
19	Sequence Number	SN	5 bit		
20	Reader Address Lease	RAL	5 bit		
21	ID Master	IDM	5 bit		
22	ID RAN Lease	IDRANL	2 bit		
23	Reservation Bits	RB	2 bit		
24	Check Bits	CB	2 bit		
25	Payload	PLD	76 bit		
26	Bit unused		8 bit	2	
...	...	...	...		
	DATA PACKET		CONTROL PACKET		

User Bank

This memory block has two main sections: *(1)* the field *Data Storage Format Identifier* (*DSFID*), of 8 bits and *(2)* the remaining portion of User Bank, hereinafter defined User Memory Remaining (*UMR*), of a variable dimension. According to the EPCglobal Gen2 standard, the DSFID defines both the UMR data format and the access method (*no-directory, directory, packed object)*. We consider the *directory* access method and a novel data format associated to the newly defined *Communication Channel* AFI (see Tab. 1). In Table 1 it can be observed that a subset of AIs (Application Identifiers), defined Cluster Packed Object (CPO) repeats itself. The number of CPOs is equal to *n* (this number depends on the UMS

value in the TID Bank). A CPO can be, thus, considered like a 117 bit frame within a multi-frame structure.

Table 1 also shows that for each CPO a *Control Packet* and a *Data Packet* are defined. The fields of the *Control Packet* are used for the Addressing and the Control Phase: *(1) RAL*, address of a CR sequentially released by a MR for a specific RAN, *(2) IDM*, identifier of the MR released by a MC, *(3) IDRANL*, identifier of the RAN, *(4) RB*, a two-bit couple used in the Addressing Phase, *(5) CB*, a two-bit couple used in the Control Phase. The *Data Packet* fields, instead, are used only in the Communication Phase: *(6) IDRS*, identifier of the source reader, *(7) IDRD*, identifier of the destination reader, *(8) IDRAN*, identifier of the RAN, *(9) CS*, counter of successful read commands, *(10) CI*, counter of failed read commands, *(11) SN*, data packet sequence number, *(12) Payload*, i.e. data exchanged between MR and CR. Only two AIs in Tab. 1 have a variable dimension and refer to all CPOs: *Cluster Map* (*CM*) and *Priority Level* (*PL*). The former identifies which CPO are available and which occupied, while the latter specifies the priority associated to each CPO ("00" low priority, "01" medium priority, and "10" high priority).

Reserved Bank

The value, 32 bits long, considered for this block of memory is *Access Password*. Through it, it is possible to understand if the memory blocks of the RFID tag are write locked, i.e. whether they can be used or not by the new AFI named *Communication Channel*.

Discovery Table

This is a new data structure stored in any reader to the purpose of keeping trace of and updating all the possible paths (in terms of used tags) from an MR to each CR and vice-versa. Its fields are RAL, IDM, IDRANL, RB, CB, and *ID-Tag*. This latter is the identifier of the RFID tag, obtained from the EPC Bank.

4.2 Addressing Phase

During this phase, the MR initializes the RFID tags (in case these are not been previously initialized). During this phase, in fact, the MR and CR Discovery Tables are "populated" to the purpose of defining for each MR one or more RANs and, for each RAN more data paths between the same MR-CR couple (this means using different RFID tags to exchange data between the same device couple). Only a subset of AIs of the CPO are taken into account for the RAN initialization and for the address handling within each RAN. These are: RAL, IDM, IDRANL, RB.

Once the MR has received the unique *ID Master* address from the closest MC, it queries all the RFID tags within its coverage range. For each selected tag, the fields *UMI*, *Toggle,* and *AFI* of the EPC Bank and the field *DSFID* of the User Bank are checked to understand: *(a)* if the UMR structure is already initialized and thus ready to be used as a communication channel; *(b)* if the tag is initialized for the first time by an MR; *(c)* if the tag is suitable to be used as a communication channel; *(d)* if the tag already initialized and used for other AFIs can be re-utilized by an MR as a communication channel.

If the selected tag is already suitable to be used with the *Communication Channel* AFI, then the MR access the UMS of the TID Bank to understand how many *n* potential CPOs can be used (and their associated AIs) based on the UMR illustrated in Tab. 1. The next step is the initialization of the AIs to use. The Application Identifiers *RAL*, *IDM*, *IDRANL* and *RB* will assume suitable values during the process of management of the MR and CR Discovery Tables and of establishment of the possible paths (i.e. tags) to use for the data exchanges.

4.3 Communication Phase

During this phase operations of *write* and *read* to and from the involved tags, finalized to MR and CR data exchanges, occurs. Both *unicast* (i.e., an exchange of messages between an MR and a single CR) and *broadcast* (i.e., the exchange of messages sent by the MR to all the CRs listening to the same tag) communications are foreseen. In the *unicast* case, the involved CR, for each received message will also release the memory resources engaged on the tag (this also works as a kind of acknowledgment mechanism). As it is likely that the exchange of a message will involve more tags, then the CR may be required to release resources on several tags.

In the *broadcast* case, it is not possible to release the engaged resources following a read operation by the CR because the CR does not know which CR has already read the data on the tag/tags (broadcast transmission without acknowledgment). To this aim, a procedure, the *Memory Resource Management* (MRM), is implemented that through the use of suitable counters will take care of the release of the communication resources (i.e. memory on tags). A further operation handled by this procedure is the *priority* handling, in case a MR has the necessity of writing an urgent information onto currently saturated tags.

4.4 Control Phase

During this phase, the Discovery Tables are updated. It can be started both by either an MR or a CR to check for the presence of a given MR/CR/Tag within a RAN during the time. Therefore, three cases can be considered: *(i)* control phase verified by the MR, *(ii)* control phase verified by the CR and *(iii)* control phase with tag verification.

(i) Control Phase verified by the MR

An MR that wants to refresh one entry in its Discovery Table sends the AI triplet *RAL – IDM - IDRANL* with *CB* asset to "01". A CR, which recognizes the same triplet in its Discovery Table, modifies the *CB* to "00" on the tag to communicate to the MR that it is still present in the RAN. If the MR verifies that the CB remains equal to "01" for a given threshold time interval, then it may decide to de-

allocate all the entries relevant to a given CR in its Discovery Table.

(ii) Control Phase verified by the CR

If a CR wants to refresh one entry in its Discovery Table then sends the AI triplet *RAL – IDM - IDRANL* with *CB* set to "10". An MR that recognizes the same triplet in its Discovery Table set the CB to "00" on the tag to confirm its presence to the CR. Oppositely, if the CR verifies that the CB is still set to "10" for a given threshold time, then it may decide to de-allocate from its Discovery Table all the entries relevant to a given MR .

(iii) Control Phase with tag verification

When an MR and/or a CR verifies that a tag is no more within its operational range, then it deletes all the entries relevant to that tag from its Discovery Table. It has been defined a minimum number of queries (Thr_{number}) that an MR and/or a CR has to consider before updating its Discovery Table.

4.5 Memory Resource Management

The Memory Resource Management algorithm, takes care of handling the UMR to optimize the usage of the memory capacity of the tags and, at the same time, to guarantee fairness in the sharing of the tags among readers. It is used both in the Communication and Control phases, and, specifically, it: *(1)* controls the PL of each data message to allow the re-usage of CPOs; *(2)* checks the CS and CI fields, specifically in the condition $(CS - CI) > 0$, in case of broadcast messages to manage the release of useless (because already accessed by all CRs) resources on a tag; *(3)* associates a validity time interval to the entries of the Discovery Tables; *(4)* monitors the minimum number of queries, Thr_{number}.

5. PERFORMANCE EVALUATION

The EPCglobal Gen2 standard [13] regulates the interaction between interrogator and tags through three procedures. Each of these procedures comprises one or more commands:

- *Select.* It consists of a single command, which is used to select a subset of a tag population through a bit mask.
- *Inventory.* This operation univocally identifies a single tag (*singulation*) among the selected population and establishes the subsequent access to it. The inventory command set includes *Query*, *QueryAdjust*, *QueryRep*, *ACK*, and *NAK*. *Query* initiates an inventory round and decides which tags participate in the round (Q-protocol). *Query* contains a slot-count parameter *Q* with range (0,15). Upon receiving a *Query,* the participating tags pick a random value in the range [0, 2^Q - 1] and load this value into their slot counters. The tags that pick a zero value reply immediately a 16 bit random number (RN16). The tags that pick a nonzero value don't reply and await a *QueryAdjust* (to increase or decrease the Q value by one) or a *QueryRep* command (to repeat a previous query without changing any parameters).
- *Access.* During the access process, an interrogator may choose to access a individual tag. The access command set comprises *Req_RN, Read, Write*, *Kill*, *Lock*, *Access*, *BlockWrite*, *BlockErase*, and *BlockPermalock*.

When the Inventory phase is successfully concluded, the Read and Write command can be sent. Through a *Read operation* it is possible to read up to 256 16-bit words (512 byte) at a time. Through a *Write operation*, instead, it is possible to write a single 16-bit word at a time. The standard foresees an optional command, *BlockWrite*, which allows to write up to 256 16-bit words at a time. More specifically, most of the readers available from the market allow to write up to a 4 16-bit words.

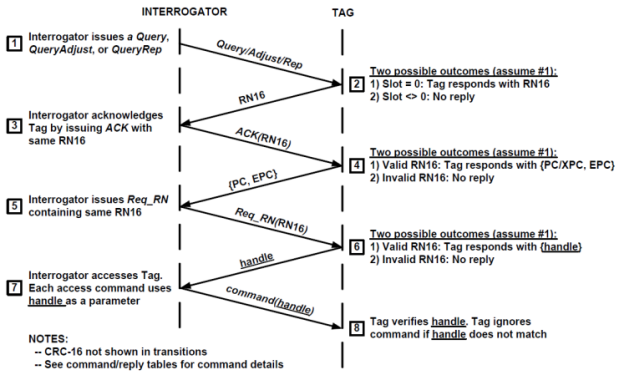


Figure 2. Inventory and access of a single tag (Courtesy of EPCglobal)

One has to notice that the EPCglobal Gen2 standard foresees a temporal windows of a maximum of 20 ms, following any Write (or BlockWrite) operation, during which the reader energizes by CW (Continuous Wave)the tag to enable the writing onto its EEPROM. This time interval is tightly related to the type of tag and, in general, for the EEPROM memories, it ranges from 4 ms to 10 ms. Anyway, if no reply is obtained from the tag in 20 ms, the reader considers the command as *failed*. Actually, new type of tags with a memory based on F-RAM (Ferroelectric Random Access Memory) technology instead of EEPROM (such as the MaxArias products declared by the Ramtron [14]) will allow to reduce the time intervals required to write a tag memory and to increase the rewrite capability (10^{14} vs. 10^6).

To obtain *upper bound* values for the Bit Rate and the Data Rate metrics, the time interval required to successfully transmit a message that occupies the whole User Bank under some ideal assumptions is computed. In an early analysis (details not given for length constraints), the QueryRep and QueryAdjust times are not considered and the only time considered is the one required to obtain the EPC code of a single tag. In other words, we consider only the "Single Tag Reply" case, while leaving the "Collided Reply" and "No Reply" cases reported in [13] for future

researches. Our current objective is, in fact, just to give an initial idea on the achievable performance levels.

In our performance evaluation study, the main parameters of the Gen2 standard are set to the values listed in Table 2.

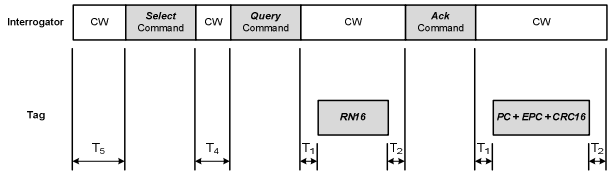


Figure 3. Interrogator – Tag communication timing diagram

We also assume that a Master Reader sends messages to a single Client Reader by using the whole User Memory of a single tag and assume to be in conditions of perfect synchronism (ideal conditions to compute a theoretical upper bound for the rate related metrics). Because the readers alternatively operate on the radio channel, thus in absence of interference and collisions, the assumed encoding can be FM0, less robust compared to the Miller encoding but with a higher Tag-to-Reader bit rate.

Table 2. Parameter Setting

Parameters	
Tari	12.5 µs
Transm. Rate R->T (Reader-to-Tag)	80 Kbps
Transm. Rate T->R (Tag-to-Reader)	160 Kbps
Preamble	
Interrogator-to-Tag Preamble	112.5 µs
Interrogator-to-Tag Frame-Sync	62.5 µs
Tag-to-Interrogator Preamble	112.5 µs
Reader Command	
Select	912.5 µs
Query	525 µs
ACK	400 µs
Req_RN	812.5 µs
Read	1150 µs
Write	1300 µs
Tag Response	
RN16	218.75 µs
PC+EPC+CRC16	918.75 µs
Req_RN Reply	318.75 µs
Timing Requirements	
T_1	62.5 µs
T_2	62.5 µs
T_3	62.5 µs
T_4	112.5 µs
T_5	1500 µs

By considering Figure 3, the values in Table 2 and the illustrated assumptions, we can estimate the time intervals to implement all the different commands and procedures illustrated in the initial part of this Section. From these intervals we are able to evaluate the two searched parameters: *Bit Rate [bit/s]* and *Data Rate [bit/s]* that is theoretically possible to achieve when two readers within a RAN exchange data over a single tag.

The performance in terms of Bit Rate and Data Rate obviously depend on the quality of the tag and of the communication channel. Among the cited operations on the tag, the write commands are the most energy and time consuming. Therefore, the highest impact on the conducted analysis is given by two parameters related to this command: T_{Write} and MaxWordBlockWrite (i.e., T_{Write}: time a tag takes to write a 16 bit-word; MaxWordBlockWrite: the number of words that the reader is able to write in a single BlockWrite command). Their values are strongly related to the class of devices (ranging from basic to high performing) available on the market.

Let us focus on 3 RFID tag classes: *(i) F-RAM Tag*, high innovative and characterized by $T_{Write} \sim 0$ ms and MaxWordBlockWrite = 128 [16-bit word], *(ii) EEPROM Tag*, products of commercial strip, characterized by T_{Write} = [4÷10] ms and MaxWordBlockWrite = 4 [16-bit word], *(iii) Standard Tag*, by this meaning virtual tag characterized by the maximum parameter values T_{Write} = [20] ms and MaxWordBlockWrite = 256 [16-bit word] defined by the EPCglobal Gen2 standard.

Figure 4 and 5 show the Bit Rate and Data Rate observed values vs. the User Memory size for a variable T_{Write} value (T_{Write} = 0,4,10,20 ms), in a reference scenario characterized by a single MR, a single CR, and a single RFID tag. The write method used is the "Write Command", conceptually equivalent to a "BlockWrite Command" with MaxWordBlockWrite = 1.

The sketched curves show that the performance level is inversely proportional to the T_{Write} value and directly proportional to the User Memory amount. Currently, the most widely diffused User Memory capacity for commercial tag is 512 bits. Consequently, the maximum values of Bit Rate and Data Rate obtained when a F-RAM Tag is used are about 2.3 Kbit/s and 1.35 Kbit/s, respectively. When considering an EEPROM Tag, with T_{Write} = 10 ms and for the same value of User Memory, a Bit Rate of about 590 bit/s and a Data Rate of about 350 bit/s are obtained (i.e., values four times lower than in the case of F-RAM Tags). One can also note that there are values of User Memory beyond which no significant increases of Bit Rate and Data Rate are observed. This suggests to avoid the use of tags with memory capacity beyond a certain limit. As an example, in case of F-RAM Tags, it is advised to consider a User Memory up to 2048 bits, while maximum 1024 bits are advised in case of EEPROM Tag with T_{Write} = 4 ms and 512 bits in case of EEPROM Tag with T_{Write} = 10 ms.

Figure 6 and 7, show Bit and Data rates when varying the User Memory for values of MaxWordBlockWrite = 4, 128, 256 (16–bit words). In the figures, the write method used is "BlockWrite Command" and the $T_{BlockWrite}$ is assumed equal to 20 ms (i.e. the maximum value foreseen by the EPCglobal Gen2 standard).

The resulting performance levels are directly proportional to the values of both MaxWordBlockWrite and User Memory. Potentially, the maximum performance is

achieved of course for the case of Standard Tag. For the latter it is assumed the possibility to write 256 16-bit words in a single solution; this results in a bit rate of about 15 Kbit/s and a Data Rate of about 9 Kbit/s. Again, by considering that 512 bit is the most widely diffused User Memory capacity in commercial tags, the maximum Bit Rate and Data Rate values of about 5.3 Kbit/s and 3.2 Kbit/s are obtained by also using F-RAM Tags. If the User Memory is increased to 1024 bits, then Bit Rate and Data Rate will be about 8.4 Kbit/s and 5 Kbit/s respectively. By using EEPROM Tags, a User Memory of 512 bits would imply a Bit Rate of about 1.25 Kbit/s and a Data Rate of about 740 bit/s.

As a last remark, please note that the performance in case of Standard Tags and F-RAM Tags significantly increase by increasing the User Memory, while the same does not happen in case of EEPROM Tags. This is mainly motivated by the relevant weight of the fixed $T_{BlockWrite}$ with respect to the increasing tag capacity.

6. INPUTS TO STANDARDIZATION ACTIVITIES

The paradigm of "communications through RFID technology", illustrated in this paper, intrinsically provides inputs to the activities of different standardization organizations. First, for its deployment, it is important to involve the ISO (*International Organization for Standardization*) and IEC (*International Electrotechnical Commission*) organizations; in fact, the specification of the AFI codes (that give information on the type of application which the tag is destined to) is included in the ISO/IEC 15961 standard. As an example, according to this standard, the AFI Code 9 is assigned to the EAN.UCC system (i.e. to the GS1) and the AFI Code 10 is assigned to the ANSI MH10.8.2. Similarly, a new *AFI Code*, among those still available, could be associated to distinguish the proposed *UMR* structure as described in this paper. This action could involve the ITU-T Study Group 17 for the definition of a new *OID* (Object IDentifier). Furthermore, the coding of the User Bank, and, more specifically, of the DSFID, is specified by the ISO/IEC 15962 standard. As a consequence, also the definition of a new *Data Format* to be transposed by GS1 is a compulsory activity to conduct.

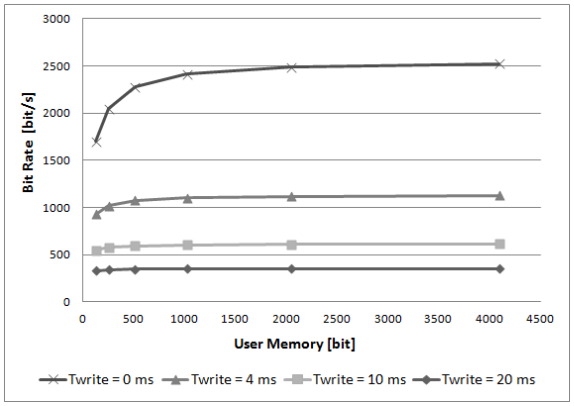


Figure 4. Bit Rate vs. User Memory utilizing "Write Command"

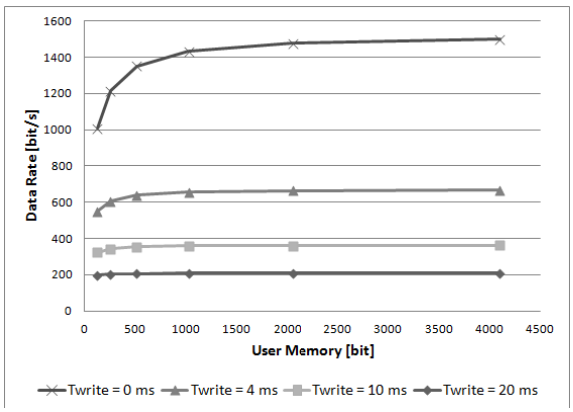


Figure 5. Data Rate vs. User Memory utilizing "Write Command"

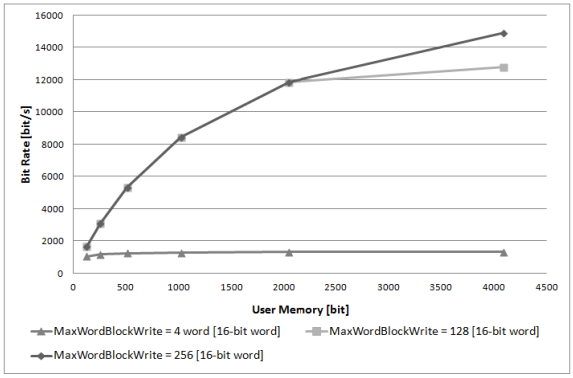


Fig. 6. Bit Rate vs. User Memory utilizing "BlockWrite Command"

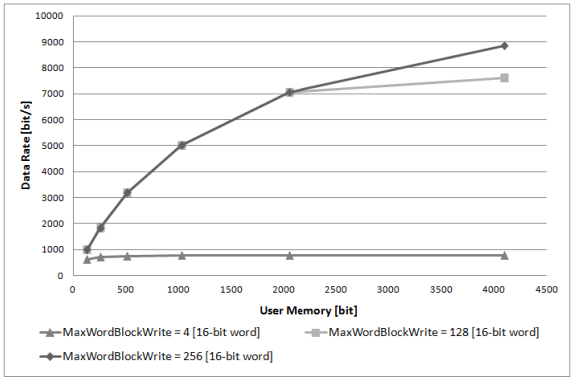


Figure 7. Data Rate vs. User Memory utilizing "BlockWrite Command"

7. CONCLUSIONS

In this paper we investigated the feasibility of a new low cost communication channel which exploits RFID tags of type EPCglobal Class1 Gen2 to support data exchanges in high pervasive environments, analogously to more traditional short-range communication technologies (WiFi, ZigBee, Bluetooth). Compliant to the EPCglobal Class 1 Gen 2 standard and thus being complementary to the EPCglobal platforms we proposed a novel AFI and a novel organization of the user memory. We evaluated two parameters, Bit Rate and Data Rate that is theoretically possible to achieve when two readers within a RAN

exchange data over a single tag considering different quality of the tag as communication channel.

Currently, our research is focusing on the performance assessment when taking into account issues that have been disregarded in the early stage of our study, such as the reader collision and the network reliability issues. In the literature several *centralized*, *distributed,* and *hybrid* mechanisms addressing reader-to-reader and reader-to-tag interference problems are proposed. The centralized mechanisms are not suitable to distributed solutions like the one we propose; therefore, our attention is on distributed mechanisms. Those available from the literature, unfortunately foresee control channels and synchronization that exploit extra hardware, thus contrasting with our first assumption of a pure RFID ecosystem. Hence the need for defining a novel Medium Access Control, which will be our next objective. Last step will be the implementation of a prototype to assess the effectiveness of the proposed solution through real experiments.

ANNEX I. LIST OF ACRONYMS

Acronym	Description	Acronym	Description
RAN	RFID Area Network	IDM	IDentification Master
M2M	Machine to Machine	IDRANL	IDentification RAN Lease
WISP	Wireless Identification and Sensing Platform	RB	Reservation Bits
GPS	Global Positioning System	CB	Check Bits
WLAN	Wireless Local Area Network	IDRS	IDentification Reader Source
MR	Master Reader	IDRD	IDentification Reader Destination
CR	Client Reader	IDRAN	IDentification RAN
WPAN	Wireless Personal Area Network	CS	Count Success
MC	Master Control	CI	Count Insuccess
ID-Master	IDentification Master	SN	Sequence Number
NAT	Network Address Translation	CM	Cluster Map
AFI	Application Family Identifier	PL	Priority Level
UMI	User Memory Indicator	ID-Tag	IDentification Tag
ISO	International Organization for Standardization	MRM	Memory Resource Management
TID	Tag IDentifier	CW	Continuous Wave
UMS	User Memory Size	F-RAM	Ferroelectric Random Access Memory
DSFID	Data Storage Format IDentifier	IEC	International Electrotechnical Commission
UMR	User Memory Remaining	EAN	European Article Numbering
AI	Application Identifier	UCC	Uniform Code Council
CPO	Cluster Packed Object	ANSI	American National Standards Institute
RAL	Reader Address Lease	OID	Object IDentifier

REFERENCES

[1] Silverio C. Spinella, Antonio Iera, and Antonella Molinaro, “On Potentials and Limitations of a Hybrid WLAN-RFID Indoor Positioning Technique”, *International Journal of Navigation and Observation*, Volume 2010 (2010), doi:10.1155/2010/397467.

[2] S Polito, D Biondo, A Iera, M Mattei, A Molinaro, “Performance evaluation of active RFID location systems based on RF power measures”, Proceedings of IEEE Personal, Indoor and Mobile Radio Communications, PIMRC 2007, Greece.

[3] Joshua R. Smith, Alanson Sample, Pauline Powledge, Alexander Mamishev, Sumit Roy, “A wirelessly powered platform for sensing and computation”,. Proceedings of Ubicomp 2006: 8th Int. Conf. on Ubiquitous Computing. Orange Country, US, Sept. 17-21 2006, pp. 495-506.

[4] R. Jurdak, A.G. Ruzzelli, and G.M.P. O'Hare. "Multi-hop RFID Wake-up Radio: Design, Evaluation and Energy Tradeoffs," in Proceedings of the 17TH International ICCCN Conference, August, 2008.

[5] Auto-Id Labs, <http://www.autoidlabs.org/>.

[6] L. Atzori, A. Iera and G. Morabito, ”The Internet of Things: A Survey”, Computer Networks, Vol. 54, No. 15, pp. 2787-2805, Oct. 2010.

[7] Complete RFID Analysis and Forecasts, 1. 2008–2018, www.idtechex.com/research/reports/.

[8] L. Battle, G. Cole, K. Gould, K. Rector, S. Raymer, M. Balazinska, G. Borriello, “Building the Internet of Things Using RFID -The RFID Ecosystem Experience”, *IEEE Internet Computing*, Vol.: 13, Issue: 3 , Pages: 48 – 55, 2009.

[9] Jongwoo Sung, Sanchez Lopez T., Daeyoung Kim, “The EPC Sensor Network for RFID and WSN Integration Infrastructure”, Pervasive Computing and Communications Workshops, 2007. PerCom Workshops '07.

[10] Qing Yang et al., “An Improved Method for Mobile Robot Localization Based on Passive RFID System”, 2012, *Applied Mechanics and Materials*, 190-191, 651.

[11] A. Iera, L. Militano, L.P. Romeo, F. Scarcello, “Fair Cost Allocation in Cellular-Bluetooth Cooperation Scenarios”, *IEEE Transactions on Wireless Communications*, Vol 10, Issue: 8, pp 2566 – 2576.

[12] Vijey Thayananthan, Ahmed Alzahrani, “RFID-based Body Sensors for e-Health Systems and Communications”, eTELEMED 2012 : The Fourth International Conference on eHealth, Telemedicine, and Social Medicine.

[13] EPC™ Radio-Frequency Identity Protocols Class-1 Generation-2 UHF RFID Protocol for Communications at 860 MHz – 960 MHz Version 1.2.0.

[14] http://www.ramtron.com/.

NON-DIRECTED INDOOR OPTICAL WIRELESS NETWORK WITH A GRID OF DIRECT FIBER COUPLED CEILING TRANSCEIVERS FOR WIRELESS EPON CONNECTIVITY

Dimitar Kolev[1], Takahiro Kubo[2], Takashi Yamada[2], Naoto Yoshimoto[2], Kazuhiko Wakamori[1]

[1] Graduate School of Global Information and Telecommunication Studies, Waseda University, Japan
[2] NTT Access Network Service Systems Laboratories, NTT Corporation, Japan

ABSTRACT

In this paper we propose an optical wireless system for indoor communication with a grid of ceiling transceivers, based on direct fiber coupling technology. The proposed network is fully compatible with EPON standard that uses point to multipoint broadcasting in the downstream and can guarantee a high speed two-way connection for multiple mobile devices. We present the transmission analysis for the both downlink and uplink and discuss the eye safety issues regarding our proposal. Furthermore, deeper analysis of the system synchronization is conducted and the distribution of the delay in the overlapping zones is presented.

Keywords— Broadband communication, EPON, eye safety, indoor optical wireless communication

1. INTRODUCTION

Recently, the users' interest in portable devices as notebooks, tablets and Smart phones has drastically increased. With the growing market of different applications for these devices such as high quality TV and radio programs, video and music on demand, games, navigation and so on, we observe an "explosion" in internet traffic. Constantly new mobile technologies as Worldwide Interoperability for Microwave Access (WiMAX - IEEE 802.16) and Long Term Evolution (LTE) - Advanced (formally submitted as a candidate 4G system to ITU-T and finalized by 3GPP in March 2011) are developed but their speed is unable to support a large number of users, concentrated in one place. Furthermore, in the license-free spectrum more often RF conflicts occur. Considering the above issues indoor optical wireless communications became very attractive because of their wide bandwidth, high security, energy efficiency and electromagnetic interference immunity. There are two distinguishable brands of optical communications – visible light communications (VLC) and infrared laser communications. VLC, also referred to as Li-Fi, relies on exchanging current light sources with LEDs and modulates their intensity in order to build a communication line. Communications, based on infrared laser diodes are more expensive and complex since they do not use the current lighting system. However, they can offer much higher data rates and full time system operation regardless of the room lighting. Furthermore, the ambient noise, which is critical for this type of communications, is much weaker in the infrared spectrum compared with the visible one [1].

It is essential to develop new standards for such optical wireless systems, which will undoubtedly dominate future networks, or to make some recommendations regarding their interconnection with the existing backbone networks and current working standards. Nowadays almost all the data is transferred as IP/Ethernet packages. Ethernet passive optical network (EPON) provides seamless connectivity for any type of IP-based or other "packetized" communications [2]. Since Ethernet devices are ubiquitous from the home network all the way through the regional, national and worldwide backbone networks, implementation of EPON is highly cost-effective. In terms of bit rate, EPON service levels for customers are scalable from T1 (1.5Mbps) up to 1Gbps. Compared with similar standards as GPON (ITU-T G.984), EPON supports an unlimited number of Optical Network Units (ONU) that makes it a better choice for indoor spaces with high user density – conference rooms, libraries, trains, etc. The EPON also supports CATV overlay on 1550nm so that the main stream on 1490nm is free for other applications' traffic.

We propose a passive indoor wireless optical system to extend the existing EPON network reach to the mobile end user. Thus we combine the functionality of a well established standard with the ability to provide high-speed mobile network access into a new generation communication network. To be able to create such a system we use the direct fiber coupling technology, because the transceivers are cheap and simple – no laser source or photo diode (PD) is used. Furthermore, such a system is optically transparent due to its independency on the wavelength, bit rate and modulation, used in the rest of the optical network.

In terms of line-of-sight the indoor communication systems can be line-of-sight (LOS) or diffusive. The main advantage of the diffusive systems is that LOS is not required and therefore the link is more stable. The main disadvantages are the high transmission power and the strong effect of multipath distortion.

The LOS system has three basic configurations – directed, in which the transmitter and the receiver are pointed directly toward each other, hybrid, in which one of the elements is pointed directly at the other which has a wide field of view (FOV) or wide beam, and non-directed, in which both

transmitter and receiver are not pointed at each other. All three configurations have their pros and cons- the directed system provides the best energy efficiency and highest speed [3], but it is easily disrupted and difficult to maintain for mobile devices; the hybrid system has a complex pointed module and poor power efficiency due to its wide beam; the non-directed system has the worst energy efficiency and bit rate performance, but is very stable in terms of obstacle disruption and no complexity is necessary for implementing with mobile devices. We propose the usage of an enhanced non-directed configuration with a grid of synchronized ceiling transceivers that increases the link speed and assures reliable connection and full coverage. The broadcasting in the indoor space resembles the EPON standard, because of its point to multipoint topology. EPON connectivity to multiple mobile devices within the coverage area can be provided by our proposal as explained in detail in Section 2. Regarding the uplink, due to the bigger size of the receiver apertures and lower ambient noise, the necessary transmit power levels are relatively low for a non-directed link, as we will show in Section 3.

In previous research often only the downlink is considered and broadcasting services are presented [4]. In terms of uplink, sometimes different technology is used- Wi-Fi, Bluetooth, etc. or the link is created regardless of its price and complexity [3]. Sometimes, a grid of ceiling transmitters is proposed forming cells on the communication plane. Such design would however, lead to handover issues. Most of the works in the field propose ceiling modules with OE/EO conversion that limits their usage and no interconnection with current standards is discussed. Systems, based on direct fiber coupling are also proposed but no multiuser solution is presented [5]. Furthermore, although the experimental data shows good system performance, the receiver aperture size is not mentioned. Our analysis shows that small apertures for mobile devices have huge impact on the system performance and the presented experimental data is not sufficient for building such an indoor optical system. We also show that in such a system the presence of EDFA is inevitable and the ASE noise is a critical factor, but so far this has not been considered.

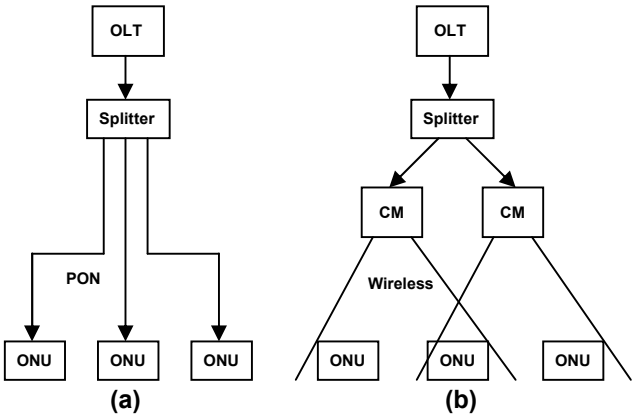


Figure 1. Comparison between: a) EPON structure; and b) proposed structure.

2. SYSTEM DESIGN

The EPON 802.3ah standard deals with the mechanism and control protocols required to reconcile the P2MP topology into the Ethernet framework. The P2MP medium is a passive optical network (PON). When PON is combined with Ethernet protocol, the network is referred to as an Ethernet passive optical network (EPON). P2MP is an asymmetrical medium, based on a tree topology. A typical EPON network is shown in Fig. 1a. In the downstream the signal from the optical line terminal (OLT) passes through a splitter and reaches all the optical network units (ONU). In the upstream direction the signal from the ONU will reach the OLT and no other ONUs. In Fig. 1b is our proposal. The downstream signal passes through a splitter but instead of an ONU, at the end of the fiber optical link, a ceiling module (CM) is connected. In the CM a wide Gaussian beam is formed in order to cover a greater area in the communication plane. The connection between the CM and the ONU is wireless optical.

2.1. Synchronization

In the proposed scenario the high speed downstream broadcasts to all devices. Due to the broadcasting the ONU can be mobile within the range of the wireless network and no handover issues are apparent. However, it is important to synchronize the CM's so that they will broadcast the same information at the same time with no delay. This can be achieved by adding extra delay in the shorter lines between the splitter and the CM. It is important to consider the delay not only in the fiber but also in the wireless part.

In Fig. 2a we show the uplink beam spot on the ceiling when it covers four receivers at the same time. The system can be designed in such a way that the four receiver apertures are positioned on the beam waist thus minor movement in any direction will result in only one or two receivers in the beam spot. When there is only one receiver in the spot there will be no multiple received signals. When there are four receivers, due to the grid pattern the distances between the center of the beam and the receivers are equal. Therefore, the delays will be equal and no signal disruption will be observed. The biggest delay will be reached when only two receivers are positioned in the beam spot and the difference in the distances from them to the center of the beam spot is maximal. This case is shown in Fig. 2c. As we can see, the maximum difference will be achieved when the distances between Rx1 and Rx2, respectively, are *(a-r)* and *r*, where *a* is the distance between the centers of the two receivers and *r* is the radius of the beam spot.

In the downlink, on the communication plane we will have the beam spots of the ceiling transmitters as shown in Fig. 2b. The neighbor spots are overlapping so that no uncovered space is available between them. We can calculate *a* as a function of *r*: *a=sqrt(2)*r*.

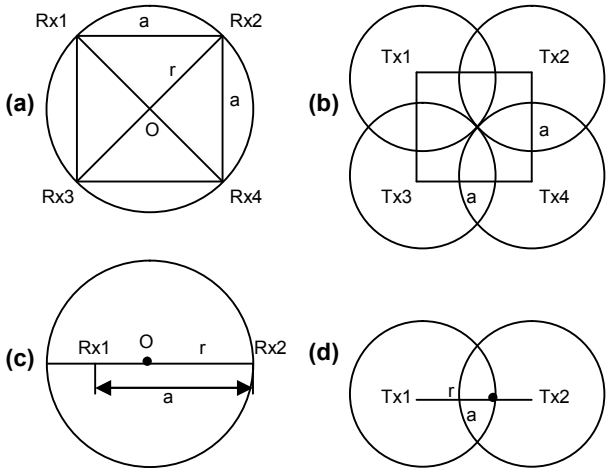


Figure 2. Cell arrangement: a) ceiling; b) communication plane; c) maximum distance in the ceiling grid; d) maximum distance in the communication plane.

Due to the form of the beam spot and the typical rectangular indoor spaces, either there will be an uncovered space close to the walls or there will be a noise coming from the wall reflections because for full coverage part of the spot will reach the walls and reflect back. We will not consider these special cases in our research. We will only estimate the delay in such systems due to the different wireless paths.

In Fig. 2d we show the case with maximum possible delay in the downlink- the distances to Tx1 and Tx2, respectively, are *r* and *(a-r)*, which is the same result as for the uplink. Assuming that the spot size is the same in the both uplink and downlink and the vertical distance between the communication plane and the ceiling can be considered as constant, we conclude that the maximum delays are equal for both ways of communication.

The distances between the two transmitters and the receiver L_1 and L_2 can be calculated as (Fig. 3):

$$L_1 = \sqrt{h^2 + r^2}$$
$$L_2 = \sqrt{h^2 + (a-r)^2} = \sqrt{h^2 + r^2\left(\sqrt{2}-1\right)^2} \quad (1)$$

We can calculate the delay $d=(L_1-L_2)/c$, where c is the speed of light. If we assume that $c=3.10^8$ m/s, the room height h=2m and r=1m we receive maximum possible delay d=0.64ns. Complete delay distribution for an overlapping zone with the above parameters for the case in Fig. 2d can be seen in Fig. 4.

The synchronized broadcasting in the downlink and the uplink with a beam spot, wide enough to assure at least one ceiling receiver in it guarantees flawless two-way communication in any place of the room for mobile users. For better understanding of the new device discovery process and multiuser communications we will discuss the EPON standard more thoroughly. Discovery is the process whereby newly connected or off-line ONUs are provided access to the PON [2]. The process is driven by the OLT, which periodically makes Discovery Time Windows (DTM) available during which off-line ONUs are given the opportunity to make themselves known to the OLT. The discovery process is shown in Fig. 5.

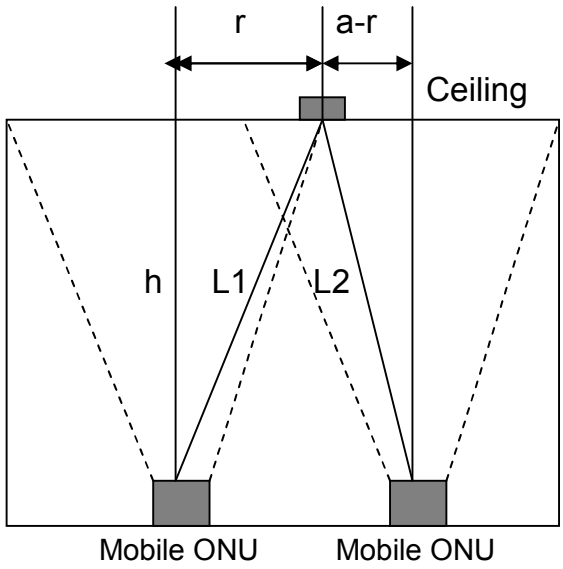


Figure 3. Different paths for different transmitters.

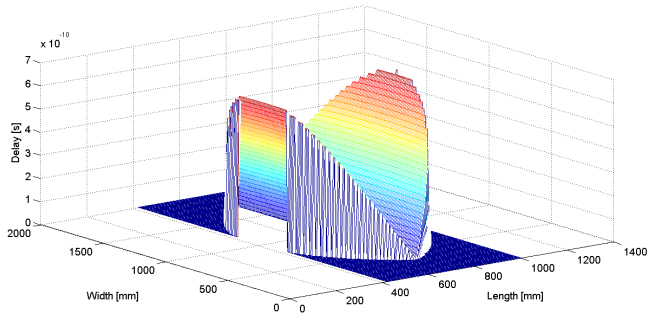


Figure 4. Delay distribution in the overlapped areas.

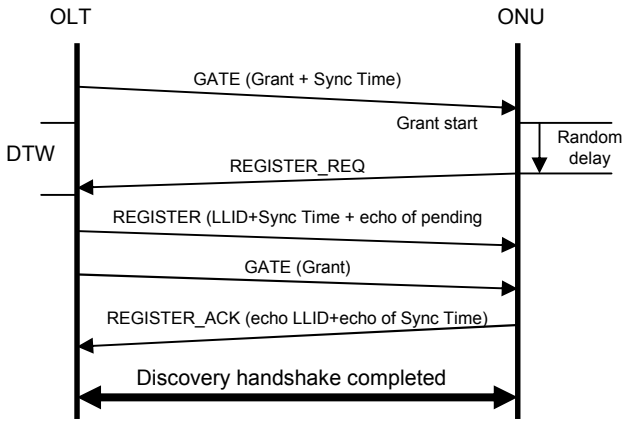


Figure 5. Discovery process.

In a period of time, which is specified by the implementer, OLT signifies that DTM is occurring by broadcasting a discovery gate message, which includes the starting time and length of the DTM. Upon receiving the message, the off-line ONUs wait for a random period of time and transmit a REGISTER_REQ message to the OLT, containing their MAC address and number of maximum pending grants. Only during the DTM multiple ONU can access the PON simultaneously and transmission overlap can occur. When the REGISTER_REQ is received, the OLT registers the ONU, allocating and assigning a new port

identity (LLID) and bonding corresponding MAC to it. Then the OLT sends a register message to the newly discovered ONU, containing the LLID and required synchronization time. Also, the maximum number of pending grants is echoed. OLT schedules the ONU for access to the PON and transmits a standard GATE message allowing the ONU to transmit REGISTER_ACK message. When the REGISTER_ACK message is received, the discovery process for the ONU is complete and normal message traffic can begin. There can be cases when the OLT requires the ONUs to go through the discovery sequence again and cases when ONUs need to deregister.

2.2. Received power and BER analysis

In Fig. 6 we show the scheme of our proposed system with the noise components, optical gain and insertion losses. In the downlink the signal is transmitted from the OLT with transmit optical power P_t. The received optical power in the photodiode (PD) P_{Ar} can be described as a function of the transmit power P_t as:

$$P_{Ar} = P_t L_{tot} G_{EDFA}, \quad (2)$$

where G_{EDFA} represents the gain of the erbium-doped fiber amplifier (EDFA) and L_{tot} represents the total insertion loss in the system. If we assume that L_c=10^($L_{c,dB}$/10) then:

$$L_{c,dB} = L_{split} + L_{coupling} + L_m, \quad (3)$$

where L_{split} is the insertion loss of the splitter, needed to separate the signal to the different ceiling transmitters, $L_{coupling}$ is the insertion loss due to direct coupling technology, and L_m is the link excess margin. There are also a fiber insertion loss and wireless propagation loss, which can be neglected due to their small values and inserted in the link margin L_m component. For a particular configuration L_c will be constant. There is also an insertion loss component L_{beam} that represents the ratio between the total power in the beam spot and the power that enters in the receiver aperture. We discuss it separately from the L_c because it strongly varies according to the beam spot size and the receiving aperture size and we would like to make a deeper study of these relationships. Furthermore, the beam spot and the receiving aperture size in the uplink differ from the ones in the downlink, which is one of the main differences in the both directions. L_{beam} can be calculated as the ratio between the total transmit power in the beam spot P_D with beam waist ω, and the received power $P_{Ar,beam}$ in a receiver aperture with diameter r_2 on distance r_1 from the center of the beam spot [4]:

$$L_{beam} = \frac{P_{Ar,beam}}{P_D} = \left[e^{\frac{-2r_1^2}{\omega^2(z)}} - e^{\frac{-2(r_1+r_2)^2}{\omega^2(z)}} \right] \frac{r_2}{8r_1 + 4r_2}. \quad (4)$$

For a simple example, if the beam spot has a beam waist ω=1m in the communication plane and the receiver with aperture diameter 20mm is located next to the beam waist, the loss is L_{beam}=-45dB.

The total loss in the proposed system will be L_{tot}=L_c L_{beam}. Considering the normal values of all the components, L_{tot} will reach values around -60dB. Normally, the transmitter output power P_t is in the range of (-5dBm ~ 0 dBm) and the receiver sensitivity is in the range of -30dBm for 100Mbps. Therefore, to assure reliable connection an optical amplifier with gain G=30dBm is required. We consider EDFA because of its low noise figure, high flat gain (>20dB) and ability to operate in WDM networks for future capacity increasing. One of the biggest disadvantages of the optical amplifiers though, is the Amplified Spontaneous Emission (ASE) noise that strongly degrades the SNR.

The current in the PD from the received optical signal is:

$$I_s = P_{Ar} \rho_{RX}. \quad (5)$$

Since the proposed system is LOS we assume that there will be no multipath distortion. In both uplink and downlink we use simple compound parabolic concentrator with wide FOV. We calculate the signal-to-noise ratio (SNR) in the proposed system by the formula:

$$SNR = \frac{\left(P_{Ar} \rho_{RX}\right)^2}{\left\langle i_N^2 \right\rangle}, \quad (6)$$

where $<i^2_N>$ is the mean square optical noise current.

The optical noise in the proposed system will consist of several components. In the fiber between the OLT and the CM the *ASE* noise from the EDFA is the dominant noise. It degrades significantly the SNR and limits the system performance. The ASE power is [6]:

$$P_{ASE} = m_t n_{sp} h\nu \Delta\nu_f. \quad (7)$$

We include factor m_t=2 in the noise calculations because two orthogonal optical polarizations can propagate in the amplifier. n_{sp} is a measure of the completeness of the population inversion for the amplifier and can be calculated by n_{sp}=N_2/(N_2-N_1), where N_1 and N_2 are populations of lower and upper laser levels, respectively. $h\nu$ is the photon energy and $\Delta\nu_f$ is the bandwidth of the optical band pass filter, which follows the EDFA and reduces the contribution of ASE to the noise. The corresponding ASE current is

$$I_{ASE} = \rho_{RX} P'_{ASE}. \quad (8)$$

It is important to consider also the fact that P'_{ASE} at the output of the band pass filter will differ with a factor of L_{tot} from the received P'_{ASE} in the PD due to the channel losses that apply not only to the optical signal but also to the optical noise in the link. Noise arises from a beating of the ASE with the optical signal and the mean square beat noise current is:

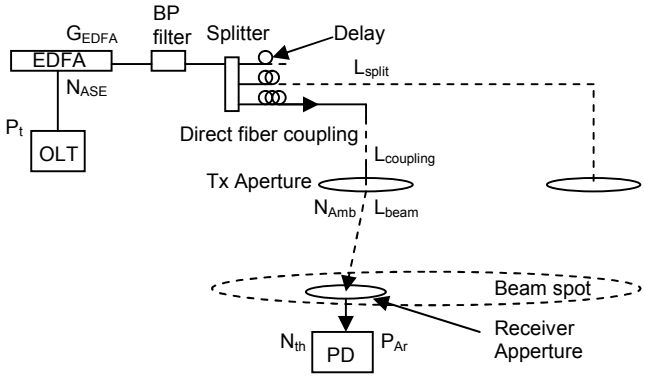


Figure 6. Link budget of the proposed system.

$$\left\langle i_{ase}^2 \right\rangle = 4G_{EDFA} I_s G_{EDFA} I_{ASE} \frac{B}{\Delta \nu_f}, \quad (9)$$

where B is the operating bit rate.

In the wireless optical part the background induced shot noise due to light sources is dominant and can be calculated by:

$$\left\langle i_{bn}^2 \right\rangle = 2e\rho_{RX} P_{bn} B, \quad (10)$$

where P_{bn} is the received background light power and e is the elementary charge.

Finally, the thermal noise in the transimpedance amplifier (TIA) following the p-i-n PD is the dominant noise component in the receiver side. Its mean square value can be calculated as:

$$\left\langle i_{th}^2 \right\rangle = \frac{4kTB}{R_{in}}, \quad (11)$$

where k is the Boltzmann's constant, T is the absolute temperature, and R_{in} is the feedback resistance.

As a final equation for the SNR in the receiver for the downlink we can write:

$$SNR_d = \frac{\left(P_{Ar}\rho_{RX}\right)^2}{\left\langle i_{ase}^2 \right\rangle + \left\langle i_{bn}^2 \right\rangle + \left\langle i_{th}^2 \right\rangle} = \\ = \frac{\left(P_{t,d} L_{tot} G_{EDFA} \rho_{RX}\right)^2}{4G_{EDFA} I_s G_{EDFA} I_{ASE} \frac{B}{\Delta \nu_f} L_{tot} + 2e\rho_{RX} P_{bn,d} B + \frac{4kTB}{R_{in}}} \quad (12)$$

In the uplink the received power will be higher due to the bigger ceiling receiver aperture and higher transmit optical power. However, some of the noise components will be different. The thermal noise will remain constant. The EDFA is next to the OLT and will serve as a preamplifier. Therefore the ASE noise power will not decrease due to attenuation as in the downlink and will fully reach the PD. The Ambient noise has smaller values in the uplink due to the absence of direct ceiling light reaching the receiver aperture but will be amplified in the EDFA. The final expression for the SNR in the uplink will be:

$$SNR_u = \frac{\left(P_{Ar}\rho_{RX}\right)^2}{\left\langle i_{ase}^2 \right\rangle + \left\langle i_{bn}^2 \right\rangle + \left\langle i_{th}^2 \right\rangle} = \\ = \frac{\left(P_{t,u} L_{tot} G_{EDFA} \rho_{RX}\right)^2}{4G_{EDFA} I_s G_{EDFA} I_{ASE} \frac{B}{\Delta \nu_f} + 2e\rho_{RX} P_{bn,u} B G_{EDFA} + \frac{4kTB}{R_{in}}} \quad (13)$$

In our proposed system an OOK modulation is employed, so the bit error rate (BER) can be written as:

$$BER = \frac{1}{2} erfc\left(\sqrt{\frac{SNR}{2}}\right) \quad (14)$$

2.3. Eye safety

Since the proposed system will operate in close contact to human users, it is important to assure user safety. Because the requirements are stricter for eye safety we will consider them and assume that skin safety requirements are also fulfilled. When discussing eye safety, there are two important parameters to define - Maximum Permissible Exposure (MPE) and Accessible Emission Limit (AEL) [7,8]. MPE shows the maximum power, on which a surface can be exposed and the Class 1 requirements will be obeyed. It is normally measured in Wm^{-2} or Jm^{-2}. When we take into consideration the eye surface A_r for a given wavelength that would receive the laser radiation we can define the maximum safe emit power as:

$$AEL = MPE.A_r. \quad (15)$$

The most important characteristics of the system are the exposure time and the wavelength. The proposal system is compatible with EPON standard, which means the downlink wavelength is 1490nm and the uplink wavelength is 1310m. For exposure time we will choose the longest possible time of 30000s (8 hours). We use a wide beam which means that not all the transmit energy will be absorbed from the eye. The maximum received power P_{eye} in the eye will be when the center of the beam spot is pointed directly into the center of the eye and it can be calculated from:

$$P_{Eye}(z) = P_T\left(1 - e^{\frac{-2R_r^2}{\omega^2(z)}}\right) \le AEL_N. \quad (16)$$

where P_T is the total transmit power, R_r is the radius of the eye and and $\omega(z)=z\tan\theta$. Actually, if $P_{Eye}=AEL_N$ then:

$$P_T(z) = \frac{AEL_N}{1 - e^{\frac{-2R_r^2}{\omega^2(z)}}}. \quad (17)$$

The wavelengths for the uplink and the downlink are in different groups regarding their effect on the human eye. Therefore, it is necessary to calculate the possible AEL levels for both of them and later take these values into account when the performance of the system is discussed.

As mentioned above, the wavelength in the downlink is 1490nm. From the MPE tables we can check the MPE for such a system – 1000 Wm^{-2}. The aperture diameter for eye irradiance is 3.5mm. If we assume that the transmitter diameter is 5cm and the eye is adjacent to the transmitting surface then we can calculate from Eq. 17 that P_T>1W would not damage the human eye. In the downlink though, we must consider that the transmitter is mounted on the ceiling and normally it is not possible to have the eye and transmitting surface adjacent. Usually, there is a distance of more than ten centimeters between the eye and the transmitting aperture. Therefore, if necessary, we can introduce a nominal hazard zone (NHZ) in which the eye safety regulations are not met but it is unlikely to have an eye exposure under normal conditions. Thus, the emit power can be further increased if needed.

In the uplink the used wavelength is 1330nm and has more serious impact on human eyes. Furthermore, NHZ zone cannot be introduced because of the close contact between the mobile device and the human eyes and the high probability of laser irradiance to reach the eye. From the MPE tables we can find the equation:

$$MPE = 18C_4C_6C_7T_2^{-0.25}, \quad (18)$$

where $C_4=10^{0.002(\lambda-700)}$, $C_6=\alpha_{max}/\alpha_{min}$ for extended source and $\alpha > \alpha_{max}$, and $C_7=8$ for λ in the interval 1200nm~1400nm. The aperture diameter for eye irradiance is d=7mm. From Eq. 18 we can calculate the MPE that MPE_{1310}= 4.8mW/mm^2. After substitution in Eq. 15 we receive AEL_{1310}=185.4mW.

To assure the system safety, it is strongly recommended that users limit the output power at the ceiling aperture to the AEL limits by standard – 10mW. Considering the link budget calculations above, assuming an OLT optical output power of 0dBm and 15-20dB loss in the network before emitting through the ceiling aperture we can conclude that for EDFA gains up to 30dB the transmitted power will be less than 10dBm and thus eye safety is achieved.

3. RESULTS AND DISCUSSION

For performance evaluation of our proposal we will show the relationship between the EDFA gain and the mean square noise currents and the BER respectively for the both uplink and downlink. The reason for examining the relation to G_{EDFA} is that the OLT and ONU output optical power is fixed and should not be subject to change in order to use the proposed system. For the downlink we assumed $P_{t,d}$=0dBm and for the uplink $P_{t,u}$=4.8dBm.

The main differences in the both directions are, as discussed in Section 2, the ambient noise level, the receiver aperture size and the transmit power. For simplicity, we assume that the noise power in the downlink is $P_{bn,d}$=10μW and in the uplink- $P_{bn,u}$=1 μW. For clear understanding of the results we have fixed the link losses $L_{c,dB}$=-20dB that are equal for both uplink and downlink, and will observe the system behavior when the beam spot and receiver aperture size is changed. Because of Gaussian power distribution, in our theoretical model calculations we consider that the receiver is located within the spot with lowest possible received power – next to the beam waist. Thus we can guarantee stable system performance in the covered area due to the higher received power. However, in a real scenario the zones that are close to the beam waist are overlapped with neighbor spots and the resulting received power and consequently the BER of the system will further improve. The values of the parameters for our calculations are given in Table 1.

Table 1. Mathematical model parameters

Parameter	Symbol	Value
Photodetector responsitivity	ρ_{RX}	0.8
Absolute temperature	T	300K
Photodetector load resistor	R_{in}	50Ω
EDFA gain	G_{EDFA}	1000
Orthogonal Polarization factor	m_t	2
Population inversion factor	n_{sp}	2.25
BP filter bandwidth	$\Delta\nu_f$	12.4x10^{9}Hz
Bit rate	B	100x10^{6}bit/s
Boltzmann's constant	k	1.3807x10^{-23}m^{2}kg s^{-2}K^{-1}
Elementary charge	e	1.6022x10^{-19}C

3.1. Downlink

There are requirements for the receiver aperture in mobile devices to be as small as possible. Considering the high level of ambient noise the required transmit power in the downlink would have very high levels.

For better understanding of the BER results we show the mean square values of the noise components and their dependence on the EDFA gain in Fig. 7.

Normally, the ASE component increases with the higher EDFA gain. Both thermal and ambient noise components remain constant since they are not amplified in the EDFA. For smaller values of the gain the thermal component is dominant. For a certain gain level though, the ASE component is equal to the thermal one and becomes dominant with further increasing of the gain.

In Fig. 8 we observe the calculated results for a 100Mbps link with a receiver aperture diameter of 50mm for three different beam spot diameters – 0.5m, 0.75m and 1m. We can observe that for small G_{EDFA} when the thermal noise is dominant the BER improves as G_{EDFA} increases. For a certain G_{EDFA} level, which differs with the configuration, the ASE noise component becomes dominant and the BER degrades because of the combined effect of thermal and ASE noise. As the results on Fig. 8 show, when the beam spot size is increased the optical power, and respectively the ASE noise that reaches the receiving aperture, is smaller and higher gain is needed before the ASE noise becomes stronger than the constant thermal noise and BER begins to increase. From the same figure we can see that the system can provide connection with BER of 10^{-9} for beam spots with radius up to 0.75m.

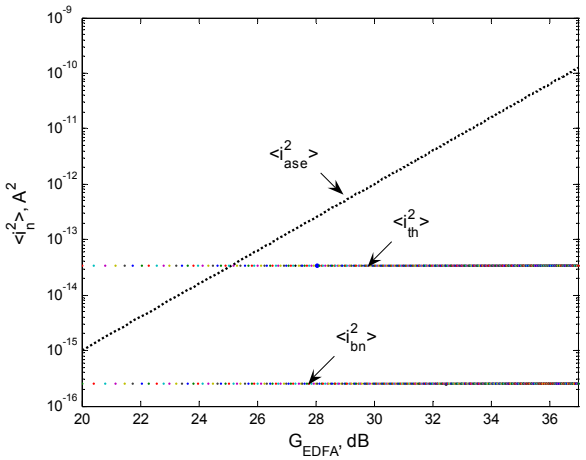


Figure 7. Mean square noise currents vs G_{EDFA} in the downlink.

The poor performance is due to the lack of an amplifier on the receiver side. In future research a transimpedance amplifier or an APD can be used in the receiver for better performance.

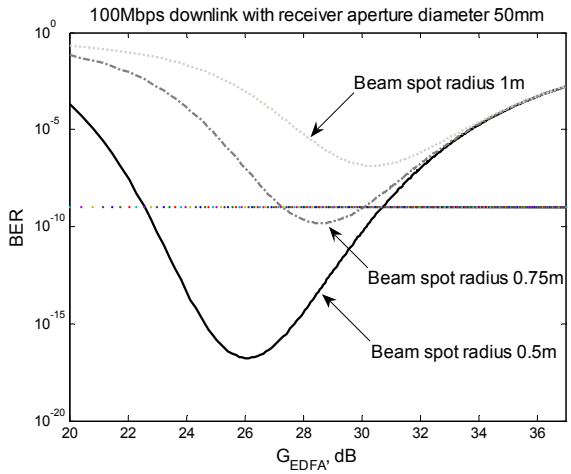


Figure 8. BER vs G_{EDFA} for different beam spot radiuses with receiver aperture diameter 50mm.

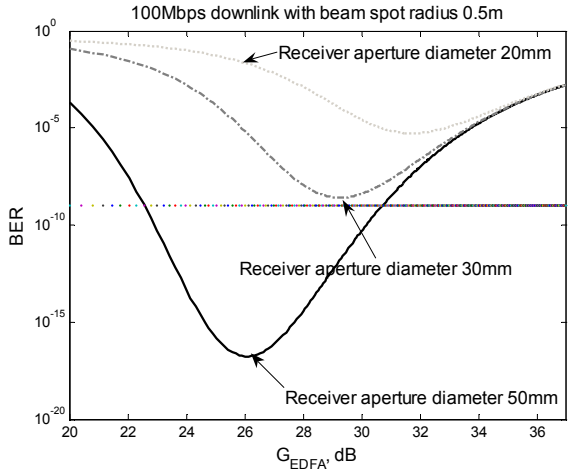


Figure 9. BER vs G_{EDFA} for different receiver aperture diameters and beam spot radius 0.5mm.

In Fig. 9 we show the relationship between the G_{EDFA} and the BER for a 100Mbps link with beam spot size diameter 0.5m for different receiver apertures – 20mm, 30mm and 50mm. We can observe that the gain levels, on which the BER starts to increase, change with the aperture size. The reason for this effect is the different quantity of received optical and noise power. When the aperture is small, a higher level of ASE noise is required to dominate the thermal noise and consequently increase the BER.

In the downlink, the maximum optical power is on the EDFA output and for optimal work it will be less than 30dBm. Considering the L_c losses in the fiber part, the optical power in the wireless part will be well below the eye safety levels.

3.2. Uplink

In the uplink the ambient noise power is much lower than the downlink because no direct light from the light sources will reach the CM (receiver FOV can eliminate the direct light from sources). Furthermore, the receiver aperture size is not limited and can be much bigger thus receiving much more power than the small aperture of the mobile devices. However, due to the shorter wavelength of the emitting signal and the higher probability of laser light to enter in the human eye, the eye safety requirements are stricter.

In Fig. 10 we show the mean square noise current dependence on the G_{EDFA}. In the uplink the thermal noise in the receiver is again constant. However, despite the ASE noise, the ambient noise is also dependent on the optical gain. This is due to the fact that the noise power, incident in the CM, will be amplified in the EDFA before reaching the OLT receiver.

In Fig. 11 we show the relationship between the G_{EDFA} and the BER for a 100Mbps uplink for two beam spot diameters- 2m and 4m, received in two apertures with different sizes- 50mm and 100mm diameter. We observe the same effect as in the downlink that for bigger EDFA gain the BER degrades. Due to the bigger aperture size, lower ambient noise and higher transmit power, a reliable link can be achieved with wider beam spots – 1.5m ~ 2m. Lower speed can further increase the energy efficiency.

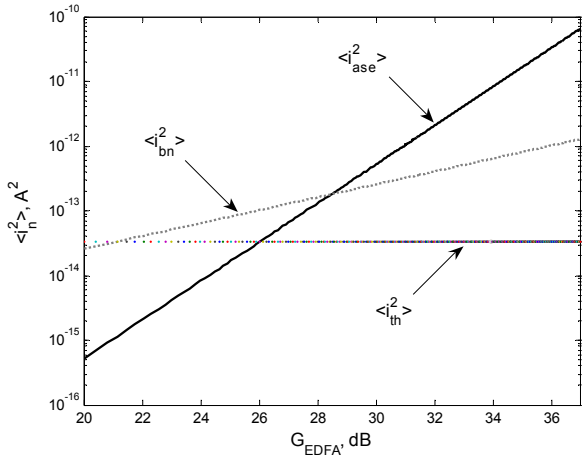


Figure 10. Mean square noise currents vs G_{EDFA} in the uplink.

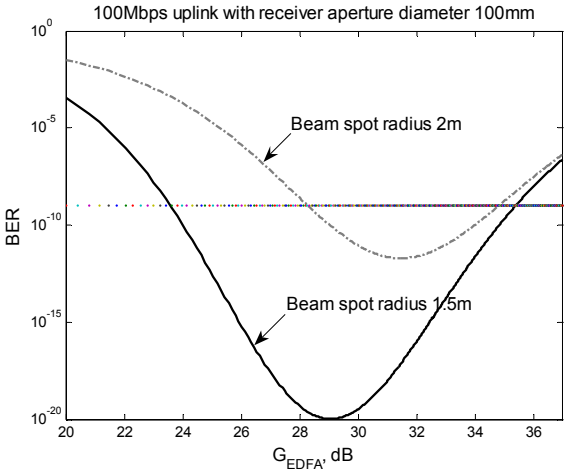


Figure 11. BER vs EDFA gain in a 100Mbps for different beam spot radiuses with receiver aperture diameter 100mm.

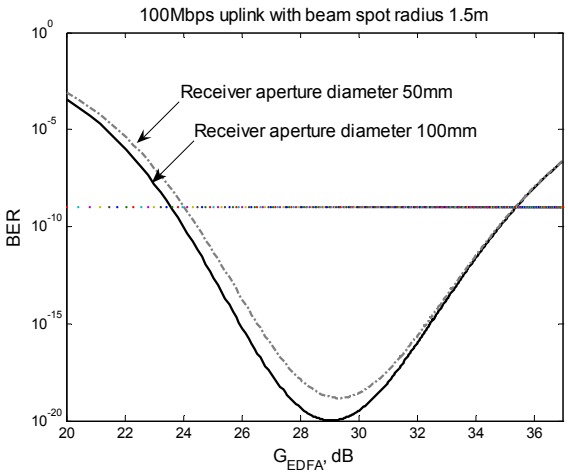


Figure 12. BER vs EDFA gain in a 100Mbps for different receiver aperture diameters and beam spot radius 1.5m.

In Fig. 12 we show the effect of receiver aperture diameters with different size. As we can observe, there is no significant difference in the performance even when the receiver aperture diameter is doubled.

In the uplink, transmit optical power is 3mW which covers the eye safety requirements calculated in Section 2.3. Furthermore, the required energy for such optical power level is much lower than the energy for typical Wi-Fi transmitter [9]. Therefore, using the proposed system in different portable devices instead of Wi-Fi would significantly lower the power consumption.

4. CONCLUSION

We have presented an indoor optical wireless system with multiple ceiling transceivers that is compatible with the EPON standard and offers multiple mobile user communication. We showed that the required power in the uplink for a non-direct configuration is relatively small compared to similar RF technologies and thus energy efficiency in the mobile side can be achieved. Furthermore, the proposed system is based on direct fiber coupling technology, which provides flexibility with the choice of wavelength, bit rate and modulation. In our mathematical model we also considered the ambient noise, thermal noise and ASE noise power in both directions, the eye safety limitations and the delay distribution in the overlapped areas. We did not consider the usage of transimpedance amplifier on the receiver side. This would further decrease the dominant thermal noise and improve system performance in terms of speed and coverage.

The proposed technology can contribute to a sustainable society with its low power consumption compared to RF technologies. Furthermore, the nature of laser limits the network coverage to a particular indoor space ensuring high security of personal data and communications and free RF spectrum. This lowers the human exposure to electromagnetic waves and allows the free resources to be used for other applications.

REFERENCES

[1] A.C. Boucouvalas, “Indoor ambient light noise and its effect on wireless optical links,” *IEE Proceedings on Optoelectronics*, vol.143 no.6, pp.334-338, 1996.

[2] IEEE standard 802.3ah – 2004.

[3] D. R. Kolev, K. Wakamori, M. Matsumoto, T. Kubo, T. Yamada, N. Yoshimoto, “Gigabit Indoor Laser Communication System for a Mobile User with MEMS Mirrors and Image Sensors,” *International workshop on Optical Wireless Communications*, Italy, 2012.

[4] D. Kolev, M. Matsumoto, “Indoor HDTV Broadcasting through Line-of-Sight Optical Wireless Link with Direct Fiber Coupling Transmitter”, *40th IIEEJ Conference*, 2012.

[5] K. Wang, A. Nirmalathas, Ch. Lim, and E. Skafidas, "High-speed duplex optical wireless communication system for indoor personal area networks," *Opt. Express*, vol.18 no.24, pp. 25199-25216, 2010.

[6] E. Desurvire, “Erbium-Doped Fiber Amplifiers: Principles and Applications,” Wiley Series in Telecommunications and Signal Processing, 1994.

[7] ANSI Z136.1-2000 standard.

[8] IEC 60825-2, 2004 standard.

[9] A. Carroll, G. Heiser, “An Analysis of Power Consumption in a Smartphone,” *Proc. of the USENIX annual technical conference*, pp. 21-35, 2010.

SESSION 3

SUPPORTING REMOTE COMMUNITIES

IMPLEMENTATION ROADMAP FOR DOWNSCALING DROUGHT FORECASTS IN MBEERE USING ITIKI

Dr. Muthoni Masinde1; *Dr. Antoine Bagula2;* *Prof. Nzioka Muthama3*

1 Hasso Plattner ICT4D Research School, University of Cape Town, muthonimasinde@yahoo.com
2 University of Cape Town, bagula@cs.uct.ac.za
3 University of Nairobi, jmuthama@uonbi.ac.ke

ABSTRACT

*Mbeere is in Eastern Kenya and it has an average of 550 mm annual rainfall and therefore classified under Arid and Semi-Arid Lands. It has fragile ecosystems, unfavorable climate, poor infrastructure and historical marginalization; the perennial natural disasters here are droughts. Of importance to this paper is the fact that despite its vast area of 2,093 km2, there is no single weather station serving the area. The main source of livelihood is rain-fed marginal farming and livestock keeping by small-scale and peasant farmers who rely mostly on the indigenous knowledge of seasons in making cropping decisions. ITIKI; acronym for **I**nformation **T**echnology and **I**ndigenous **K**nowledge with **I**ntelligence is a bridge that integrates indigenous drought forecasting approach into the scientific drought forecasting approach. ITIKI, a framework initiated by the authors of this paper was adopted and adapted from the word itiki which is the name used among the Mbeere people to refer to an indigenous bridge used for decades to go across rivers. ITIKI makes use of mobile phones, wireless sensor networks and artificial intelligence to downscale weather/drought forecasts to individual farmers. ITIKI implementation project in Mbeere commenced in August 2012; this paper describes the implementation roadmap for this project.*

Keywords— ITIKI, Indigenous Knowledge, Indigenous Knowledge Weather Forecasts, Mbeere, Drought Early Warning System

1. INTRODUCTION

The Arid and Semi-Arid Lands (ASALs) of Kenya make up more than 80% of the Country's landmass. The latter is prone to harsh weather; this, among other reasons has seen the Country consistently contribute the highest number of people affected by natural disasters in Africa for the last two decades ([5] and [6]). Effective drought early warning systems (DEWS) have high potential in making a contribution towards tackling the current cycle of droughts in Mbeere, Kenya and the larger Sub-Saharan Africa (SSA). However, successful DEWS rely on weather forecasting systems; implementation of the latter in most countries in SSA is hampered by among other things, inadequate coverage by weather stations. Further, the content, format and dissemination channels of the scientific Seasonal Climate Forecasts (SCFs) do not address the farmers' needs; the farmers have in turn continued to rely on their now endangered indigenous knowledge forecasts (IKFs) to derive critical cropping decisions ([1] and [2]).

ITIKI; acronym for **I**nformation **T**echnology and **I**ndigenous **K**nowledge with **I**ntelligence is a bridge that integrates indigenous and the scientific drought forecasting approaches. ITKI was developed as a novel bridge that combines the strengths of SCFs and IKFs to deliver a DEWS composed of four elements: (1) Drought Knowledge (2) Drought Monitoring and Prediction; (3) Drought Communication and Dissemination; (4) Response Capability.

To tackle the diverse characteristics of these two knowledge systems, ITIKI is oiled using three ICTs: (1) mobile phones; (2) Wireless sensor networks; (3) Artificial intelligence (agents, fuzzy logic and artificial neural networks) ([3] and [4]). In order to collect real data to test ITIKI, a case study of two communities in Kenya (Mbeere from eastern and Abanyole from western) was carried out. On completion of the system prototype, participants from the two communities evaluated it; based on content and format of the integrated forecasts, up to 90% of Mbeere respondents gave a score of 'excellent' and also gave commitment to participate in post-tests system's deployment phase. From these findings, a decision was reached in August 2012 to implement ITIKI among the Mbeere people.

The Mbeere people occupy the former Mbeere District, which is classified under the ASALs. With a population of 168,000, the Mbeeres are the majority in the Region which has a population of about 220,000[7]. Mbeere is about 2,093 km^2 and it is located in Eastern side of Kenya. It lies between Latitudes 0^o 20' and 0^o50' South and Longitude 37^0 16' and 37^0 56' East. Other aspects of data analysis of the case study revealed that the Mbeeres were far more disadvantaged than the Abanyoles and the Community was therefore selected for the first phase of ITIKI implementation. In order to ensure a people-drive implementation, a purely community-based approach was adopted. To this end, collaboration with Community-Based-Ogranisation, Kiritiri Orphans and Vulnerable Children Advocate (KOVCA) was established. An Advisory Board currently made up of an ITIKI expert, a representative from the Kenya Meteorological Department and a church (Catholic Church) representative was set up to advise on the project direction. This paper describes how

the various components of ITIKI have been pieced together in creating the on-going DEWS implementation in Mbeere.

The rest of the paper is organized as follows: Section 2 discusses the theoretical literature on which the paper is hinged while ITIKI's Architecture is described in Section 3. Section 4 details the methodology we have adopted and Section 5 introduces the Implementation Roadmap being followed in the Mbeere case study. Discussion, Conclusion and Further work are presented in Section 6

2. BACKGROUND LITERATURE

2.1. Droughts Prediction and ICTs

Among other things, drought prediction plays a critical role in mitigating the negative effects of droughts. Parametric indicators of drought commonly computed are: (1) duration; (2) severity; (3) location of the drought in absolute time (initial and termination time points); (4) area of the drought coverage; (5) magnitude/density of the drought computed by getting the ratio of severity to duration[8]. There are several well-developed indices for quantifying effects of droughts in terms of these parameters; among these is the Effective Drought Index (EDI) which differs from the rest of the indices in a number of ways, one being that it calculates drought on a daily basis. The others use scales such as weekly, monthly bi-monthly, and so on ([6] and [9]).

ICTs can be used to significantly improve drought prediction. The ITU acknowledged the critical role of ICTs, especially in addressing food insecurity (mostly a consequence of drought) and suggested that ICTs can be used: (1) to provide the remote sensing infrastructure, such as Wireless Sensor Networks(WSNs); (2) as the equipment (software and hardware) for analysis of drought data, including statistics, modeling and mapping, for example; laptops, servers, databases, GIS, data mining and neural networks; (3) as the communication infrastructure to disseminate the relevant information to farmers/consumers, for example Internet and mobile phones[10].

2.2. Seasonal Climate Forecasts

The main climate variables of interest for societal applications are atmospheric temperature, rainfall and humidity[11]. The current approaches used for producing Seasonal Climate Forecasts (SCFs) include the use of: (1) Physically based dynamical global/general climate models GCMs); (2) Regional Climate Models (RCMs); (3) Empirically based statistical; and (4) A combination of dynamical and empirical models. All these models generally produce forecast information at 'coarse' spatial resolution (of the order of 100–200 km), which is presented as the probability of the seasonal rainfall being in the 'above normal', 'below normal', or 'normal' compared with historical trends ([12] and [13]). This may have contributed to the current status where the utilization of such forecasts in SSA is still dismal. In West Africa, for example, efforts to disseminate and apply forecasts are at an experimental stage[14]. There is need for downscaling the forecasts produced by climate models to the desirable level of details required in real-life application models ([15] and [16]).

2.3. Indigenous Knowledge on Droughts

IK a body of knowledge existing within or acquired by local people over a period of time through accumulation of experiences, society-nature relationships, community practices and institutions, and by passing it down through generations[17]. In IK drought forecasting, the local weather and climate are assessed, interpreted and predicted by locally observed variables and experiences using combinations of plant, animals, insects and meteorological and astronomical indications [18]. The entry point for the forecasting is the amassed knowledge of exact arrival of the rainy season. IK on drought forecasting in the tropics falls into six general categories: (1) patterns of seasons (cold, dry, hot, rainy and so on); (2) animal, insects and bird's behaviour; (3) astronomical; (4) meteorological; (5) human nature and behaviour; and (6) behaviour of plants/trees, for example fruit and flower production[3].

Researchers ([1], [19] and [20]) today concur that IK and modern science weather forecasts complement each other; they are not mutually exclusive but significant discordance between the two is still apparent. Clear understanding and careful integration of IK present opportunities especially in the dissemination process of weather forecasts to farmers in SSA because this supports ways that are culturally appropriate and locally relevant to the people. There is a common departure that, generally (not just in climate and weather information), integrating IK into modern science can improve livelihoods ([17], [21], [22], [23], [24] and [25]).

On the question whether IK needs modern science, there is evidence that IK has been eroded and is slowly disappearing. Extreme variations never witnessed before by community members bring IK into disrepute; integrated approaches aimed at giving the communities several levels of risk-preparedness are desirable. Further, using IK alone, it is difficult to forecast beyond a season (say beyond two years); in modern science, this can be achieved by employing technologies such as Artificial Neural Networks. Some terminologies used in IK may sometimes be ambiguous, for example '*abundant rainfall*' may mean rainfall for the day or a season. Finally, unlike modern science, climate change may be difficult to foretell using IK alone.

2.4. Early Warning System for Droughts and ICTs

In [26], early warning (EW) is defined as "*the provision of timely and effective information, through identified institutions, that allows individuals exposed to hazard to take action to avoid or reduce their risk and prepare for effective response.*" Effective early warning systems consist of four components; (1) gathering of the risk knowledge;

(2) monitoring and predicting the situation; (3) communicating the warning messages; (4) responding to the warning [27]. The phenomenal role of ICTs in all the four components cannot be overemphasized; this include remote sensing that enables real-time detection of hazards, SMS technology that allows for direct and individualized delivery of disaster alerts and the instantaneous access of diverse and voluminous information via the Web, just to mention a few.

3. ITIKI ARCHITECTURE

3.1. Overview

ITIKI was realised in form of an *Integrated Drought Early Warning System (DEWS) Architecture*; this framework provided the blue-print for the implementation of the system that integrates ICTs with the indigenous knowledge on droughts. The Framework design was guided by the four components that make up an effective early warning system.

3.2. Features of ITIKI

The overall goal of ITIKI was to come up with a ***relevant***, ***affordable***, ***sustainable***, ***integrated***, ***resilient***, ***useable***, ***effective***, ***generic***, and ***micro-level*** early warning system for droughts for the Sub-Saharan Africa and Africa at large. Below is how each of these attributes is achieved in our integrated framework:

3.2.1. Indigenous Knowledge

Going by the phrase by Stern[28], "*The effectiveness of forecast information depends strongly on the systems that distribute the information, the channels of distribution, the recipients' models of understanding and judgment about the information sources, and the ways in which the information is presented*." One way of achieving an ***effective*** early warning system for droughts is therefore to put into consideration the targeted users' coping strategies, cultural traits and specific situations. In the case of the Sub-Saharan Africa, this is easily achieved by incorporating the local people's indigenous knowledge on weather/climate forecasting([17], [21], [29] and [30]).

3.2.2. Effective Drought Index

As the name suggests, the Effective Drought Index (EDI), is a ***very effective*** index compared to other drought indices. Its uniqueness stems from the fact that it provides spatial and temporal distribution of droughts on a daily-basis[9]. EDI computes the intensity of droughts by using cumulative precipitation as a weighting function of time and also gives the Available Water Resources Index (AWRI); the latter is a measure of hydrological drought and can be used to assess the quantity of soil moisture. By incorporating it into our drought early warning system framework, it makes it possible to ***quantify and qualify droughts in micro scale*** (time and spatial distribution) as well as in ***absolute terms***[6].

3.2.3. Wireless Sensor Networks

A deeper look into the problem of early warning system for droughts in SSA reveals a grave situation where the meteorological institutions the National Meteorological Services (NMSs) charged with weather forecasting rely on weather stations that are thousands of kilometres apart ([31] and [32]). This sparse network makes it difficult to provide locally relevant information necessary for scaling weather information down to the local (say village level) communities. Furthermore, weather stations are very expensive and their operation may be difficult to sustain in many developing countries where the lack of expertise and high cost of maintenance may hamper operation after funding from donors. In our framework, the now readily available versatile and WSNs-based weather stations were employed to fill the gap. Further WSNs are used to automatically extend the available climate maps and prediction through (1) collection of climate data; (2) analysis of this data; and (3) modeling of climate change in the remote villages.

3.2.4. Mobile Phones

Africa has achieved a mobile phone penetration level much higher than that of computers[33]. With well-designed solutions, the use of these phones can be extended from the traditional use, as mere communication devices, to computing devices on which the much needed e-applications can be executed. Our DEWS utilises this window of opportunity by using the mobile phone to not only disseminate drought alerts but also as an input device for the IK. This way, the system is both ***affordable*** and ***sustainable***.

3.2.5. Artificial Intelligence

In order to create an ***integrated*** system that can juggle all these myriad moving parts at the macro and micro-level, some reasoning was necessary; use of intelligent agents achieved this. Further, IK on weather and drought is so rich; it has been said to be holistic[34]; in order to model this aspect of IK and ensure preservation of its richness, we employed the use of Fuzzy Sets[35]. In order to build a ***complete*** early warning system, forecasting/predicting future droughts was crucial; ***Artificial Neural Networks (ANNs)*** were used for this purpose.

3.3. ITIKI Architecture

ITIKI Architecture is shown in Figure 1 below:

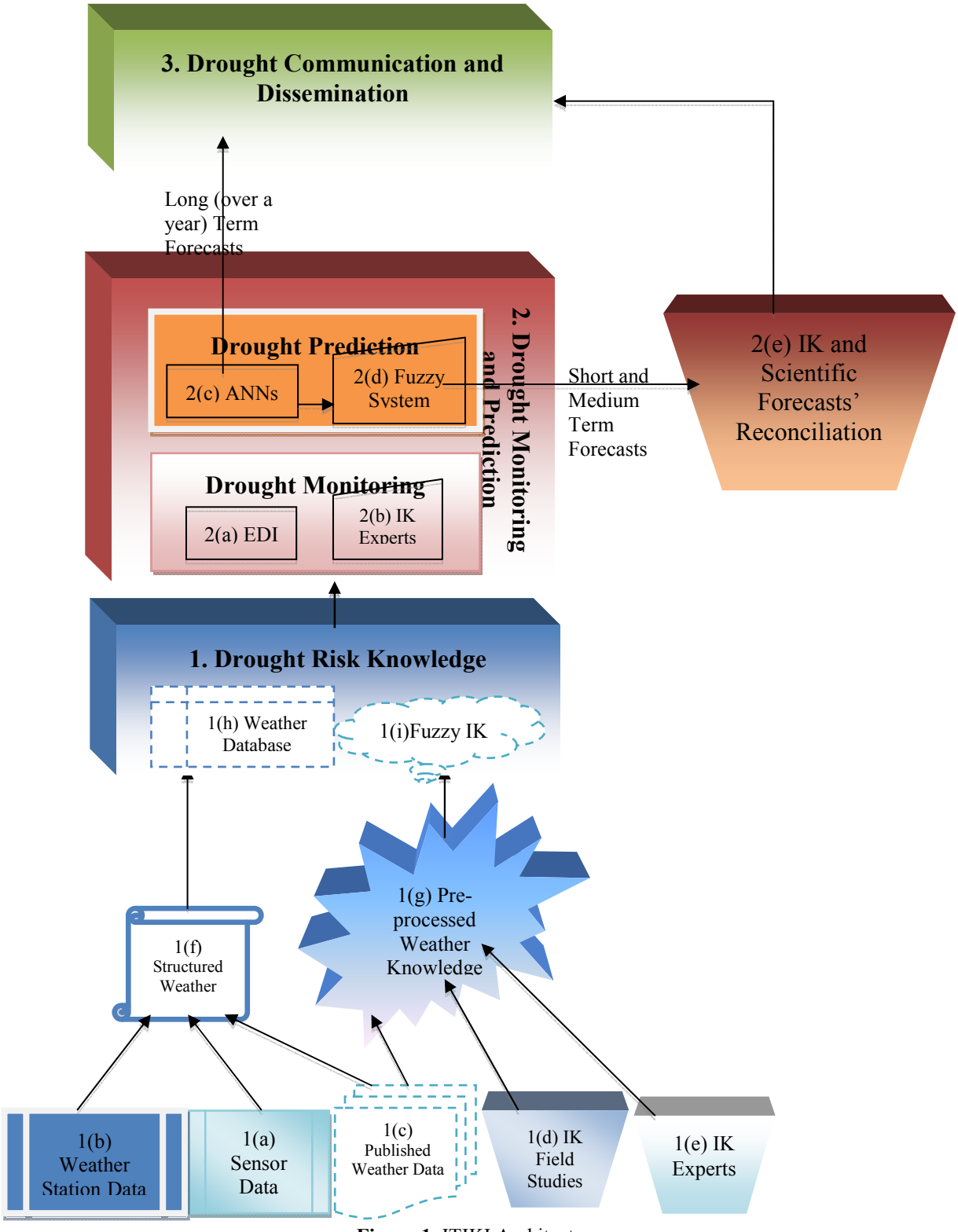


Figure 1. ITIKI Architecture

3.3.1. Element 1: Drought Risk Knowledge

1(a): Using wireless sensors that are capable of sensing temperature, humidity, atmospheric pressure, wind (direction and speed), precipitation and soil moisture, weather data is automatically collected and sent to a structured store ***1(f)*** in form of text messages (SMS).

1(b): Rainfall data observed from rainfall stations (contains only rain gauges, the Mbeere case) stations is manually entered into the system and stored in the same database as the sensors' data.

1(c): Other data elements (IK) are retrieved from various publications available in print and on-line. These are in form of limited studies on IK in Mbeere and SCFs by KMD. Out of these, the structured elements are stored in ***1(f)*** while the unstructured ones are stored in ***1(g)***.

1(d): IK on droughts collected during various field studies is stored in 1***(g)***.

1(e): This is the real-time IK from the IK Experts.

1(h) and **1(i):** the structured data is stored in a database ***1(h)*** while the pre-processed indigenous knowledge is represented as Fuzzy Sets ***1(i)***.

3.3.2. Element 2: Monitoring and Prediction

This was implemented using two sub-components: (1) Drought Monitoring that pre-processes the data to detect suggestive patterns as well minimise duplicates and other errors. This is achieved through EDI Monitor (**2(a)**) and IK Experts (**2(b)**) ; and (2) Drought Prediction using Artificial Neural Networks ((**2(c)**) and Fuzzy Logic System ((**2(d)**). In **2(e)**, the resulting forecasts are reviewed by both the scientists and IK Experts after which 'reconciled' forecasts are generated and passed to the Dissemination component. This is partially a manual activity where the meteorologists and the IK experts sit to reconcile SCFs and IKFs. However, the short-term forecasts (a few hours to two weeks) do not need the manual 'reconciliation'; the system intelligently reconciles the two (from IK and from ANNs) and sends them to the Drought Communication and Dissemination Component. Further, in line with fuzzy system, for purses of 'recovering' IK's original meaning/format, the output 2(e) is passed through 1(i) for Defuzzification

3.3.3. Element 3: Forecasts Dissemination

Mobile phones are used to send customized forecasts in form of test message and where possible, free phone calls to the farmers. Other forecasts are posted on websites while others are generated in audio formats that can be broadcasted via community radios stations and visual displays on strategically located village digital billboards. Though not implemented, the Framework is designed to support natural language processing to allow for translation of the forecasts into the local languages.

4. ITIKI IMPLEMENTATION - MBEERE CASE

4.1. Methodology

4.1.1. About the Mbeere People

With an average of 750mm (most parts receive less than 550 mm) annual rainfall, Mbeere is classified under Arid and Semi-Arid Lands (ASALs). A further classification of the ASALs places Mbeere under Category C; 50-85% of the land is arid (Republic of Kenya, 2008), Its terrain is characterised by scattered outcropping hills and its extensive altitudinal range of the area influences the temperature, which ranges from 15^{o}c to 30^{o}c. The main source (over 80%) of livelihood is rain-fed marginal farming and livestock (agro-pastoralists) keeping. This being the case, the farmers here rely mostly on the knowledge of seasons in making cropping decisions. Like most parts of Kenya, there are two main rain seasons experienced in Mbeere; the March-April-May (MAM) long rains and the October-November-December (OND) short rains. Crops grown include maize, sorghum, millet, beans, cowpeas, green grams, pigeon peas, cotton and tobacco on farms of average of 3.5 Ha. Livestock kept include Cattle (Zebu mostly), goats, sheep, poultry, donkeys and bees (http://www2.kilimo.go.ke).

4.1.2. Sample Data and Sampling Phases

The data used in this research was collected in 2 phases:

Phase I: This took place between August and December 2010 and the aim was to identify the prevalent IK indicators used by the Mbeere people. A guided interview involving 44 respondents was carried out with the help of representatives from the community.

Phase II: This was carried out between June and July 2012 and the objective was to evaluate the usability and relevance of the integrated drought monitoring system. Like in Phase I, guided interviews were conducted and the same (as for Phase I) respondents were approached. During this survey, sample output from the integrated system (output from the mobile application) was demonstrated to the respondents for feedback.

4.1.3. Data Analysis

Weather Observations Equipment: There is no synoptic station in Mbeere; the area is served by one located over 20 kilometres away. The location of this station does not do justice in reporting the weather for the entire Mbeere because it is located in a completely different climatic zone. The explains why over 90% of Mbeere respondents gave a value of between 0 and 1 (out of 3) for both accuracy and relevance of weather forecasts issued by KMD.

Respondents' Geographical Distribution: The respondents in both surveys were drawn from several villages covering over a third of the region

Knowledge and SFCs: Most respondents knew about the services but did not link it to KMD.

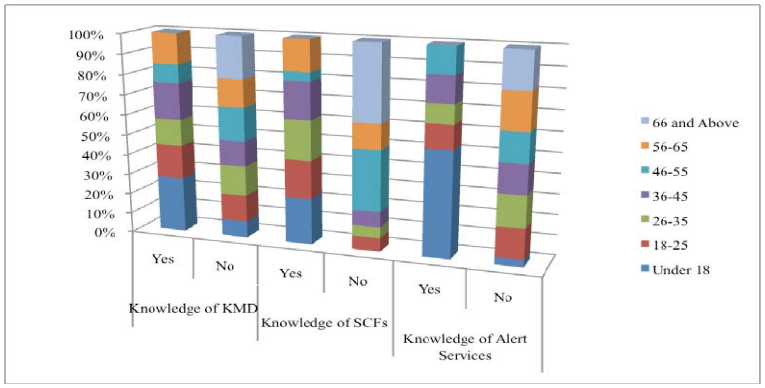


Figure 2. Respondents' Knowledge of SCFs

Distribution by Gender and Age: The respondents were mostly semi-literate small-scale/peasant agro-livestock female farmers over 20 years old.

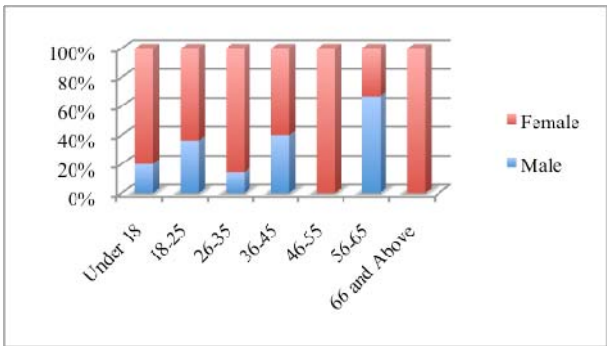


Figure 3. Respondents' Distribution by Age and Gender

Distribution by Level of Education: There were way more literate men than women; the number of illiterate people dropped with decrease in age.

Mobile phone ownership and use: There are generally more men who own phones than women, hoewever, all (except one) the people that did not have a phone did use phones mostly from relatives.

Phones for Weather Forecasts' Dissemination: The lack of relevant weather information among the respondents motivated their willingness to receive forecasts over the mobile phone. 89% of them said that they would like to receive weather information via mobile phones

Other Findings

- More than 50% of these preferred that such forecasts be sent to them via an authority such as Chiefs and Village Elders. Further, over 80% of the respondents preferred that the messages in the local language (Kimbeere);
- Like in most other communities/regions, the indigenous weather/drought indicators among the Mbeere people are associated with seasons. The diversity, level of details and systematic nature of the indicators from the respondents confirmed that the community has very rich IK systems that help the people cope with and adapt to the environment; for example, mixed agriculture. An elaborate list of these indicators was compiled and classified according to five main seasons: January-February dry season, March-April-Mary long rains, June-July Cold Season, August-September dry season and October-November-December short rains;
- The number and category of IK indicators reported by the respondents varied with gender and age brackets. For example, women mostly till the land and they go to fetch water, hence they are able to notice more indicators; and
- Using various examples, respondents expressed concerns that some IK indicators are no longer easy to notice because of biodiversity degradation.

5. IMPLEMENTATION ROAD MAP

5.1. Objective

The objective of the implementation project is to make use of ITIKI to downscale weather and drought forecasts to an individual small-scale farmer in Mbeere.

5.2. Implementation Strategy

(i) A focus group made up of 12 people representing 12 villages was formed in the first week of September 2012; these were selected from the list of 44 people that participated in piloting the ITIKI's prototype;

(ii) Weekly focus group discussions are held during which reporting and deliberations on various weather/drought indicators are done;

(iii) Though not very relevant to Mbeere, the October-November-December Seasonal Climate Forecasts issued by the Kenya Meteorological Department have been translated, customized and disseminated to the Mbeere people via ITIKI prototype;

(iv) In collaboration with KMD, plans are under way to install and operationalize a minimum of 5 rainfall stations;

(v) Once funds are available, plans under way to install and operationalise a minimum of 5 sensor-based weather stations;

(vi) Use ITIKI to create a drought early warning system using data from (i), (iv) and (v) for Mbeere

5.3. Current Operational Structure

This is purely a community-based initiative and the Mbeere people (through the representatives in the focus group) will spearhead the project. As such, a Community-Based-Ogranisation; KOVCA (Kiritiri Orphans and Vulnerable Children Advocate) will be the main outfit to spearhead the implementation. An Advisory Board initially made up of an ITIKI's expert (Muthoni Masinde), a representative from KMD and a church (Catholic Church) representative has been set up to advise on the project.

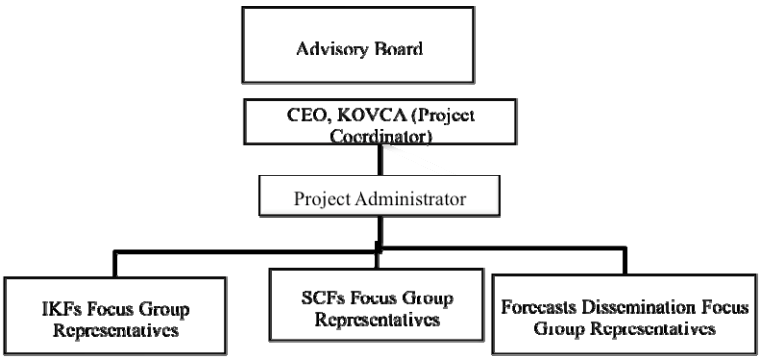


Figure 4. Project Operational Structure

6. CONCLUSION AND FURTHER WORK

Mbeere people occupy a semi-arid area that has hostile terrain and unpredictable rainfall. The main source of livelihood is subsistence crop farming and livestock keeping. So dependant is the community to rainfall that a good season translates to a healthy community and doom and misery otherwise. Modern weather forecasts are alien to the Community; there is no single weather station in the region. As such, the Community has continued to rely on IK to reach critical cropping decisions. With extreme weather variations being witnessed in the area and elsewhere in the globe, it has become increasingly difficult to use IK alone and a survey in the Community revealed dire need of reliable and relevant weather and drought forecasts. This made the authors reach a decision to pilot the recently developed ITIKI[3] framework and prototype among the Mbeeres. Starting off with a step-by-step explanation of ITIKI Architecture, we have explained the 'how', the 'what' and the 'when' of the pilot project currently underway.

We have just started; the rainfall stations will be installed before the next rain season due to start by mid-October 2012. For the first time, the Seasonal Forecasts by KMD have been translated into Kimbeere and widely disseminated to the Community. A follow-up study to find out if this improved the usage of the forecasts is scheduled for early 2013. Further, the number of weather parameters observed will be improved by the planned installation of sensor-based weather meters along-side the rainfall stations.

Lastly, during the development of ITIKI, it emerged that there is no single source known to the authors that provides a systematically collected and stored IK on weather. ITIKI implementation in Mbeere will provide one such information source.

REFERENCES

[1] Mugabe, F.,T., Mubaya, C.,P., Nanja, D., H., Gondwe, 2010. Use of indigenous knowledge systems and scientific methods for climate forecasting in southern Zambia and north western Zimbabwe. Zimbabwe Journal of Technological Sciences, 1 (1).

[2] Ziervogel, G., Bithell, M., Washington, R. And Downing, T., 2005. Agent-Based Social Simulation: a Method for Assessing the Impacts of Seasonal Forecast Application Among Smallholder Farmers. Agricultural Systems, 83 (1), pp. 1-26

[3] Masinde, M. And Bagula, A., 2012. ITIKI: bridge between African indigenous knowledge and modern science of drought prediction. Knowledge Management for Development Journal, In Press (2012), pp. 1-19.

[4] Masinde, M., Bagula, A. And Muthama, N., 2012. The Role of ICTs in Downscaling and Up-scaling Integrated Weather Forecasts for Farmers in Sub-Saharan Africa, In: The Fifth ICTD, March 12-15 2012, ACM Digital Library, pp. 122.

[5] Deely, S., David, D., Jorgelina, H. And Cassidy, J., 2010. World Disasters Report – Focus on Urban Risk. 1 edn. Geneva, Switzerland: International Federation of Red Cross and Red Crescent Societies.

[6] Masinde, M. And Bagula, A., 2011. The Role of ICTs in Quantifying the Severity and Duration of Climatic Variations – Kenya's Case, Proceedings of ITU Kaleidoscope 2011: The Fully Networked Human? - Innovations for Future Networks and Services (K-2011), 12-14 December 2011, IEEE Xplore, pp. 1-8.

[7] Kenya National Bureau Of Statistics, 2009. The Kenya Census 2009: Population and Housing Census Highlights. Government Press

[8] PANU, U.S. and SHARMA, T.C., 2002. Challenges in drought research: some perspectives and future directions. Hydrological Sciences-Journal.

[9] Byun, H. And Wilhite, D.A., 1999. Objective quantification of drought severity and duration. Journal of Climate, 12(9), pp. 2747-2756.

[10] ITU-T, 2008. ITU-T Technology Watch Briefing Report Series. 4, February. http://www.itu.int/dms_pub/itu-t/oth/23/01/T230100000040001PDFE.pdf: ITU.

[11] Jury, M., R., 2008. Predicting Climate Variability in Southern Africa. In: H. VIRJI, F. CORY, F. AMY and S. MAYURI, eds, Climate Variability , Water Resources and Agriculture Productivity: Food Security Issues in Tropical Sub-Saharan Africa. 1st edn. SCOWAR, pp. 375-380.

[12] BARRON, E. and SOROOSHIAN, S., 1997. Assessing the Impacts of Climate on Regional Water Resources. 12th Mission to Planet Earth Observing System Investigators Working Group Meeting 1997, pp. 1.

[13] LAU, L., YOUNG, R., A., MCKEON, G., SYKTUS, J., DUNCALFE, F., GRAHAM, N. and MCGREGOR, J., 1999. Downscaling global information for regional benefit: coupling spatial models at varying space and time scales. Environmental Modelling & Software, 14 (6), pp. 519-529.

[14] Roncoli, C., 2002. Reading the Rains: Local Knowledge and Rainfall Forecasting in Burkina Faso. Society & Natural Resources: An International Journal, 15 (5), pp. 409-427.

[15] Ghile, Y. And Schulze, R., 2008. Development of a framework for an integrated time-varying agrohydrological forecast system for Southern Africa: Initial results for seasonal forecasts. 34 No.3, July 2008. Sout h Africa: Water Resource Commission.

[16] Hansen, J., W., 2002. Realizing the potential benefits of climate prediction to agriculture: issues, approaches, challenges . Agricultural Systems, 74 (3), pp. 309-330.

[17] Sillitoe, P., 1998. The Development of Indigenous Knowledge: A New Applied Anthropology. Current Anthropology, 39 (2), pp. 223-252.

[18] Boef, W.D., Kojo, A., Kate, W. And Anthony, B., 1993. Cultivating Knowledge; Genetic Diversity, Farmer Experimentation and Crop Research. 1st edn. London: Intermediate Technology Publications.

[19] Mercer, J., Kelman, I., Taranis, L. And Suchet-Pearson, S., 2010. Framework for integrating indigenous and scientific knowledge for disaster risk reduction. Disasters, 34 (1), pp. 214-239.

[20] Ziervogel, G. And Opere, A., 2010. Integrating meteorological and indigenous knowledge-based seasonal climate forecasts for the agricultural sector: Lessons from participatory action research in sub-Saharan Africa. 2010. http://web.idrc.ca/uploads/user-S/12882908321CCAA_seasonal_forecasting.pdf: IDRC.

[21] Brokensha, D., W., Warren, D., M. And Oswald, W., 1982. Indigenous Knowledge Systems and Development. American Anthropologist, 84(3), pp. 671-672.

[22] Thrupp, L., A., 1989. Legalising local knowledge: from displacement to empowerment for Third World people. Agriculture and Human Values, 6 (1989), pp. 13-24.

[23] Flora, C., 1992. Reconstructing agriculture: the case for local knowledge. Rural Sociology, 57 (1992), pp. 92-97.

[24] Richards, P., 1993. Cultivation: knowledge or performance? In: M. HOBART, ed, An Anthropological Critique of Development: the Growth of Ignorance. 1st edn. London: Routledge, pp. 61-78.

[25] VIRJI, H., CORY, F., AMY, F. and MAYURI, S., 1997. Climate Variability, Water Resources and Agricultural Productivity: Food Security Issues in Tropical Sub-Saharan Africa. Workshop on Climate Variability Prediction: START/WCRP/OSTROM/SCOWAR.

[26] Richard-Van, C., Maele, Gaëlle, S. And Lisa, M., June,12, 2011-last update, Weather, Water And Climate Information Provide Early Warnings That Save Lives [Homepage of World Meteorological Organisation], [Online].Available [May 13, 2012]. http://www.wmo.int/pages/mediacentre/factsheet/Earlywarning_en.html

[27] ISDR, 2006. Developing Early Warning Systems: A Checklist. Third International Conference on Early Warning: From concept to action, EWC III, pp. 1-13.

[28] STERN, P.C. And WILLIAM, E., EASTERLING, 1999. Making Climate Forecasts Matter. Washington D.C.: National, Research, Council - National Academy Press.

[29] Fernando, D. ,A.,K.. And Jayawardena, A., W., 1998. Runoff Forecasting Using RBF Networks with OLS Algorithm. Journal of Hydrologic Engineering, 3 (3), pp. 203-209.

[30] ORLOVE, B., Roncoli, C., MERIT, K. and ABUSHEN, M., 2009. Indigenous climate knowledge in southern Uganda: the multiple components of a dynamic regional system. Climate Change, 100 (2), pp. 243-265.

[31] Jarraud, M., 2008. Guide to Meteorological Instruments and Methods of Observation (WMO-No. 8). 7th edn. Geneva 2, Switzerland: World Meteorological Organisation.

[32] EAC, S., 2008. Enhancing Capacities of the Meteorological Services in Support of Sustainable Development in the East African Community Region Focusing on Data Processing and Forecasting Systems. Arusha, Tanzania: East Africa Community Secretariat. ITU 2010; The World in 2010: ICT Facts and Figures; Available: www.itu.int/ITU-D/ict/material/FactsFigures2010.pdf

[33] Berkes, F. And Mina, K., Berkes, 2009. Ecological complexity, fuzzy logic, and holism in indigenous knowledge. Futures, 41 (1), pp. 6-12.

[34] Zadeh, L. A. 1965. Fuzzy sets,. Information and Control, 8 (1965), pp. 338-353.

A SUSTAINABLE INTEGRATED-SERVICES COMMUNITY LEARNING CENTER

Prasit Prapinmongkolkarn1, Supavadee Aramvith1, Chaodit Aswakul1,
Anegpon Kuama2, Sucharit Koontanakulvong3, Ekachai Phakdurong4
[1]Department of Electrical Engineering, Chulalongkorn University,
[2]Social Research Institute, Chulalongkorn University,
[3]Water Resource Research System Unit, Faculty of Engineering, Chulalongkorn University,
[4]THAICOM PLC, Bangkok, Thailand

ABSTRACT

This paper proposes the concept of integrating the community learning center with e-health facility and natural disaster warning system. The model for sustainability and ubiquity of ICT facilities in community has been achieved through three years of experiences in implementation of universal service obligation (USO) schemes in Thailand. From the beginning, the community learning centers have been designed with the principle of sustainability, flexibility, easy-to-use, cost saving and local participation concepts. With the country's lesson learned in the recent great flood last year and to prepare our country for future natural disasters, it is natural that the community learning center is proposed to extend its conventional services with real-time information and data service system for flood warning. This new service of the center can expectedly cowork with the conventional national television broadcasting, radio, mobile phone, satellite and amateur radio services. It is our belief that such integrated-services community learning center concept, the first of its sort, will enhance the education of people by bridging the digital divide in USO, to improve health care and wellness of people by telehealth service, as well as to make our country ready for unforeseen natural disaster crisis in the future.

Keywords— Community learning center, universal service obligation, telehealth, natural disaster warning, flood-risk evaluation system

1. INTRODUCTION

In the advent of globalization and rapid telecommunication technology development, various facets of telecommunication systems, providing telecommunication services - voice, data and video as well as information service system - have been developed to serve the needs of people both in urban and rural areas.

Thailand's National Broadcasting and Telecommunications (NBTC) realized the ICT needs of people and, since 2009, NBTC has implemented several universal service obligation projects throughout Thailand in order to bridge the digital divide and to provide equal access opportunity for people living in remote areas. The USO projects cover from telehealth centers, community public telephone and internet center, school internet and community learning centers. Ample experiences learned from the operation of these USO projects have led to conclude a sustainable functionality model of country's community learning center.

Innovative broadband models for digital inclusion have been explored, put into trial and actual use and evaluated recently in Thailand [1]. For instance, a telehealth system has been implemented in Phang Nga Provincial Hospital interconnecting with health center and community hospital in the remote four islands in Andaman Sea [2]. The implementation based on WiMAX and CDMA 2000 for inter-island communications was successful as a demonstration project and its sustainability as a service is being tested after the expiration of the contract of funding by NBTC. In addition, the IPSTAR, an IP-based broadband satellite, has been deployed both in distance learning and tele-consulting or remote diagnostics in rural areas effectively since a few years ago after its commencement of service in 2006.

A sustainable community has to be environmentally friendly and immune from or less affected by natural disaster. In recent years, Thailand and other Asia Pacific countries have been constantly affected by earth quakes, tsunami and big floods. Hence, most of the national regulators, for example, Ministry of Internal Affairs and Communications (MIC) and NBTC have initiated natural disaster warning systems operating in both on-line and offline modes at the central administration. To make our country ready for unforeseen natural disaster crisis in the future, and given the success of our ongoing community learning centers, we believe that some of the essential features available from the central database of the natural disaster warning system should be possibly disseminated to community learning center in the near future. The aim of this paper is to propose such a framework of sustainable integrated-services community learning center with various features of services made available at the center via its provided internet-enabled infrastructure.

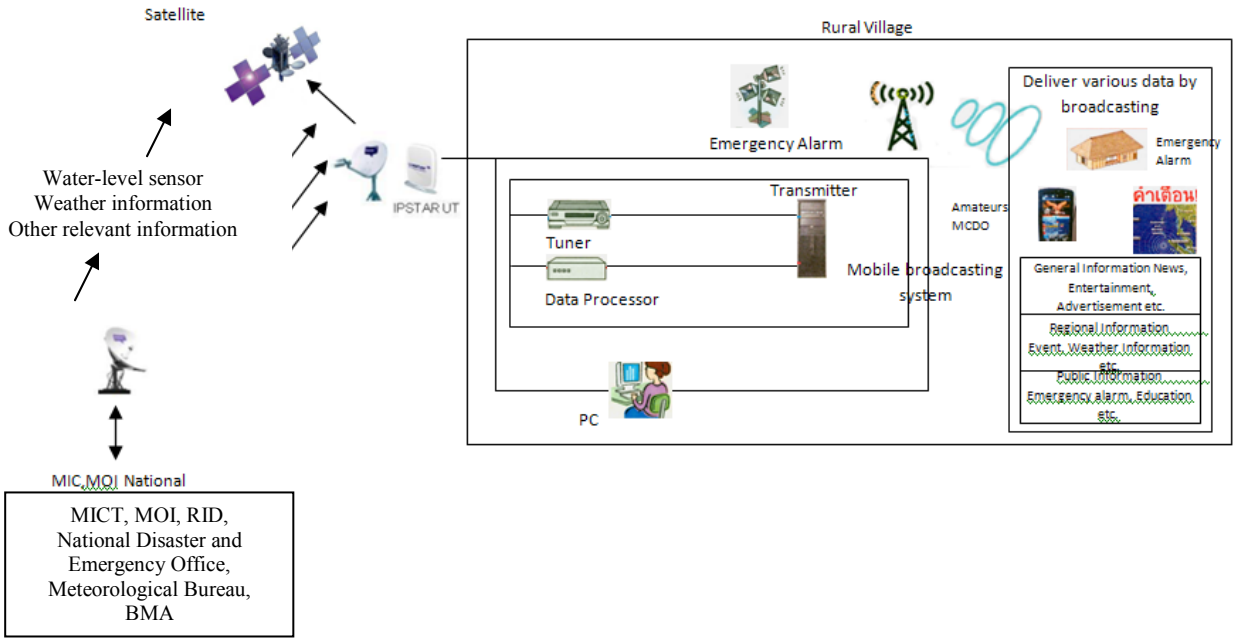


Figure 1. Overall architecture of integrated-services community learning center

The organization of this paper is as follows: overall system architecture of sustainable, integrated-services community learning center is presented in Section 2. Methodology and philosophy adopted in first setting up a community learning center have been found crucial to its future sustainable and here are described in Sections 3 and 4. In Section 5, our IPSTAR system -an alternative enabling technology for connecting the center- for distance learning and e-health applications is described. Section 6 summarizes on the flexible, easy-to-use and cost effective natural disaster warning system, here proposed to be integrated with the community learning center's ongoing services. Finally, concluding remarks are given in Section 7.

2. OVERALL SYSTEM ARCHITECTURE

The system consists of a sustainable community learning center as a central role with integrated-services functionality providing internet applications e.g. e-health, e-education via the IPSTAR broadband satellite communication (or other wired cheaper broadband medium whenever applicable) as depicted at the left hand side of Figure 1. In addition, at the right hand side as shown in Fig 1, the data and information service software for Flood Risk Evaluation system for Thailand (Flood REST) can be installed to use the same or separate server from community learning center depending on budget.

The Flood REST has been developed by the research team of Water Resource Research System Unit, Faculty of Engineering, Chulalongkorn University in collaboration with Royal Irrigation Department (RID) and Ministry of Information and Communication Technology (MICT) and Bangkok Metropolitan Administration (BMA) under the pressure of great flood starting from September 2011 until December 2011. The system comprises BMA Canal Monitoring System and flood level prediction from the Water Measurement System, and Thai Crisis Reporting System (MICT and Chulalongkorn University), which identifies the flood areas in special flood map reporting system. Such information can be integrated into the nation-wide network of our community learning center. Wherever possible, the developed system have been designed by open-source, ITU and IEEE standards. Standardization of necessary basis technologies is important to ensure that all components can function well together especially in our system which must be integrated from many different components.

3. ORIGIN OF COMMUNITY LEARNING CENTER

Currently, Thailand had still yet to gain experiences in rolling out the Internet centers to rural or non-economically justifiable areas. However, the country has been determined to change this situation. In 2009, based on the initiative of the former National Telecommunications Commission (NTC) and, presently the National Broadcasting and Telecommunications Commission (NBTC), Thailand's universal service obligation (USO) has been revived with the aim of providing equal access opportunities for people living in rural areas to reach the Internet and necessary information technology infrastructure.

Towards this aim, Chulalongkorn University has been commissioned to carry out the NBTC Community Learning Center trial project [3]. The objective was to learn of ways to prepare our country in the sequel scaling-up of building internet-enabled community learning centers throughout the whole Kingdom of Thailand. To test out novel ideas as well as to set up a role model of Thai community learning center, the research team members at the university as well as the team of the Office of NBTC have been responsible

for studying, designing, building and evaluating proper operational and management models for the sustainability of community learning centers.

In this regard, five pioneering community learning centers have been started up in Thailand's rural areas. These trial sites of community learning center have been located selectively in five provinces, namely, Lamphun, Chaiyapoom, Pichit, Suratthani, and Krabi, which cover all regions of Thailand. Hence, the study could take into account sample community representatives with varieties in physical/social constraints and needs. Thus, in carrying out the project, we planed our activities with three stages

Firstly, at each selected village, even before the learning center was built, dialogue exchanges and learning activities have been created and tried out together by our research team of Chulalongkorn University and by collaborative involvements from people in their own community. This stage has been proven essential to synergize the momentum of community dreams, which would play subsequent roles in provoking villagers' eagerness to learn of new technologies. Also this stage is concerned with the design and construction of physical infrastructure e.g. rooms, buildings, equipments, software and contents that are to be integrated into the whole learning center of the village. At the end of first stage, the community could then have the center at their village. This stage took about half a year.

Secondly, with the support and understanding of community leaders, the administrative committee has been elected by villagers to manage their own community learning center. The principle is to try to involve potential stakeholders and community leaders into this administration for a transparent management with good governance. In addition, the Holy Grail was to make villagers not only understand how to use the telecommunication and internet services available at the constructed community learning center. But also, and more importantly, the goal was to enable villagers with tools that in turn would expectedly enable their own unique local values.

With people becoming eager to learn, equipments ready for usages, the center's activities have been planned and evaluated. The success of those activities depend much on the level of villager participation at their community learning center not only as users, but also as owners of this common community facility. By feeling of ownership from villagers as a whole, the sustainability of the community learning center has been envisioned plausible. Activities have been created to reflect needs of local villagers, which differ from one center to another. And innovative ideas in utilizing the facilities in the center must be continuously invoked mutually amongst villagers.

Thirdly, after the whole year of activities run by individual centers, in the trial project, the central focus has been shifted towards how to create the network of centers in all learning centers. This is a crucial stage and utmost important for the sustainability of community learning center as a whole. And currently we are at this stage of building such network.

4. USO/UA ROLE OF COMMUNITY LEARNING CENTER

In the ITU context [4], the definition and scope of universal service is the choice that individual countries must be able to decide for themselves. That is, each member state has their own right to detail the suitability of their requirement for universal services. For developing countries, due to their limitations, the extent of universal services might yet to be realized. These differences in the country readiness have raised another important line of attempts under the general framework of universal access [5][6]. Particularly, the policy of universal access (UA) framework is aimed at a more readily realizable target in that everyone in the country must be able to have a service access, with a certain degree of convenience and to some basic telecommunication services like telephone service.

Based on the five pilot communities who have so far been able to start up their centers, we have learnt lessons worth sharing and disseminating. In contrast to the belief that UA must be the first step towards the elaborate USO regime for a country to modernize herself, from experiences in this trial project, the opposite school of thoughts has been found more practicable and better matching with the reality of our own people and expectedly the people in other developing countries.

We thought of USO as rather the preliminary step in bridging the digital divide between those people who have and those who have no chances in getting accesses to the telecommunication and internet service infrastructure. USO roadmap is often concerned with only how to increase the penetration rate of information communication infrastructure to people at large and too often the focus is on building physical infrastructures to ignite the USO dream.

In this trial community learning center project, we have tried at first that same line of approach as the USO paradigm with the project's focus on the internet aspect of USO. But later the findings suggested that it was relatively easy, given sufficient budgets, to build physically universal telecommunication services throughout the country. It is however much harder, once those services have been available, to bridge on another big gap of people's feelings.

Villagers with less educational backgrounds and no computer literacy often feel of and look at this new information technology from a distance with fears. Fear of not being able to learn new tools. Fear of humiliating themselves. Fear of breaking unintentionally the provided equipments. Only with those fears being overcome would the available internet services by USO be really made accessible to the people. For this reason in the paradigm of connecting not only equipments but people, we think USO is the preliminary step, and the universal access or UA should be viewed as the ultimatum.

We also learnt a great deal from the villagers' local wisdoms. Every community has their own strengths but, in many cases, without realizing that. We believed by networking people from different villages with different backgrounds, villagers would naturally learn of their own

uniqueness. Only by observing others could oneself be able to know better of one's inner beings. And such dialogues across the geographical barrier can be facilitated by the realm of internetworking those community centers.

In our project, the five pilot learning centers have gradually realized of their own strengths and started to build activities of their community centers around those. For instance, the center at Krabi has come to realize their local tourism activities as well as their village handicraft products. With a long history of community, filled with told legends, our community learning center at Ta Tong in Suratthani has set their goal in extending their activities towards the realization of their local museum of historical artifacts. Other centers share similar enlightened findings of their own unique values.

By strengths of local communities and activities enabled by new technology, the community learning center is believed to open up a new gateway of knowledge to help Thailand prepare its best at its roots towards the country's role in the ASEAN and globalised future. And currently, based on the success stories in the trial project, Thailand is now this year on her way in building up about nine hundred community learning centers more throughout the whole country. Figure 2 summarizes the frame of thoughts used in building the country's sustainable national learning society from network of the community learning centers being built.

Figure 2. Framework of sustainable national learning society

5. E-HEALTH ROLE OF COMMUNITY LEARNING CENTER VIA BROADBAND SATELLITE IPSTAR

Learning society cannot be realized easily unless the community people are in good health. However, in Asia Pacific Region, most of the population is living in rural areas where medical facilities and care are either inadequate or non-existing. Many hospitals, health centers and medical schools are located in cities but the lack of network connectivity and centralized medical database makes it difficult to provide effective service to people in rural areas. In this environment, IPSTAR, an IP-based broadband satellite can provide healthcare service to rural villages at very low cost

From the health center and clinic in rural areas, tele-consulting and VDO conference via IPSTAR can be used to access to a city hospital and to a medical school for training and education and to provide home telehealth to villagers at their community learning center as shown schematically in Figure 3**.** The system comprises IPSTAR user terminal and 0.84 meter outdoor antennas. The service fee for small IPSTAR UT at community learning center is about 200 US$ per month

The e-health equipment can be either what we have developed as a simple personal computer with a high precision camera or more complicated equipment like special remote diagnostic terminal with tele-consulting, and video conference services. The plausibility of the operations of telehealth systems is based on the efficient management of the supervisory consulting terminal and the availability of consulting physicians at the main hospital which interconnects with telehealth centers in the rural areas. The wellness support platform using the mobile terminal has also been proposed to simplify the management of data for the users while reducing the cost of data collection for healthcare service [7].

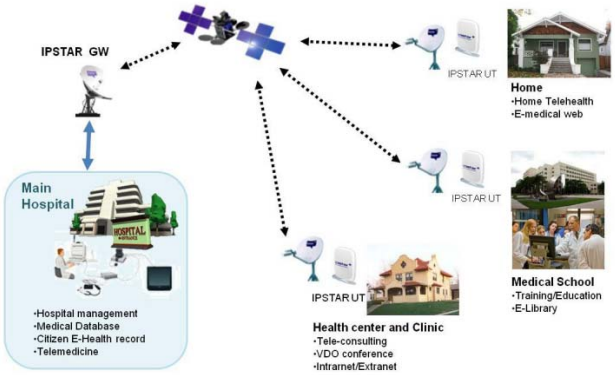


Figure 3. IPSTAR for E-health services

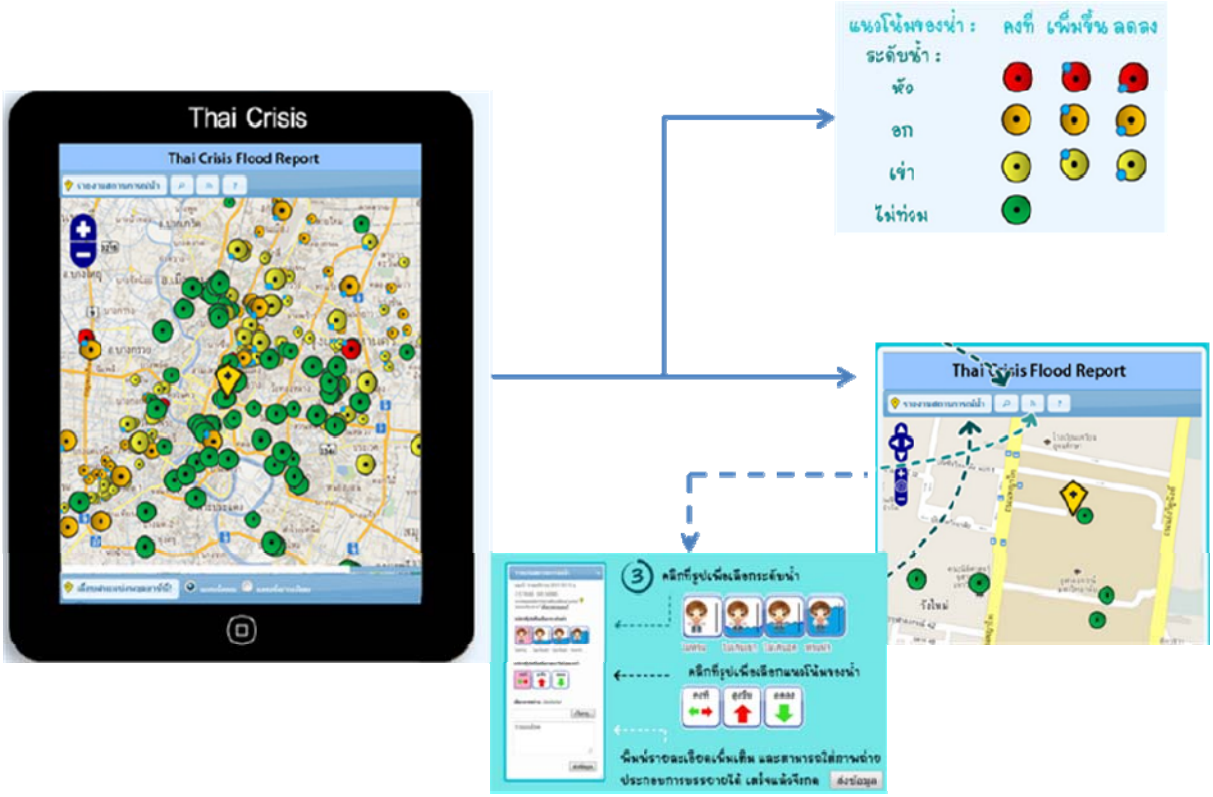


Figure 4. Tablet showing application of Flood REST

6. FLOOD AND NATURAL DISASTER WARNING ROLE OF COMMUNITY LEARNING CENTER

Flood Risk Evaluation System for Thailand (Flood REST) [8][9] mentioned in Section 1 uses the market available, easy-to-extend platform of "Google Map". On top of this platform, Flood REST employs the ground elevation data and the real-time monitored water level data from runoff stations in the relevant vicinity to fit into the Flood REST model, which constantly predicts flood level dynamics in the area of interest. The model output available for access by people ranges from the expected water level in canals, on roads, above the ground height, all of which information is essential for warning people if their area will be likely affected by the flood.

In Social Flood Map Monitoring System and Thai Crisis Reporting System, maps are marked by circles with different colors or grey-scale intensity (in black and white) to distinguish the three steps of flood conditions with an example shown in Figure 4. In this map, water level measured from water elevation marks and the pre-installed level sensor network are displayed to indicate easy-to-understand pictures of different flood levels, namely, head-level flood, breast-level flood, knee-level flood and no-flood conditions. Such data is displayed location-wise so that responsible officials, people, and rescuer volunteers on the flood sites can be best informed.

Most of the flood information systems are based on remote sensing (satellite) digital imagery, photogrammetry and field survey or sensor system-level monitoring. Data is in a standard on-line format and the output display devices can be both mobile handsets and computer tablets. The tablet is more appropriate than a computer notebook for water resource management on site of emerging floods for various merits, e.g., with smaller size and lighter weight, sharper display, touch screen, picture rotatable, easy to present in subgroup, long-standing battery and more viable applications. Tablet can be operated in on-line and offline mode where in the offline mode the tablet can offer file based GPS location and area map. Such application can also be made accessible through the network and computer equipment at the constructed community learning center for wider dissemination of useful data to villagers at large.

7. CONCLUSIONS

In this paper, we have presented a model of sustainable integrated-services community learning center which combines the community internet, telehealth/telemedicine, distance learning and flood/natural disaster warning system within the same unified managerial paradigm. The sustainability of this community learning center begins with the design phase that encourages local people to actively participate in determining the requirements and design of facilities and infrastructure to be established at the center. Then the local value and sense of ownership must be

upheld. To improve the sustainable usefulness, all the community learning centers nation-wide should be interconnected via activities, projects, local needs, and national directives. Training of local people is another important activity and must be regularly carried out in order to allow the community learning center to grow its fruitfulness towards not only its by-design functionality of mere internet service provision, but also its by-local-invention functionality of genuine healthy and well-informed learning society. Only with that goal would the full potential of information technologies be realized by the people, who would gradually and steadily apply the technologies in their agricultural planning, marketing, life-long education, health-care improvement as well as to better equip themselves with unforeseen threatening natural disasters.

At this early stage at least, wherever appropriate, open-source, ITU and IEEE standards have been used in software and hardware development. The community learning center is being developed to set up its own website in parallel to the nation-wide next-generation mobile phone infrastructure roll-out. People are being educated and more especially the young and the olds are eager to learn of the new technologies. We believe there is a long journey towards the country's sustainable integrated-services community learning center network. But with the right frame of thoughts and our learned lessons in the past trials, we have a promising hope of actual realization for economic gains and well-beings of the community as a whole.

ACKNOLEDGMENTS

The authors would like to express their sincere thanks to the National Broadcasting and Telecommunications Commission of Thailand for supporting the USO projects cited in this paper and to the Ministry of Internal Affairs and Communications (MIC), Japan, and the Ministry of Information and Communication Technology, Thailand, for helpful discussions

REFERENCES

[1] S. Aramvith, P. Prapinmongkolkarn, A. Kongchanagul "Innovative Broadband Models for Digital Inclusion," Proceedings of ITU Kaleidoscope 2009, Mar del Plata, Argentina, 31 August – 1 September 2009, pp.117-123.

[2] S. Aramvith, P. Prapinmongkolkarn and T. Chalidaphongse, "Wireless Communications for Health Applications and Services in Rural Thailand," Proceedings of ISMAC 2001, Manila, Philippines, September 7-9, 2010.

[3] C. Aswakul et. al., Chula-Unisearch Final Report, NTC Learning Center Project, 2011 (in Thai).

[4] ITU-D Case Library of Rural Application Focus Group 7, http://www.itu.int/ITU-D/fg7/case_library/categories.asp

[5] A. Pentland, R. Flentcher and A. Hasson, "DakNet: Rethinking Connectivity in Developing Nations," Computer Magazine, Institute of Electrical and Electronic Engineers, January 2004.

[6] R. Rajora, "Bridging the Digital Divide: Gyandoot, the Model for Community Networks," TATA McGraw Hill, New Delhi, 2002.

[7] H. Takeim S. Horiguchi, T. Shimizu, Y. Hayashi, A. Takahashi and Y. Tomizawa, "Wellness Support Platform Using Mobile Terminals," NTT DOCOMO Technical Journal, Vol. 11, No. 2, pp. 9-16.

[8] D. Jampanil, S. Koontanakulvong and S. Sakulthai, "Community Based Participation in Integrated Water Resources Development: Lessons learned from Case Study of Rayong Province, Thailand," Proceedings of the 7th International Symposium on Social Management Systems, , Colombo, Sri Lanka, 14-16 September, 2011.

[9] S. Koontanakulvong, el. al., Flood Risk Evaluation and Flood Reporting System via GIS and Online Network, Research Report, submitted to Chulalongkorn University, Mar 2012 (in Thai).

SESSION 4

RESOURCE DISCOVERY AND MANAGEMENT

S4.1 System design and numerical analysis of adaptive resource discovery in wireless application networks

S4.2 Design and Implementation of virtualized ICT resource management system for carrier network services toward Cloud computing era

S4.3 Harmonized Q-Learning For Radio Resource Management In LTE Based Networks

SYSTEM DESIGN AND NUMERICAL ANALYSIS OF ADAPTIVE RESOURCE DISCOVERY IN WIRELESS APPLICATION NETWORKS

Wei Liu, Takayuki Nishio, Ryoichi Shinkuma

Kyoto University
Graduate School of Informatics
Kyoto, Japan

ABSTRACT

In this paper, we propose an adaptive resource discovery method in heterogeneous wireless application networks. The adaptive method uses either centralized mode or flooding mode to discover available resources according to different network status. The proposed adaptive method is used to reduce energy consumption in resource discovery process. We establish theoretical energy model for both modes. A heuristic algorithm is designed to implement the proposed adaptive method. It is also proved to be energy efficient through extensive evaluations.

Keywords— adaptive resource discovery, energy efficient, heterogeneous wireless application networks

1. INTRODUCTION

In last few years, wireless communication technologies have been developed a lot. Different kinds of wireless networks like 3G, Bluetooth, WLAN (Wi-Fi) and WiMAX converges into a unified heterogeneous wireless network. As a result, user can choose different network according to different requirements. At the same time, hardware and software technologies enable modern wireless terminals to integrate more types of resource compared with before, e.g., computation resources, communication resources, sensor resources, software application resources and so on. Detailed definition and categorizing of resource can be found in [1]. Through wireless communication, resources distributed in different node can be connected dynamically and utilized opportunistically. We call the emerged wireless resource platform as wireless application networks. In order to enhance functionality and improve performance, nodes in wireless application networks could share different kinds of resource with each other.

For example, in Figure.1, the central node only owns 3G resource. It can borrow humidity sensor resource and GPS resource from other nodes to enhance its functionality. It can also borrow more 3G resource to improve the quality of communication. However, in order to realize this idealized picture, there still needs a way to discover available resources in the surrounding nodes. This paper introduces an energy efficient adaptive resource discovery method to solve this problem.

Many research works about resource discovery have been published [2, 3, 4, 5, 6, 7, 8]. However, existing works are not able to adapt their mechanisms based on different network status. Apart from that, more resources consume more energy. Battery capacity becomes a bottleneck of wireless applications. As a result, more and more research works [6, 7, 8] aim at providing energy efficient resource discovery method. However, there are two main problems in existing works: (1) most of them saved energy through sacrificing other important quality metrics like correctness and coverage of response without providing a formal quantitative analysis; (2) they only considered resource discovery and energy consumption in a homogeneous network like 3G cellular or WLAN ad-hoc alone. Obviously, energy consumption in heterogeneous networks is more realistic and important in modern community.

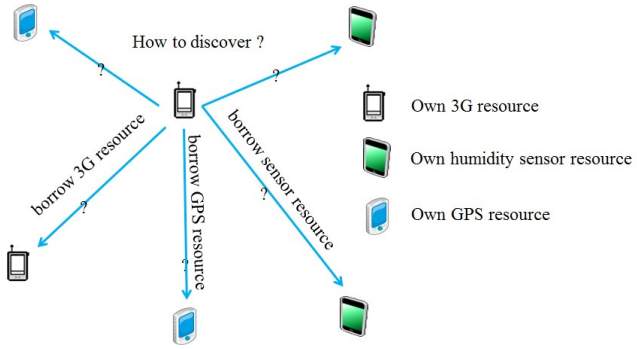


Figure 1. Application scenario

In this paper, we proposed an adaptive resource discovery method which transforms between two different resource discovery modes. Contributions of this paper include: (1) According to our best knowledge, this is the first research work introduce adaptive resource discovery solution based on method transforming(radical tuning defined in [9]) and the first work considers about resource discovery in the background of heterogeneous wireless application networks. (2) The proposed adaptive method is used to reduce energy consumption in resource discovery process, theoretical energy and response quality models are established. (3) A heuristic algorithm is designed to implement the adaptive method and proved to be efficient through evaluations.

In the rest of this paper, we introduce assumptions and system model briefly in section 2. Adaptive method and energy models are analyzed in section 3. In section 4, the heuristic algorithm is introduced. In section 5, we verify our adaptive method through evaluations. Related works are discussed in section 6. In the last section, conclusions and future works are given.

2. SYSTEM MODEL

In this paper, we assume that wireless nodes are in wireless application networks including both 3G cellular and WLAN ad-hoc network like Figure.2 shows. In 3G network, there is a central base station which is able to communicate with all the nodes. We call it Central Resource Broker (CRB) in the background of resource discovery. Apart from that, every node can communicate with nearby nodes through WLAN ad-hoc network. The assumed communication abilities are common for current smart phones or other wireless devices. To discover resources, nodes either maintain a resource repository in CRB through widely-covered 3G or flood resource requests in the area through short-ranged WLAN ad-hoc.

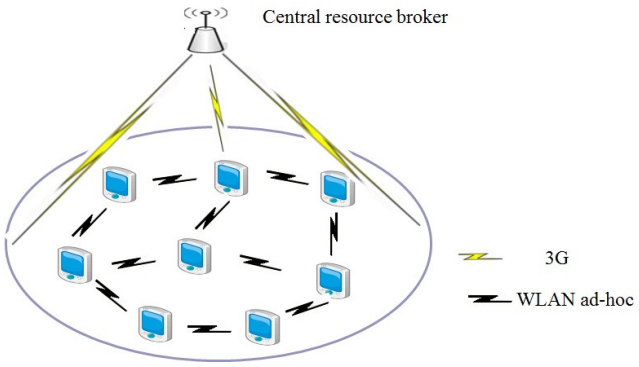


Figure 2. System architecture

Different method consumes different amount of energy. In this paper, we try to minimize energy consumption through transforming between two methods according to network status while maintain the expected resource information availability (defined in 3.2) be higher than a threshold R_{thresh}. The problem is defined as:
Objective: Minimize energy cost in resource discovery
Constraint: E[Resource Information Availability] $\geq R_{thresh}$

3. ADAPTIVE METHOD AND MODEL DEFINITION

3.1. Adaptive method

As described in section 2, nodes can discover available resources through 3G and WLAN ad-hoc networks. There are two basic modes in the defined wireless application networks:

(1) Centralized mode: In this mode, all available resource information is stored in CRB. If a node wants to allocate resource, it first checks whether the resource is available in itself. If not, it sends a resource request to CRB through 3G. CRB returns the identifications of nodes in which required resources are available.

(2) Flooding mode: In this mode, no resource information is maintained in CRB. If a node wants to allocate resource, it checks whether the resource is available in itself. If not, it floods resource request in the area through WLAN ad-hoc.When a node which owns the required resource receives the request, it sends its identification back to the requestor to notify available resources

In the proposed adaptive method, resource discovery method transforms between centralized and flooding modes automatically according to network status. At first, time is divided into consecutive time slots. Then at the end of each time slot, nodes send statistics to CRB according to their experiences in the last time slot. CRB estimates energy consumption for both modes based on collected statistics, chooses energy efficient method and notifies each node to use it in the next slot.

3.2. Resource information availability and maintenance

Available resources in one node would change during time. It is affected by factors like task processing, environment change, remaining battery power and so on. In this paper, we only consider about task processing factor which is generated from two sources:

(1) Nodes allocate resource for task from itself;

(2) Nodes allocate resource for task from other nodes;

Task processing occupies resources and the end of processing releases occupied resources.

In this paper, Resource Information Availability (RIA) which reflects the quality of responded information is defined as: The possibility that the response to a request includes all available resource information correctly. It includes two aspects: correctness and coverage. Energy is consumed to maintain RIA.

3.2.1. RIA in centralized mode

In the centralized mode, because of the widely coverage area of 3G, all nodes can send their resource information to CRB. Coverage of RIA is perfectly maintained. However, when available resource changed, nodes should update resource information stored in CRB to maintain the correctness of RIA. Energy consumption for RIA maintenance is analyzed. First, we consider about energy consumption for maintaining one type of resource named A. All model parameters listed in Table.1 are for one time slot.

The processing time for one task request with all resource A in the nodes is:

$$T' = \frac{S}{R} \tag{1}$$

The capacity of all the nodes which shows the maximum number of task requests for resource A that can be processed

Table 1. Parameters for RIA maintenance

S	expected task size of a request for resource A
R	sum of resource A in all nodes
λ_{A-o}	number of generated task for resource A from other nodes
λ_{A-s}	number of generated task for resource A from itself
λ_A	number of generated task for resource A $(\lambda_{A-o} + \lambda_{A-s})$
T	length of time slot
T$^{'}$	processing time for one task with all resource A
N	number of task can be processed
$F_{A-regist}$	expected number of updating for resource A
F_{regist}	expected number of updating for all resource types

is:

$$N = \frac{T}{T'} \tag{2}$$

Depending on the relationship between A and N, there are two situations:

(1) When $\lambda_A > N$, N determines $F_{A-regist}$:

$$F_{A-regist} = 2 \times N = \frac{2 \times R \times T}{S} \tag{3}$$

(2) When $\lambda_A \leq N$, λ_A determines $F_{A-regist}$:

$$F_{A-regist} = 2 \times \lambda_A \tag{4}$$

The factor of 2 indicates updating is needed at both allocating and releasing time of resources.

Obviously,F_{regist} is the sum of updating for all types of resource:

$$F_{regist} = \sum F_{I-regist} \tag{5}$$

where I stands for different resource type

3.2.2. RIA in flooding mode

In the flooding mode, on-demand flooding is adopted and nodes are assumed to be uniformly distributed in the area. For on-demand flooding, nodes know their available resources well when they receive a request. Correctness of RIA is perfectly maintained. However, the requestor has to set a large enough TTL value to ensure request packets arrive at every node in the area in order to maintain the coverage of RIA.

To estimate proper TTL value,we define P_i as the expected percentage of newly found nodes in the i-th hop, e.g., P_1 = 40% in Figure.3 since 40% of nodes are found with TTL increases from 0 to 1. Specifically, we define $P_0 = 1/N_{node}$ which means the percentage occupied by the requestor itself, where N_{node} is the number of nodes in the area. As a result, TTL value k should satisfy the following equation asymptotically in order to provide a complete coverage:

$$\sum_{i=0}^{k} P_i = 1 \tag{6}$$

where k is the minimum value satisfying the equation. Detailed method of calculating P_i is described in [10].

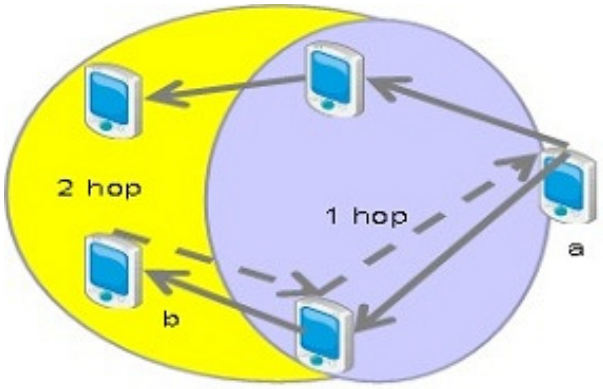


Figure 3. RIA in flooding mode

3.3. Tradeoff between RIA and energy consumption

In the previous analysis, we assume RIA is perfectly maintained. However, nodes may be tolerant to lower RIA in order to save energy. In this section, we analyze the tradeoff between these two metrics.

3.3.1. Tradeoff in centralized mode

In section 3.2.1 a node updates resource information right after available resource changed. Instead of that, it could wait for a period of time before updating. With this strategy, because F_{regist} becomes smaller, less energy is consumed. As

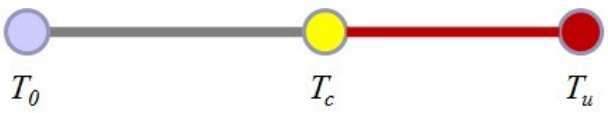


Figure 4. Tradeoff in centralized mode

Figure 4 shows, T_0 is the time of previous resource updating; T_c is the time when the available resource changed. The node doesn't update information in CRB until T_u. Requests generated between T_c and T_u would be responded with outdated information by CRB. If T_u is further from T_c, the expected RIA from T_0 to T_u becomes lower. On the other hand, if there is another change of resource between T_c and T_u, only one updating is executed other than 2 in perfect case. As a result, F_{regist} becomes lower, less energy is consumed.

In the following analysis, resource requests from other nodes are assumed to be a Poisson arriving process. Let H represents a random variable with definition:

$$H = \begin{cases} 1 & \text{response to a request is correct} \\ 0 & \text{response to a request is outdated} \end{cases} \tag{7}$$

We need to find the expected value E[H] for one request:

$$E[H] = \int_0^T f(t) \times E[H|T^{'} = t]dt \tag{8}$$

T is the length of period from T_0 to T_u. $T^{'}$ is a random variable of T_c and $f(t)$ is the probability density function of $T^{'}$. If we assume Poisson process as the request arrival model, the arriving of a request uniformly distributes in T conditioned on a request happens within this period.

$$f(t) = \frac{1}{T} \tag{9}$$

$E[H|T^{'} = T_c]$ is calculated as:

$$E[H|T^{'} = T_c] = \frac{T_c - T_0}{T_u - T_0} \tag{10}$$

Let $\alpha = (T_u - T_c)/(T_c - T_0)$. E[H] is computed as:

$$E[H] = \frac{T_c - T_0}{T_u - T_0} = \frac{1}{1+\alpha} \tag{11}$$

Nodes can keep α constant through choosing T_u according to T_0 and T_c. In order to keep expected RIA larger than R_{thresh}, α should satisfy inequality:

$$\alpha \leq \frac{1 - R_{thresh}}{R_{thresh}} \tag{12}$$

The resulted expected updating number is:

$$F_{regist}^{'} = \frac{F_{regist}}{1+\alpha} \tag{13}$$

3.3.2. Tradeoff in flooding mode

In section 3.2.2, the TTL value of request is assumed to be large enough to cover every node in the area. Larger TTL value means a larger probability to discover required resources. But larger TTL value also consumes more energy to relay the request. Users may also agree with a smaller TTL value to save energy with lower expected RIA.
The evaluation method is nearly the same as before except that only the following inequality needs to be satisfied:

$$\sum_{i=0}^{k} P_i \geq R_{thresh} \tag{14}$$

where k is the minimum value satisfying the inequality.

3.4. Energy consumption model

In this section, we analyze energy consumption model of two modes. We define N_{resp} as the average number of response to one request.λ_o is the number of generated task from other nodes for all types of resource.
The energy consumption in the centralized mode can be calculated as:

$$\begin{aligned} E_{central} =& (\lambda_o \times E_{3G-trans}) + (N_{resp} \times \lambda_o \times E_{3G-recv}) \\ &+ (F_{regist} \times E_{3G-trans}) \end{aligned} \tag{15}$$

$E_{3G-trans}$ and $E_{3G-recv}$ are the energy consumption of transmission and receiving through 3G. The total consumption of $E_{central}$ includes three parts: sending resource request to CRB $\lambda_0 \times E_{3G-trans}$; receiving resource response from CRB $N_{resp} \times \lambda_0 \times E_{3G-recv}$ and maintaining RIA $F_{regist} \times E_{3G-trans}$.
In the flooding mode, we first notice that the expected out degree d of a node is equal to the number of newly found nodes in the first hop:

$$d = P_1 \times N_{node} \tag{16}$$

The average distance H from resource requestor to the provider is:

$$H = \sum_{i=0}^{k} P_i \times i \tag{17}$$

In this mode, nodes rebroadcast a request when receive it first time and TTL is larger than 0. Otherwise, nodes discard it. W_{trans} and W_{recv} are energy consumption of transmission and receiving through WLAN ad-hoc. Because every newly found node rebroadcasts the request except for TTL=0, energy consumption for request relaying is $(\sum_{i=0}^{k-1} P_i \times N_{node}) \times W_{trans}$; All neighbors of the relaying node receive the request, the energy consumption of receiving resource request is $(\sum_{i=0}^{k-1} P_i \times N_{node}) \times d \times W_{recv}$; The energy consumption of relaying and receiving response is $N_{resp} \times (W_{trans} + W_{recv}) \times H$. The energy consumption of one request in the flooding mode is:

$$\begin{aligned} E_{flooding}^{'} =& \sum_{i=0}^{k-1} P_i \times N_{node} \times W_{trans} + \sum_{i=0}^{k-1} P_i \times N_{node} \\ &\times d \times W_{recv} + N_{resp} \times (W_{trans} + W_{recv}) \times H \end{aligned} \tag{18}$$

The energy consumption of all requests in a time slot is:

$$E_{flooding} = \lambda_o \times E_{flooding}^{'} \tag{19}$$

4. HEURISTIC ALGORITHM

In this section, we implement the proposed adaptive method through a heuristic algorithm. According to energy consumption models, three statistics are needed to estimate energy consumption of two modes in a time slot, these are:

(1) Number of resource request for other nodes λ_o
(2) Number of resource updating F_{regist}
(3) Average number of response for each request N_{resp}

At the end of each time slot, distributed nodes send three statistics to CRB according to their experiences in the time slot. CRB processes raw data (e.g., summing up λ_o from each node) and stores records of previous N_{slot} time slots in a list. After initialization,"Check" part tests whether network status is too dynamic to be predicted. If there are N_{trans} transitions of discovery method in the kept records, network status is assumed to be unpredictable. Centralized mode is used until network status become relatively regular.

In "Large_diff", if the energy consumption ratio of two methods is less than C_{thresh} the better method is chosen directly. If all conditions above are not satisfied, CRB uses average value of previous N_{trend} records to predict network status in "Prediction" part. CRB chooses energy efficient method and sends decision to every node. If the transforming is from flooding to centralized mode, nodes should update their resource information at first to ensure RIA. The framework of the algorithm is shown below:

heuristic_algorithm ($N_{slot}, N_{trans}, C_{thresh}, N_{trend}, P_{thresh}$)

Initialize:
  preprocess raw data;
  create new record r;
  insert r into list and delete outdated record;

Check:
  if (there are N_{trans} or more transitions in the record list)
    choose centralized mode;
    goto Make_choice;

Large_diff:
  if (better consumption / worse consumption $\leq C_{thresh}$)
    choose better method;
    goto Make_choice;

Prediction:
  //X_{i-j} means statistic X_j in the i-th record
  $\lambda_o = \frac{\sum_{i=1}^{N_{trend}} \lambda_{i-o}}{N_{trend}}$

  $F_{regist} = \frac{\sum_{i=1}^{N_{trend}} F_{i-regist}}{N_{trend}}$

  $N_{resp} = \frac{\sum_{i=1}^{N_{trend}} N_{i-resp}}{N_{trend}}$

  use λ_o , F_{regist}, N_{resp} to estimate energy consumption;
  if (better consumption / worse consumption $\geq P_{thresh}$)
    don't change method;
  else
    choose better method;

Make_choice:
  send decision to every node;

In the heuristic algorithm, N_{slot} defines how long the records of passed time slots are kept. It depends on the characteristics of network status. The ratio of N_{trans} and N_{slot} shows the dynamic degree of network status. This is used to prevent meaningless transition which wastes energy. C_{thresh} is just a threshold. If the discrepancy between two methods is quite large, it is highly probable that the current method should be remained except for unpredictable heavy status change. N_{trend} reveals the smoothness of network status change. In wireless network, the N_{trend} should be kept small to ensure quick reaction to status change. Because transforming also consumes energy, P_{thresh} prevents transforming when the difference of two methods is small.

5. NUMERICAL EVALUATION

In this section, we verify our adaptive method and heuristic algorithm through numerical evaluations. In the evaluation, we assume nodes are uniformly distributed in a rectangular area. Nodes can discover available resources through 3G in the centralized mode or WLAN ad-hoc in the flooding mode. We compare energy consumption of our adaptive method with pure centralized and flooding modes. Since the effectiveness of the adaptive method comes from method transforming, improving centralized or flooding mode itself is not our focus. As described in section 6, the adaptive method also benefits from the integration of improved version of both modes.

In the following evaluations, 1 unit of resource can process one task within a time slot. As a result, 2 units of resource process one task within half of a time slot and so on. For energy consumption, since only the comparison among different methods is concerned, we adopted result in [11] where the ratio of 3G and WLAN ad-hoc is $20/1$ without a specific unit. All nodes generate task request with an equal probability. Unless clarified specifically, the adaptive method chooses centralized mode when it starts up and the expected RIA is 0.95 for both modes. The heuristic algorithm is initialized with $< N_{slot} = 5, N_{trans} = 3, C_{thresh} = 0.5, N_{trend} = 1, P_{thresh} = 0.95 >$. Details of parameter selection are out of the scope of this paper. It relies on the principles described in section 4 and network characteristics. Unmentioned parameters are listed in Table 2.

Table 2. Parameters for evaluation

Rectangular area	$1000m \times 1000m$
Number of Nodes	100
WLAN ad-hoc range	$250m$
3G transmit $E_{3G-trans}$	20
3G receive $E_{3G-recv}$	10
WLAN transmit W_{trans}	1
WLAN receive W_{recv}	0.5
Resource information size	1 KB
Response identification size	0.1 KB

5.1. Relationship between resource and request

We first show the relationship between resource amount and request number. Intuitively, if there are more resources in each node, there is less chance for it to ask for resource in other nodes. It means little energy would be consumed in the flooding mode. However, for centralized mode, the amount of available resource in a node changes even if the task is processed by itself. Energy will still be consumed for RIA maintenance. As resource amount grows, more requests are needed to cover the maintenance cost of CRB. This is proved by Figure.5-7 where the cross-point of request number grows with resource amount.

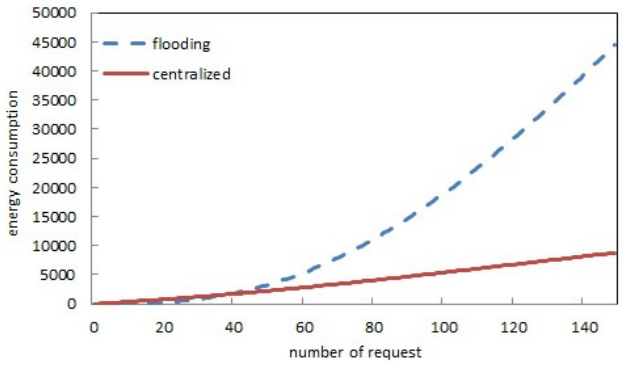


Figure 5. Relationship between resource and request 2-units

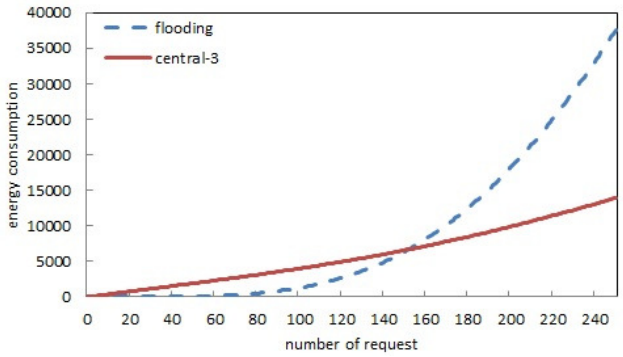


Figure 6. Relationship between resource and request 4-units

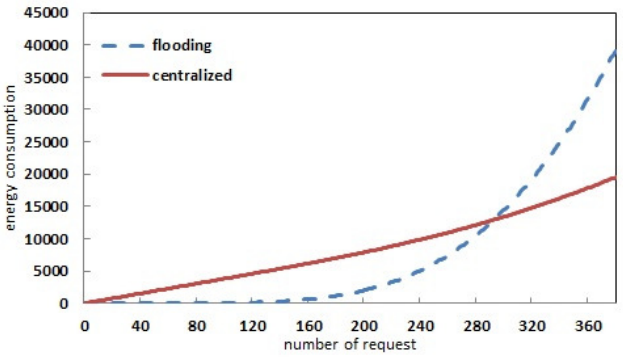


Figure 7. Relationship between resource and request 6-units

5.2. Performance in extreme situations

Then we verify the performance of our adaptive method when the network is in extreme situations. There are three extreme situations:

(1) Flooding mode only;
(2) Centralized mode only;
(3) Too dynamic to be predictable;

In this evaluation, every node owns 4 units of resource. For situation (1) the number of request is always 80 in one time slot. For situation (2) the number of request is always 200 in one time slot. For situation (3) the number of request increases from 80 to 200 in one time slot and decreases back to 80 in the next. According to Figure.6 when the number of request is 80, flooding mode is preferred. If the number of request is 200, centralized mode is better. Figure.8-10 show the energy consumption of three methods in 20 time slots of three extreme situations.

As figures show, the energy consumption of our adaptive method is near to the better choice in all three situations. The small difference comes from both prediction errors and transforming consumption. As time goes on, the energy consumption of adaptive method converges to the better choice.

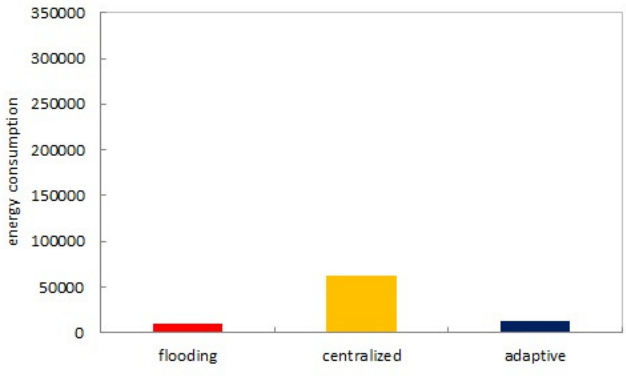


Figure 8. Energy consumption in flooding mode only

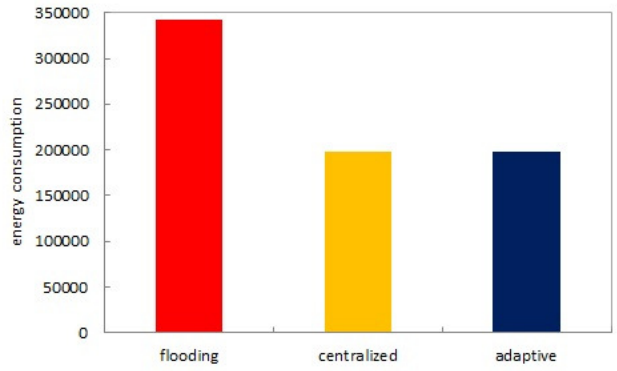


Figure 9. Energy consumption in centralized mode only

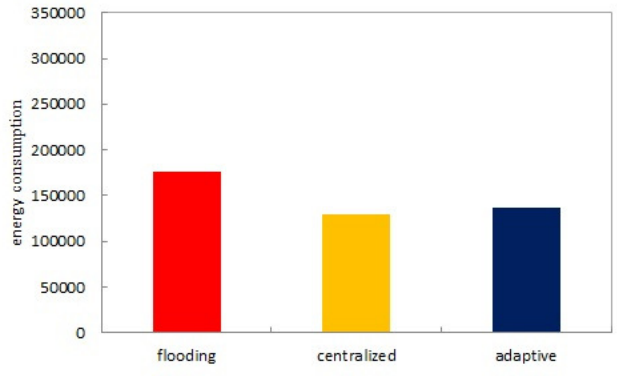


Figure 10. Energy consumption in extreme mode only

5.3. Comprehensive examples

In this subsection, we use two comprehensive examples to show the effectiveness of our adaptive method. In example 1, there is one type of resource named A. Every node owns 6 units of resource A. For network state s_1, 200 requests are generated in one time slot. For state s_2, 400 requests are generated in one time slot. According to Figure.7, s_1 prefers flooding mode and s_2 should choose centralized mode. We assume that network stays in s_1 for 10 time slots. Then the number of request increases 100 in each time slot until reaches s_2. It stays another 5 time slots in s_2 before come back to s_1 in reverse direction. This process has a period of 17 time slots. Energy consumption of different methods in one period is shown in Figure.11. For the idealized consumption, we assume CRB knows future network status well and always make correct prediction. Our adaptive method performs better than centralized (77.1%) and flooding modes (58.1%). It is near to the idealized consumption (102.1%). The discrepancy between adaptive method and idealized situation is due to prediction errors.

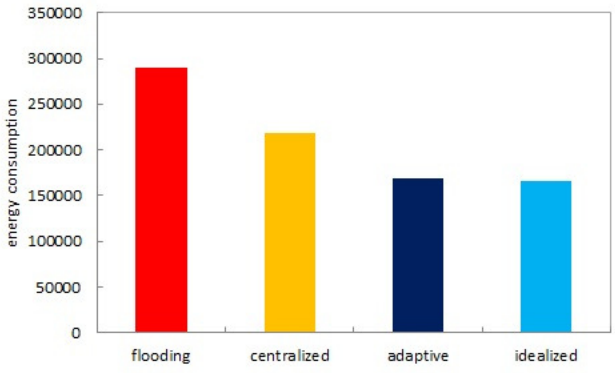


Figure 11. Energy consumption of example 1

In example 2, there are two types of resource A and B. Each node owns 2 unit of resource A and 6 units of resource B. In reality, resource A represents scarce resource e.g., sensor or GPS resource. Resource B represents common resource e.g., 3G resource. For state s_3, there are 70 requests for resource A and 10 requests for resource B. For state s_4, there are 10 requests for resource A and 70 requests for resource B. For example, in reality, during working hours, there are more needs for resource A, in s_3. But in spare time, there are more needs for resource B, so in s_4. The network stays in s_3 for 5 time slots. Then it changes to s_4 for another 5 slots and continues this process forever.

The energy consumption for different methods in one period (10 time slots) is shown in Figure.12. The adaptive method plays better than centralized (74.8%) and flooding modes (56.7%). It is near to idealized consumption (113.7%). The discrepancy is larger than example 1 because there is no intermediate state in this example.

According to previous evaluations, although the amount of reduced energy consumption depends on network characteristics, our adaptive method performs near to the idealized consumption. It also prevents wasting of energy caused by meaningless transforming in highly dynamic situations.

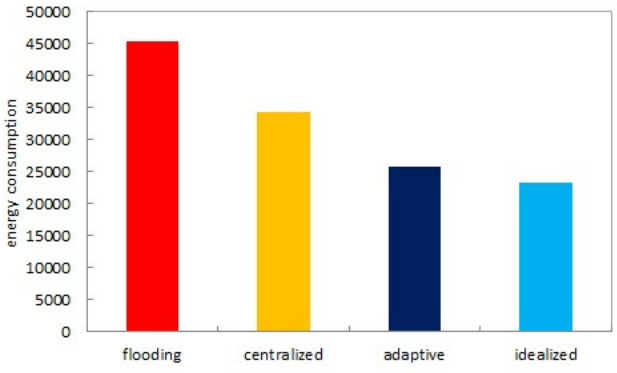


Figure 12. Energy consumption of example 2

6. RELATED WORKS

Due to the importance of resource discovery, there are many existing related works. They can be roughly divided into two categories: directory-aided and directory-less. In directory-aided strategy [2, 3, 4, 5], resource directories are used to facilitate resource discovery. [2] is the only work which considers about the combination of different discovery methods. However, it always chooses centralized mode when available and does not consider about adaptivity to different network status.

In directory-less strategy, flooding based method is used. However, energy consumption of flooding increases exponentially with request number. To solve this problem, several improved strategies like probability based [6], semantic based [7] resource discovery were proposed. However, these works either decrease RIA because not all nodes are covered or depend on semantic distance calculation which is still not reliable now. Crossing-layer strategy which binds resource discovery with routing protocol was introduced in [8]. Although the energy consumption is reduced, it is highly integrated with routing protocol which limits its compatibility. It should be mentioned that, improved centralized or flooding mode is easy to be integrated into the proposed adaptive method when they become mature. Since they perform better than their basic modes, the effectiveness of our adaptive method can also be improved. [9] gives a comprehensive survey for this area.

As a result, as far as we know, there is no existing work that is similar to our adaptive resource discovery method. The importance of adaptive resource discovery bases on method transforming is also emphasized in [9].

7. CONCLUSION & FUTURE WORKS

In this paper, we present an energy efficient adaptive resource discovery method in wireless application networks. It transforms between centralized and flooding modes to discover available resources according to network status. We

also design a heuristic algorithm to implement the adaptive method and prove its effectiveness through numerical evaluations.

Our work is only the first step of widely usage of this adaptive strategy. As discussed in previous sections, it could integrate improved methods to replace their basic companions. It could also be used to optimize other metrics like response time or RIA. The only requirement of these metrics is that they prefer different discovery strategy under different network status.

REFERENCES

[1] K. Vanthournout, G. Deconinck, and R. Belmans, "A taxonomy for resource discovery," *Personal and Ubiquitous Computing*, vol. 9, no. 2, pp. 81–89, 2005.

[2] SIG Bluetooth, "Specification of the bluetooth system, version 1.1," *http://www. bluetooth. com*, 2001.

[3] E. Guttman and J. Veizades, "Service location protocol, version 2," 1999.

[4] F. Sailhan and V. Issarny, "Scalable service discovery for manet," in *Pervasive Computing and Communications, 2005. PerCom 2005. Third IEEE International Conference on*. IEEE, 2005, pp. 235–244.

[5] C. Canali, M.E. Renda, P. Santi, and S. Burresi, "Enabling efficient peer-to-peer resource sharing in wireless mesh networks," *Mobile Computing, IEEE Transactions on*, vol. 9, no. 3, pp. 333–347, 2010.

[6] Z. Gao, X.Z. Yang, T. Ma, and S.B. Cai, "Ricffp: an efficient service discovery protocol for manets," *Embedded and Ubiquitous Computing*, pp. 786–795, 2004.

[7] M. Klein, B. Konig-Ries, and P. Obreiter, "Service rings-a semantic overlay for service discovery in ad hoc networks," in *Database and Expert Systems Applications, 2003. Proceedings. 14th International Workshop on*. IEEE, 2003, pp. 180–185.

[8] J.A. Garcia-Macias and D.A. Torres, "Service discovery in mobile ad-hoc networks: better at the network layer?," in *Parallel Processing, 2005. ICPP 2005 Workshops. International Conference Workshops on*. IEEE, 2005, pp. 452–457.

[9] C.N. Ververidis and G.C. Polyzos, "Service discovery for mobile ad hoc networks: a survey of issues and techniques," *Communications Surveys & Tutorials, IEEE*, vol. 10, no. 3, pp. 30–45, 2008.

[10] P. Vellore, P. Gillard, and R. Venkatesan, "Probability distribution of multi-hop multipath connection in a random network," in *Global Telecommunications Conference, 2009. GLOBECOM 2009. IEEE*. IEEE, 2009, pp. 1–5.

[11] N. Ristanovic, J. Le Boudec, A. Chaintreau, and V. Erramilli, "Energy efficient offloading of 3g networks," in *Mobile Adhoc and Sensor Systems (MASS), 2011 IEEE 8th International Conference on*. IEEE, 2011, pp. 202–211.

DESIGN AND IMPLEMENTATION OF VIRTUALIZED ICT RESOURCE MANAGEMENT SYSTEM FOR CARRIER NETWORK SERVICES TOWARD CLOUD COMPUTING ERA

Yoshihiro Nakajima, Hitoshi Masutani, Wenyu Shen, Hiroyuki Tanaka,
Osamu Kamatani, Katsuhiro Shimano, Masaki Fukui, and Ryutaro Kawamura

NTT Network Innovation Laboratories
1-1 Hikarinooka, Yokosuka-shi, Kanagawa, 239-0847 Japan
e-mail: nakajima.yoshihiro@lab.ntt.co.jp

ABSTRACT

This paper describes the design and implementation of a virtualized information and communications technology (ICT) resource management system called "Management Engine" (ME) for carrier network services to realize flexible service operation and dynamic resource accommodation between multiple services in the cloud computing era. To facilitate network services using virtualized ICT resources in a carrier network, a virtualized ICT information model is designed that expresses the relationship and mapping between physical resources and virtual resources for failure handling and analysis required in network carrier operations and management. A disaster recovery scenario to guarantee high-priority voice communication service in case of a large-scale natural disaster is used to examine MEs capability and functionality for providing next generation mobile network service over an OpenFlow network and virtualized servers. As a result, it is found that ME performs both integrated ICT resource management and inter-service dynamic resource accommodation. Further research areas and standardization issues ascertained from prototype experiment results are presented.

Keywords— Network management system, Network virtualization, Disaster recovery, Resource information modeling

1. INTRODUCTION

Today, new network services are going in and out of fashion much more rapidly than in the last decade. Service providers need to keep coming up with short- and long-term service improvements in terms of both performance and features. In addition, since it is very difficult for service providers to forecast the traffic demands of a service, it is very hard to deploy physical network equipment and IT equipment separately on a disaggregated basis for every service.

In data centers, cloud computing technologies such as OpenStack [1] and ProtoGENI [2] facilitate user-triggered provisioning and reconfiguration of IT resources, such as virtualized server and storage, through web portals or script programming. Since these technologies give users programmability and flexibility for IT resources in data centers, they allow both small start-ups and dynamic traffic-amount-based resource reconfiguration. As a result, service providers can handle momentary burst traffic demands, which often come up during the Christmas selling season and product launches, using on-demand elastic provisioning of IT resources in accordance with the current traffic. Therefore, they do not need to keep a large IT infrastructure for handling maximum peak traffic. This kind of cooperation scheme enables users to realize not only cost saving but also high flexibility in service operation.

Network carriers face the need to reduce OPEX/CAPEX while ensuring variety, reliability, availability, and management flexibility in network services. However, they have been deploying not only many individual networks with function-specific network nodes for each service but also server systems on which session controls and application servers such as IMS and VoD run, with the aim of achieving high performance and reliability. Under todays rapidly changing business situation, the current carrier network architecture does not provide users with sufficient flexible control capabilities for items such as resource provisioning, and does not control application programming interfaces (APIs) as well as the cloud computing architecture.

Meanwhile, disaster recovery and disaster-tolerant network services are required in many countries, especially in Japan. When a huge earthquake and tsunami devastated the north part of Japans east coast on March 11, 2011, the telecommunication infrastructure was seriously damaged and voice communication service was incredibly congested, with more than ten times the normal call traffic volume [3]. To ensure that robust voice communication service will be provided in the event of a major disaster, it is essential to dynamically reconfigure carrier networks, servers, and systems to enhance call acceptance performance for momentary burst call attempts. It is also essential to build a robust fiber/path network.

Resource virtualization of network and IT equipment in carrier network architecture is a key technology in this regard. In recent decades, networks and IT resources have been virtualized through, for example, multiprotocol label switching (MPLS) and virtual routing and forwarding (VRF) for networks, hypervisor-based virtual machine middleware such

as KVM and Xen, and containers for IT resources. However, the networks themselves have not been as fully virtualized as servers, since most network functions are strongly bonded to the physical network nodes. This has made it hard to use existing virtualization technologies to decouple the functions. Software-defined networking technologies, such as OpenFlow networks, are used to tackle the problem of improving management flexibility, and to enable slicing of isolated networks, which is a form of network virtualization, by decoupling the control plane from the forwarding plane from the network nodes [4–6].

Management and resource control for the virtualized ICT resources remains one of the most significant challenges for network carriers. In particular, virtualized network information modeling has not yet been discussed to any extent. Such modeling must express the relationship and mapping information between physical equipment and virtual function equipment because virtualized ICT resource management requires decoupling the functionality from the entity of the physical ICT resources. Furthermore, cooperation in the area of ICT integrated resources, including the use of resource access APIs and semantics, is required to achieve today's network services. Various issues need to be addressed to achieve flexible network service management of virtualized ICT resources. These issues include the need to model virtualized ICT resource information and describe resource access interfaces, abstracted network topologies, and control and service policies.

The purpose of this study is to establish a virtualized ICT resource information model, one that takes carrier network maintenance and operation requirements into account, in order to leverage proactively virtualized ICT resources in carrier network services. In this paper, we first summarize discussion points about exploiting virtual ICT resources from a carrier network service standpoint, and existing carrier network architectures from the standpoint of the resources-services relationship. We then describe our design of a virtualized ICT resource management system called "Management Engine" that implements a virtualized ICT resource information model. To examine MEs effectiveness and adaptability from a network operator perspective, we show how it uses both an OpenFlow network and server virtualization to achieve integrated ICT resource management and service cooperation between two mobile network services on a mobile network prototype. Further improvements and standardization topics are summarized on the basis of knowledge obtained in validating the prototype.

The contributions of this study include:

- Designing a virtualized ICT resource information model to take advantage of ICT virtualization technology in carrier network service.
- Designing a modular/layered management system architecture for virtualized ICT resource to accommodate various new resource types and protocols.
- Using a mobile service prototype with a disaster recovery scenario to demonstrate flexible service cooperation to ensure dynamic resource accommodation between services with virtual ICT resources.
- Summarizing future standardization and improvement issues in the virtual ICT resource management field.

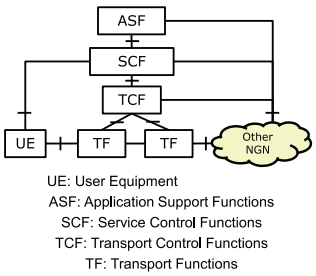


Figure 1. NGN architecture.

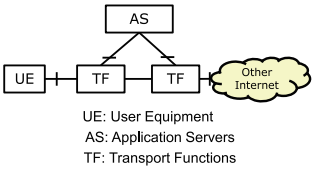


Figure 2. Internet architecture.

In the next section, we survey and summarize existing network architectures and research issues for virtualized ICT resource management. Section 3 describes our design of 1) a carrier network service architecture for virtualized ICT resources and 2) the ME management system that implements a virtualized ICT resource information model. A prototype implementation for mobile network service and its demonstration in case of disaster recovery are reported in Section 4. Section 5 discusses future research areas and standardization issues on the basis of perception derived from the prototype system and experiments. Finally, in the last section we conclude the paper with a brief summary of key points.

2. REQUIREMENTS FOR VIRTUALIZED ICT RESOURCE MANAGEMENT

Before describing our design of a carrier network service architecture using virtualized ICT resources, we summarize existing network service architectures from the resource type and service-resource cooperation features.

Next generation network (NGN) requirements, architectures, and protocols have been actively discussed in ITU-T and published as a set of NGN recommendations. It is on the basis of these recommendations that NGN services are now being deployed. Discussions on network architectures are now being focused on future networks, and it is necessary to clarify the differences between future networks and NGNs with respect to functional entities. The ITU-T Y.2012 [7] standard defines NGN functional requirements and architecture [8], and Figure 1 shows the schematic structure of the latter. A service control function, resource control function, and transport gateway function are explicitly defined in order to provide NGN services with the expected quality, security, and reliability.

From a similar perspective, the schematic structure of the Internet architecture could be described as shown in Figure 2. In the Internet architecture it can be considered that the service control function, resource control function, and transport control function are not explicitly placed but are implicitly included in the application servers and packet transport

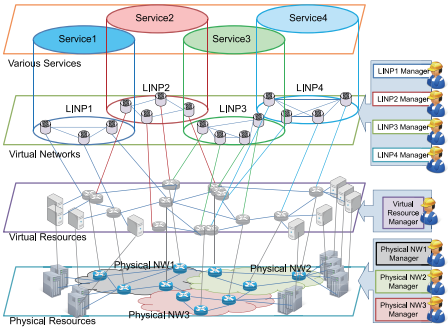


Figure 3. Network virtualization architecture of Y.3011.

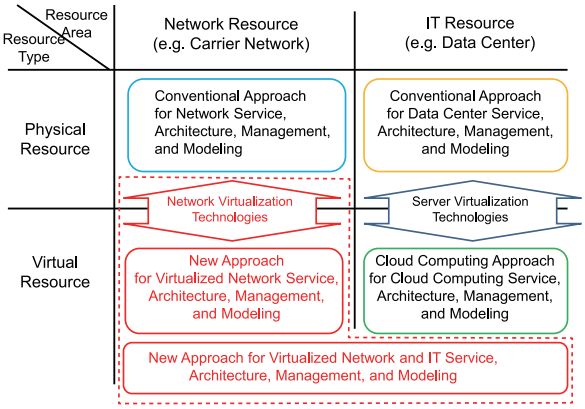


Figure 4. Overview of virtualization technology.

functions. The protocols between the application servers and user equipment can be proprietary, and the protocol processing functionality can be installed in the application servers and user equipment in response to requirements for new services.

In both NGN and Internet architectures, application service functions, which are assumed to be run on IT resources, and network functions are separately managed. In other words, they are not integrated. In addition, the cooperation function between application service and ICT resources is limited, making it impossible to realize optimal global optimization and control.

Various efforts have been made to define future networks to overcome the existing network problems efficiently. Recommendation Y.3001 [9] aims to clarify future network objectives and design goals, and Y.3011 [10] describes the framework of network virtualization for future networks. Figure 3 shows the framework of network virtualization architecture. From a network architecture standpoint, as discussed in the process of defining NGN architectures, future network architectures should clarify several design concepts, including ICT resource support and control functions and service-cooperation control functions.

To design a carrier network service architecture on virtualized ICT resources, we summarize requirements from a service and virtualized resource management perspective. Figure 4 shows the overview of virtualization technologies. We focus on the research issues marked in red in the figure.

Current carrier network services do not use virtualized resources because it is hard to trace the assignment relationship between physical and virtual resources from the management perspective. To incorporate virtualization technologies into carrier network service, the virtualized network information model should clarify what a virtualized network is, what function it provides, and what resources it has. In addition, a virtualized ICT resource information model is required to describe the traceability and mapping relationship between virtual and physical resources. Cloud computing technologies focus on managing virtualized and physical IT resources in data centers. However, they support only limited cooperation control features between the IT resources and the network.

To realize the architecture we propose, one issue to be addressed is how to achieve flexible and close cooperation between virtualized network resources and virtualized IT resources. In designing the architecture, we used the approach described in ITU-T Y.3011, and thus the architecture is similar to that described in the recommendation. Moreover, we extended it to take advantage of virtualized IT resources to guarantee maximum flexibility of service operations. The ME management system we propose should achieve the following requirements:

- multi-tenant support
  The virtualized network should support isolated networks and include computing resources on top of shared physical ICT resources. Users can exploit delegated control functionalities in network and computing resources, e.g., policy routing, filtering, and QoS control in the network.
- cooperation control and resource provisioning
  User-triggered integrated ICT resource provisioning and control should be supported without interference from other service networks' operations and control. In addition, multilayer-cooperation control between ICT resources and network services should be supported.
- traceability and mapping relationship information between virtualized and physical resources
  A proposed architecture's management system must provide useful information to operators to perform failure-analysis and fault-recovery operations in the same way as in current network service management systems.

3. CARRIER NETWORK SERVICE ARCHITECTURE DESIGN

The goal of this study is to establish a carrier network service architecture on top of virtualized ICT resources in carrier networks to encourage flexible network service operations and dynamic service reconfigurability for service providers and carrier operators. The design scope described in this paper includes ICT resource management, ICT resource control, and an abstracted ICT virtualization information model from

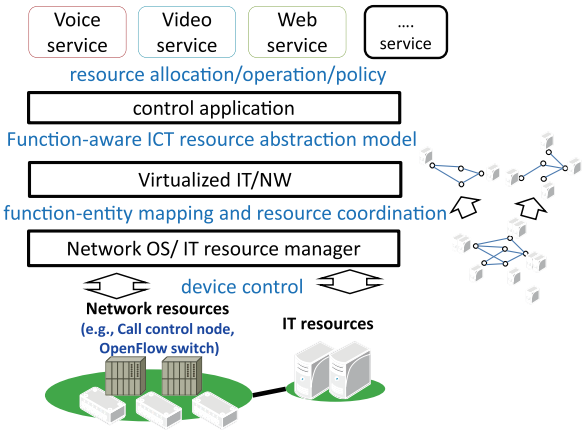


Figure 5. Management layer model.

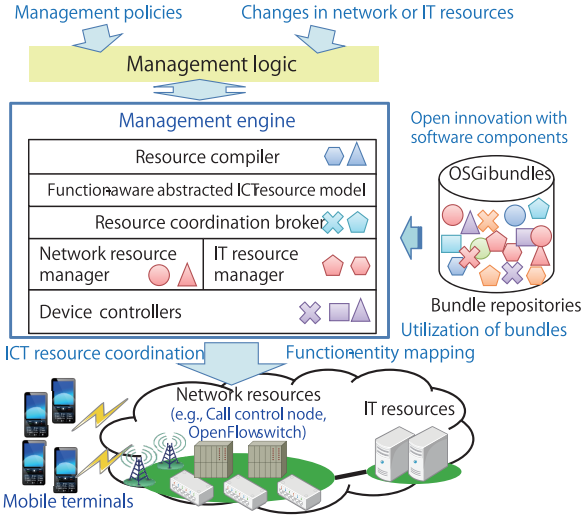


Figure 6. Architecture overview of Management Engine.

the aspects of management, but does not include accounting and security.

Figure 5 shows the concept design of the proposed network architecture. We introduce the ICT resource virtualization layer, which is the new neck of hourglass to provide the virtualized ICT resource abstraction model for service providers. The infrastructure operator in the carrier network maintains the physical network and computing resources to provide virtual ICT resources requested by users or service providers. Service providers and users may build their own service management system on top of the virtual ICT resources provided by the infrastructure management system. Controls and provisioning from service providers are performed with public programming interfaces with the information described in the abstracted ICT virtualization resource information modeling with control permission delegated to virtualized ICT resources. Monitoring information for users is given in the same manner.

3.1. Management System Design

Since the management system plays the most important role in this architecture, we used the following approaches in designing it:

- Layering/modularizing architecture
- Utilizing software component technologies
- Leveraging software-defined network technologies

A layered and modularized architecture allows a management system to exploit the benefits of technology innovations with minimum development and modification of them. Using software component technologies makes it possible to easily replace small granular modules during runtime. This enables the management system to be used and developed over a long time period.

In the existing resource management systems of network carriers and service providers, wide use has been made of computing resources and physical entity-based resource management of the network equipment, e.g., NMS and EMS. One solution to achieve sustainable use and evolution for a management system is to decouple the system nodes functionality from its physical resource entity.

To decouple the systems functionality from its ICT physical resource entity, the system needs both a function-aware ICT resource abstraction model and a multi-layer coordination technique to allocate functions to ICT resources. In addition, flexible ICT services can be provided regardless of the physical resource configuration if the management system leverages both a software-defined network's programmability and ICT resource virtualization functionality.

Figure 6 shows the architecture overview of the ME management system we propose for the proposed network service architecture. In this figure, the management logic is a component to handle management policies and operation orders, which in carriers and service providers are described with much more abstracted notation, including operation and ICT resources.

3.2. Modules

The ME system consists of seven principal modules. Each module communicates with other modules through a function-aware ICT resource abstraction model. It is assumed that all module implementations are replaced with other implementations that have the same functionalities and APIs. The detail roles of each principal module are listed as follows:

Resource compiler The resource compiler module converts a request from the management logic component into a resource allocation request or a resource control request. It is assumed that more abstracted operations are performed in the management logic component. The resource compiler accepts user requests from the management logic.

Function-aware virtualized ICT resource informationmodel This model is used for all the procedures carried out in

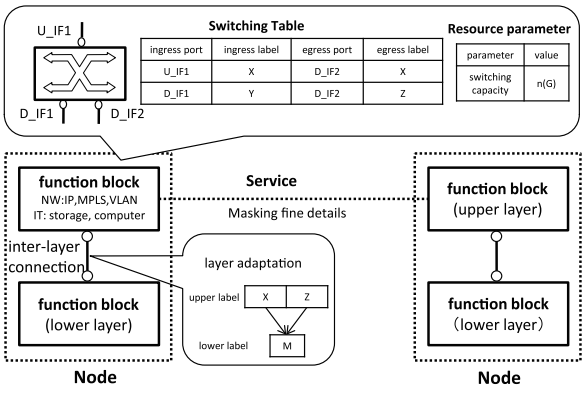


Figure 7. Component details of ICT integrated resource abstraction model.

ME. Module-specific properties and information are superimposed to handle the differences between the model and the actual ICT equipment.

Resource coordination broker The resource coordination broker module searches the management logic component for the optimal combination of network and IT resources to meet resource requirements. It also performs multi-layer-aware resource coordination to assign functions to the ICT resource entity in cooperation with the network resource manager module and the IT resource manager module.

Network resource manager The network resource manager module handles general network management, which includes fault monitoring, configuration management, and performance monitoring.

IT resource manager The IT resource manager module performs general IT-related management.

Device controller The device controller module plays a wrapper role to absorb the differences between the ICT resource abstraction model and the actual ICT equipment information model, which vary depending on the resource types and vendors.

3.3. Function-aware virtualized ICT resource information model

The proposed model is designed to provide unified ICT resource management from the functionality standpoint. Existing resource information models, e.g., ITU-T G.805 [11], TM Forum Information Framework (SID) [12], and NDL [13], have limited resource information and focus on a specific network domain. Since the existing models mainly focus on the modeling for physical network equipment, it is impossible for them to perform unified calculations or processing for multi-layer-capable ICT resource coordination searches, and for allocating functions to ICT resource entities. We designed the model by referring to ITU-T G.805, TM Forum SID, and NDL. Moreover, we extended it to handle network resources, IT resources, and the unified calculations needed for combining ICT resources.

The proposed model consists of three components: function block, layer adaptation, and service. Figure 7 illustrates the relationship of the three components and each component's properties in the function-aware ICT integrated resource abstraction model. The detailed functionalities of these components are summarized as follows:

Function block The function block notation represents network protocol, function (e.g., IP, MPLS, VLAN), and IT resource type (e.g., computing, storage). The global routing/forwarding path can be controlled by specifying a label-based forwarding rule or a modifying rule in each logical switch instance at a function block. The unified ICT resource management is performed by managing the resource properties of a function block, such as switching performance, storage capacity, and processing performance.

Layer adaptation The layer adaptation notation represents the hierarchical relationship between the functional blocks. Flexible changes in layer adaptations are permitted. In other words, a set of the underlying functional blocks can be swapped without side effects to the blocks, and the inter-layer connection above the blocks can be swapped by changing only the inter-layer connection relationship.

Service The service notation represents the logical link between the functional blocks in the same layer. If one-by-one correspondence exists between two services and their input/output labels, which is expressed by using a service's properties (e.g., IP address in IP service) in a function block switching table, two services in the same layer are logically connected. More abstracted service notation can be made acceptable by omitting the internal function blocks or the underlying function blocks and services to provide a flexible resource view.

4. ARCHITECTURE IMPLEMENTATION FOR MOBILE NETWORK SERVICES

We developed a prototype ME system to manage both network services and ICT resources over a next generation mobile network prototype to examine ME's service control capability and flexibility for unexpected events, such as network failures and rapid changes in traffic demands.

The prototype mobile network was designed by referencing the architecture of the 3GPP Long Term Evolution (LTE) mobile network, which is an all-IP based network. Two types of services, voice communication service and video-on-demand service, run over the prototype mobile network. These services are enabled by using dynamically allocated network and IT resources (e.g., a virtual machine instance), while fixed allocated resources using the maximum capacity-based design are used in current mobile networks.

The ME manages both network resources and IT resources. Since dynamical control and allocation are indispensable for the network and the IT resources, we used an OpenFlow network for network virtualization and KVM for IT resource virtualization.

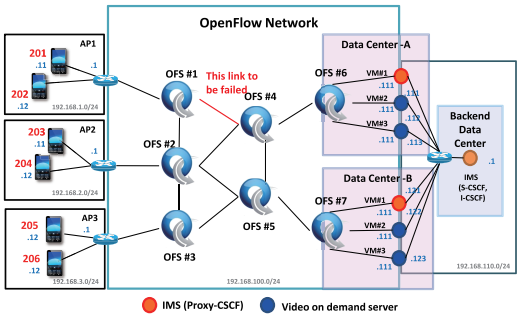


Figure 8. Prototype network and server configuration.

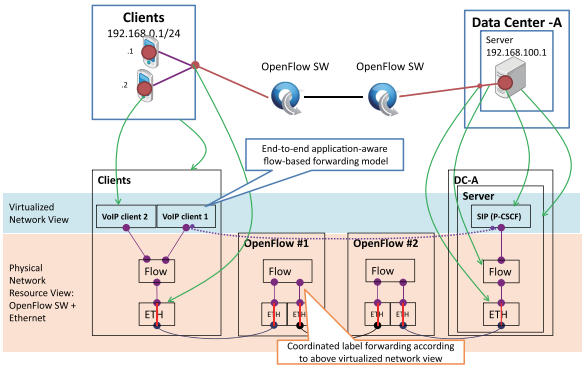


Figure 9. Model notation of VoIP service over OpenFlow network.

4.1. Prototype Configuration

Figure 8 shows the ICT resource configuration of the mobile network prototype used for these experiments.

An IP multimedia subsystem (IMS) with dynamic scale-out/scale-down extension is used for voice communication service, and a Flash over HTTP video server is used for video on-demand service. There are two data centers, in which the P-CSCF of IMS and an HTTP video server program are executed on three virtual machine instances, one backend data center, where the S,I-CSCF of IMS runs, and three wireless access points that connect several mobile phones. The OpenFlow network provides an application-aware virtual path between these data centers and mobile phones.

Figure 9 illustrates the voice communication service modeling over the OpenFlow network. In this demonstration, we defined the layer-4-level client-to-server path as the minimum elements of network virtualization view for operators. A layer-4-level client-to-server path is regarded as a bearer between client and server in 3GPP LTE. The physical resource entities in the underlying layers, including Ethernet switching and OpenFlow switching, are systematically configured by the function assignment request of a layer-4-level client-to-server path from operators.

For VoIP clients and VoD Viewer, we used the standard VoIP software and Web browser initially installed in Android version 2.3. The service side interface of all servers uses the same IP address and the same MAC address in order to achieve transparent access from a mobile phone without any modification to the mobile phone configuration. In this configuration, only control messages from ME are needed to perform dynamic service scale-out and scale-down.

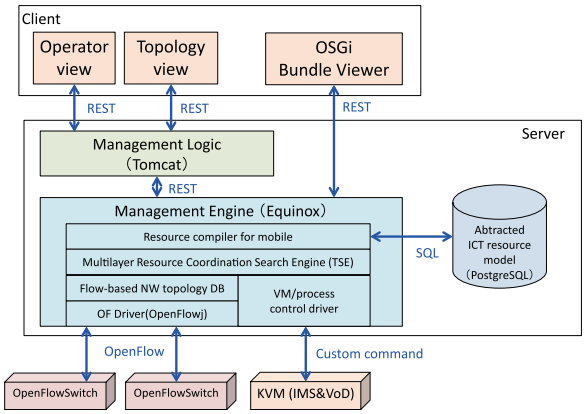


Figure 10. Implementation overview of Management Engine.

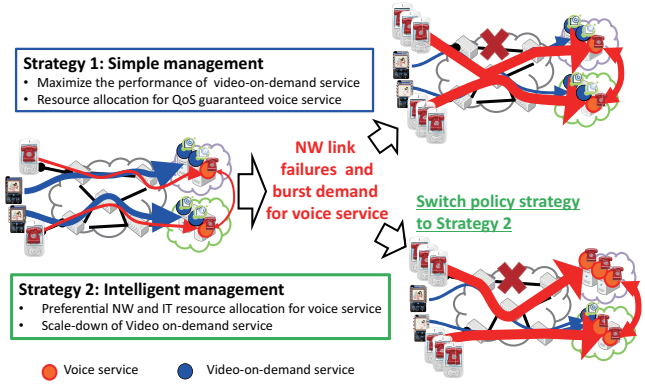


Figure 11. Difference between two operation policies.

4.2. Management Engine Implementation

Figure 10 shows the ME implementation for these experiments. All ME modules in this figure are developed as OSGi bundles. An OpenFlow driver, which exploits the OpenFlowj library, controls seven OpenFlow switches. Flow-based policy routing specified by using layer-3 and layer-4 header properties (e.g., protocol type, port number), is performed. Since a multi-layer resource combination search engine is needed for ICT-tightly-coupled resource coordination and function assignments for the mobile network service, we used our novel multi-layer ICT-resource-integrated threaded path/resource search engine (TSE) [14].

Layer-3 and layer-4 flow properties are used to store all the flow tables in each OpenFlow switch as label switching tables. If a client or server sends voice or video streaming flow, each OpenFlow switch sends a PACKET_IN message of the OpenFlow protocol to ME to acquire flow rules calculated by ME in accordance with operator control requests. The VM/process control driver on ME issues a shell-script based message to servers to control the ICT resources.

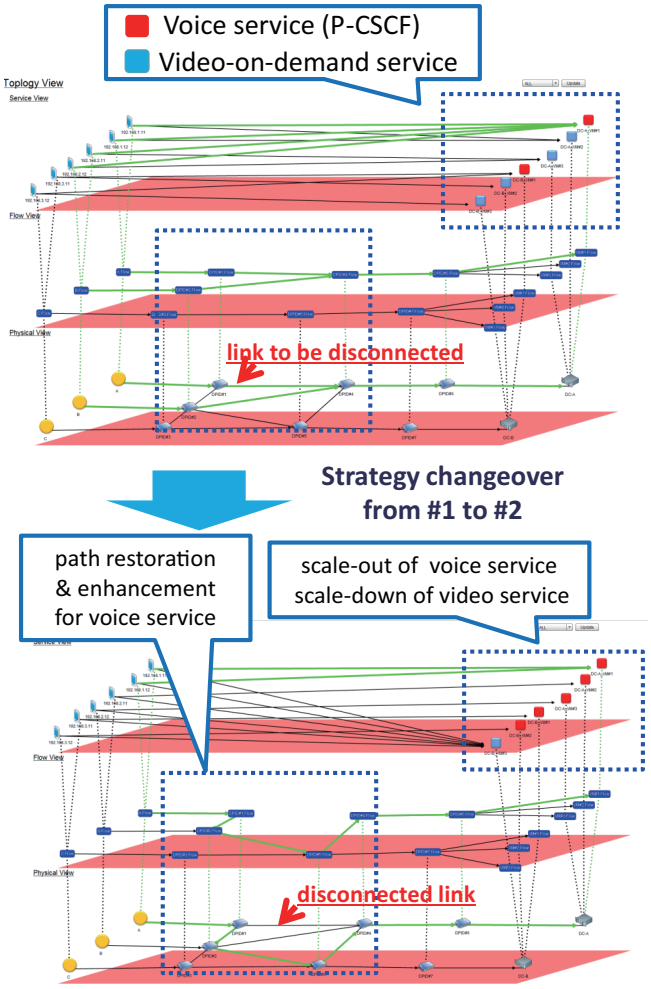


Figure 12. ICT resource and topology changes produced by strategy change from #1 to #2.

4.3. Operation strategy policy

We have implemented two operation policy strategies as the implementation of the TSE module to swap policies in accordance with event situations. By swapping a TSE OSGi bundle in ME, an operator can select a suitable strategy for traffic demand changes. Figure 11 shows the conceptual difference between the two operation policies when simultaneous situation changes occur. The two strategies are summarized as follows:

Strategy #1: Simple management The resource assignment is optimized to enhance the scalability of the bandwidth-consuming VoD service. This enables bandwidth-aware network path allocation to be performed. The network and server resources are assigned to ensure the minimum QoS for voice service.

Strategy #2: intelligent management Developed for emergency operation in the event of a natural disaster such as a major earthquake. Since it is assumed that burst demands for voice service will be made, preferential resource allocation and service scale-out for voice service are carried out while video service may be scaled down. Delay-aware network path allocation can be performed with path restoration to avoid link failures.

4.4. Demonstration

We conducted a demonstration for the following scenarios to examine ME flexibility from the management standpoint.

1. Normal state: Strategy #1 is chosen.
2. Network problems due to a major earthquake: A link failure occurs at the red line in Figure 8 and burst emergency call demands for voice service may be predicted.
3. Enforcement for voice service with network restorations: Strategy changeover from #1 to #2.

Figure 12 shows the ICT resource and network topology changes between scenarios #1 and #3. The whole processing time, which includes the calculation time for resource coordination, allocation, OpenFlow network control, and service control, takes less than five seconds after the operator switches the operation policy over from #1 to #2. Twelve virtual paths are assigned over seven OpenFlow switches. After the configuration was completed, the mobile phone was able to connect the voice service and the video on-demand service instantly without any mobile configuration change. In this way we found that the ME can provide flexible operation and management for ICT-tightly-coupled network services by swapping the operation policy strategy for an OSGi bundle.

However, about a 30-millisecond-delay is incurred in installing flow rules on OpenFlow switches because an OpenFlow switch refers to flow rules in the OpenFlow flow DB in ME when a new flow is coming to the switch.

5. DISCUSSION

In this section we discuss future research areas and standardization issues to improve the virtualized ICT resource management and management system architecture. The major future research areas and standardization issues are described as follows:

- Unified abstracted ICT resource information model

  In this paper, we have described our design for abstracted ICT resource information modeling with the focus on OpenFlow-based flow switching capability, resource capability, and ICT integrated resource searching. However, this description is not thorough enough to present programmability and adaptability for flexible network virtualization. The DMTF and TM Forum associations have attempted to establish ICT resource information modeling from the node entity level to the protocol level for network control information to accommodate model including network carrier specific network nodes, protocols, such as tunneling and control, and OAM entities.

- Standardized API for virtualized ICT resource control and provisioning

  Standardization of APIs is an important issue in the carrier network field as a means to lower entry barriers for service providers and encourage new service creation. To achieve standardization it is necessary to consider the APIs that have been proposed or used in cloud computing so far. The DMTF association has attempted to standardize APIs for ICT resource control and provisioning from a data center management standpoint. In contrast, OpenStack, CloudStack, and Amazon EC2 have defined their own APIs for cloud computing platform control and provisioning.

- Operational policy description language

  Arbitrating virtualized ICT resource accommodation among services is very important when large-scale natural disasters occur. It is necessary to have an operation policy description language to arbitrate dynamic resource provision to ensure highly prioritized network service, e.g., voice communication service in case of a major earthquake. In the existing operation procedure, operators handle decisions on resource allocation and path restoration for achieving high priority service. To handle these matters faster and more flexibly, and to reduce operation costs including those for operator personnel, one challenging issue is establishing an operational policy description language that allows automatic system handling.

6. CONCLUSION

We have designed and implemented a virtualized ICT resource management system called Management Engine for carrier network services with a virtualized ICT information model that expresses the relationship and mapping between physical and virtual resources. A demonstration with a prototype mobile network service under a disaster recovery scenario confirmed the ME systems flexible operation and capability.

Our future work will include standardizing a virtualized ICT resource information model and other management aspects, such as the monitoring and propagating of alert information in virtual networks.

7. REFERENCES

[1] "Openstack," http://www.openstack.org/.

[2] "Protogeni," http://www.protogeni.net/trac/protogeni.

[3] Masaru Fujino, "Ict responses to the great east japan earthquake, 9 months later," 2011, US Telecom Association Boarding Room.

[4] N. McKeown, T. Anderson, H. Balakrishnan, G. Parulkar, L. Peterson, J. Rexford, S. Shenker, and J. Turner, "Openflow: enabling innovation in campus networks," *ACM SIGCOMM Computer Communication Review*, vol. 38, no. 2, pp. 69–74, 2008.

[5] D. Staessens, S. Sharma, D. Colle, M. Pickavet, and P. Demeester, "Software defined networking: Meeting carrier grade requirements," in *Local & Metropolitan Area Networks (LANMAN), 2011 18th IEEE Workshop on*. IEEE, 2011, pp. 1–6.

[6] T. Koponen, M. Casado, N. Gude, J. Stribling, L. Poutievski, M. Zhu, R. Ramanathan, Y. Iwata, H. Inoue, T. Hama, et al., "Onix: A distributed control platform for large-scale production networks," *OSDI, Oct*, 2010.

[7] ITU-T Y.2011, "General principles and general reference model for next generation networks," Tech. Rep., ITU-T, 2004.

[8] ITU-T Y.2012, "Functional requirements and architecture of next generation networks," Tech. Rep., ITU-T, 2010.

[9] ITU-T Y.3001, "Future networks: Objectives and design goals," Tech. Rep., ITU-T, 2011.

[10] ITU-T Y.3011, "Framework of network virtualization for future networks," Tech. Rep., ITU-T, 2012.

[11] ITU-T Recommendation G.805 , "Generic functional architecture of transport networks," Tech. Rep., ITU-T, 1995.

[12] TM Forum, "Gb922 information framework (sid) suite release 9.5," Tech. Rep., TM Forum, 2011.

[13] F. Dijkstra, B. Andree, K. Koymans, J. Van Der Ham, P. Grosso, and C. de Laat, "A multi-layer network model based on itu-t g. 805," *Computer Networks*, vol. 52, no. 10, pp. 1927–1937, 2008.

[14] K. Yamada, Y. Tsukishima, K. Matsuda, M. Jinno, Y. Tanimura, T. Kudoh, A. Takefusa, R. Takano, and T. Shimizu, "Joint storage-network resource management for super high-definition video delivery service," in *National Fiber Optic Engineers Conference*. Optical Society of America, 2011.

HARMONIZED Q-LEARNING FOR RADIO RESOURCE MANAGEMENT IN LTE BASED NETWORKS

Dhananjay Kumar[1], Kanagaraj N N[2] and Srilakshmi.R[3]

[1]Department of Information Technology, Anna University, MIT Campus, Chennai, India.
Email: dhananjay@annauniv.edu
[2]Alcatel-Lucent India Limited, Chennai, India.
Email: kanagaraj.n.n@alcatel-lucent.com
[3]Department of Information Technology, Anna University, MIT Campus, Chennai, India.
Email: srilakshu14@gmail.com

ABSTRACT

The efficient management of radio resource is highly imperative so as to meet the vast application requirements in future high speed wireless networks such as Long Term Evolution-Advanced (LTE-A). The current research on applying machine learning algorithms either focuses on packet scheduling in infrastructure network or in cognitive radio in ad-hoc environment. Our study on spectrum usage indicates that there is a lot of room for optimization of spectrum in a multi-operator scenario of LTE systems which covers large customer over a vast geographical area. In this paper, we introduce the concept of Harmonized Q-Learning (HQL) for the radio resource management in LTE based networks that efficiently manage its resource pool dynamically. The multi-operator system is modeled on the game theory based Q-Learning. Our system level simulation of the proposed algorithm shows higher throughput while meeting the real-time resource requirement of each player.

Keywords— *Cognitive radio, LTE, Q-Learning*

1. INTRODUCTION

With dramatic increase in wireless data traffic (e.g., web browsing, media streaming etc.), the system developers working with Long Term Evolution (LTE) and LTE-Advanced (LTE-A) are bound to seek innovative resource management techniques. The Cognitive Radio (CR) with suitable machine learning algorithm is expected to provide a robust solution to the management of scarce spectrum resources [1]. As per the ITU-R broader guidelines "The International Mobile Telecommunications (IMT) system using a cognitive radio system must operate in accordance with the radio regulations, local and regional administration rules governing the use of a particular band" [2].

The LTE-A is required to provision peak data rate of 1 Gbps with average spectrum efficiency of 3.7 bps/Hz/Cell [3]. It is obvious that LTE-A will face severe spectrum shortage unless some non-conventional technique for usage of radio resource is adapted [4]. This also justifies the research interests towards application of cognitive radio in future mobile radio systems. To analyze the pattern of spectrum usage in a wide band (1.7 GHz – 2.3 GHz), spectrum samples were collected at different time interval and one such is shown in Fig.1. To get further insight into the practical scenario we also observed it in narrowband of 1.79 GHz-1.84 GHz (Fig.2). As can be seen in Fig.1 & 2, there is lot of unused spectrum (hole) in the band of observation although the cellular mobile service in the city (Chennai) is served by multi-operator systems. This result motivated us to carry out further research in this spectrum optimization.

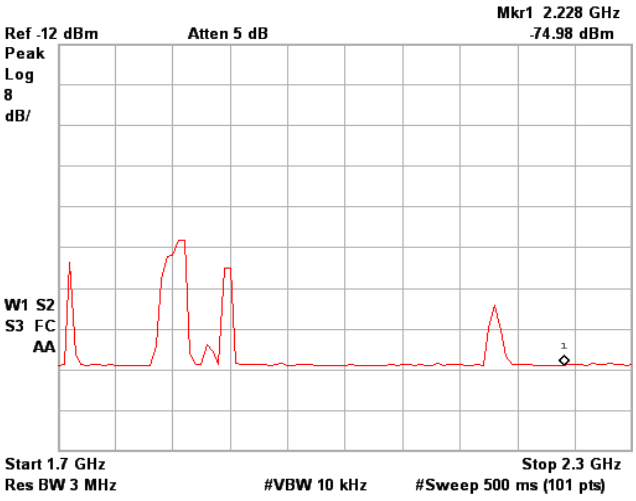


Figure 1. Measurement taken at Adyar, Chennai in wide band (1.7 GHz-2.3 GHz).

In a typical multi-operator set-up of cellular networks, it is highly unlikely that each service provider will have similar number of subscribers in their different location areas. In other words, if an operator does not need all Resource Blocks (RBs) in a given time and location, other operators can share this unused RBs based on the concept of CR and vice-versa.

A suitable learning algorithm can be developed in multi operator environment of cellular network that permits

optimum sharing of spectrum. This scenario can be modeled with game theory based learning where each player tries to attain equilibrium [5].

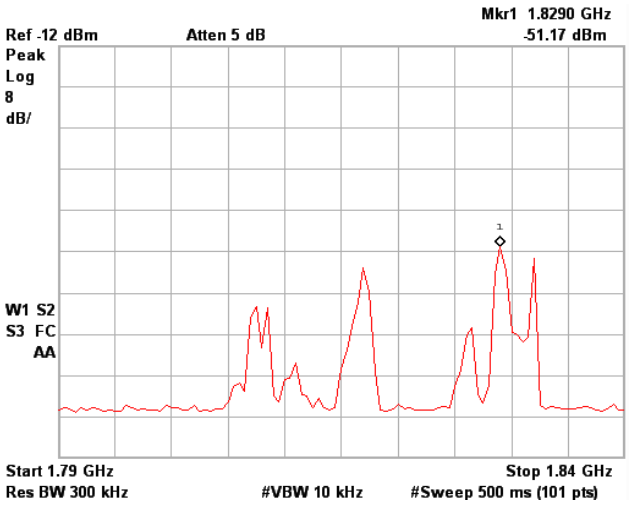


Figure 2. Measurement taken at Adyar, Chennai in 1.79 GHz- 1.84 GHz.

Our earlier work [6-7] on CR system was focused on optimization of allocation of sub-carrier and power in OFDMA based network considering primary user behavior. In a larger context of spectrum sharing, optimization of resource allocation at subcarrier is not enough. Particularly, for system operating in a large geographic area with non-uniform distribution of user, the mechanism for the maximization of resource utilization can be implemented with some machine intelligence in Medium Access Control (MAC) layer. In this paper, we develop reinforcement learning with game theory to manage the licensed and free (cognitive) RB of an LTE based network. Every evolved Node-B (eNodeB) is required to employ mobile terminal assisted co-operative sensing to estimate the availability of cognitive RB. The eNodeB corresponding to the different operator needs to co-ordinate among them to avail the free RBs.

Our proposed algorithm addresses the learning problem in Multi Agent System (MAS) using Nash Q-learning which is mainly appropriate to be applied for RB allocation in LTE networks. First we explore single agent Q-learning which is solved using Markov Decision Process (MDP) and it is extended for multi agent context. The purpose of an agent in MAS is to maximize its own expected reward in each Transmit Time Intervals (TTI). In MAS the action of one agent affects the reward of other agents, and therefore the optimal behavior of each agent need to be trained. Game theory is a suitable solution needed to predict the strategies that agents will choose to play in a particular game.

The purpose of adapting game theory is to find out equilibrium with its information and make decision following the equilibrium strategy. Since the behavior of MAS may change as they also learn and adapt, a suitable strategy need to be developed. The learning process in the system will depend upon the plan and policy of each agent. We introduce Harmonized Q-Learning (HQL) algorithm to adapt the action based on past and current situation that tackles the problem of learning and coordination using two modes, (i) simultaneous play mode allows each agent to get an opportunity of sharing the best RBs equally in each TTI, and (ii) alternate play mode allocates the resource blocks for an agent in each TTI depending on the traffic demand.

In the proposed HQL algorithm, each agent has to maintain m Q-tables and the total space requirement is $m|X| \cdot |A|^m$, where X is the state space, A is the action space, and m is the number of agents. The HQL algorithm performs better in terms of space complexity, and is linear in the number of states, polynomial in the number of actions, but exponential in number of agents.

The rest of the paper is organized as follows. Section 2 deals with related work. System architecture that includes learning in cognitive environment in presented in Section 3. The Q-Learning is modeled in Section 4. Section 5 discusses formulation of resource allocation in CR based LTE networks. Results and discussion is presented in Section 6. This paper concludes in Section 7.

2. RELATED WORK

The intelligent mechanism in packet scheduling has been advocated by many to make the radio resource management more efficient. In [8], throughput and fairness is improved by applying Q-learning which finds out the suitable rule for each TTI. The non-negotiation based Q-learning and Robinson Monro algorithm [9] is applied for collision avoidance in multi-channel multi user cognitive radio system which enables only one secondary user to access the channel. In [10-11], the Q-learning is applied to bid the multiple frequency bands at same time for transmission of data across cognitive nodes and multiple frequency channels are selected with or without considering the channel state information and primary user traffic.

The cognitive radio concept was proposed in LTE Networks and the subcarrier and power allocation is performed using K-Nearest Neighbor Algorithm [12]. The cognitive based spectrum sharing between DVB and LTE-Advanced using autocorrelation based advanced spectrum sensing method was proposed in [13]. Distributed Q-learning is used to avoid interference between macro cell, femto cell, and neighboring femto cells in [14]. Centralized and distributed dynamic spectrum leasing architectures [15] were proposed where the power level of primary user is reduced by using secondary user as relay node. In [15], the unused spectrum is allocated based on Hungarian algorithm and the primary and secondary users learn and revise their actions based on reinforcement learning.

In [16], Q-learning is applied in cognitive wireless mess networks where the multiple secondary users are allocated with best power levels in a non-cooperative manner. This work is in line with our proposal but for mesh networks. Our proposed HQL algorithm introduces cognitive radio concept in LTE Networks that assigns RBs based on learning and coordination in multi agent frame work.

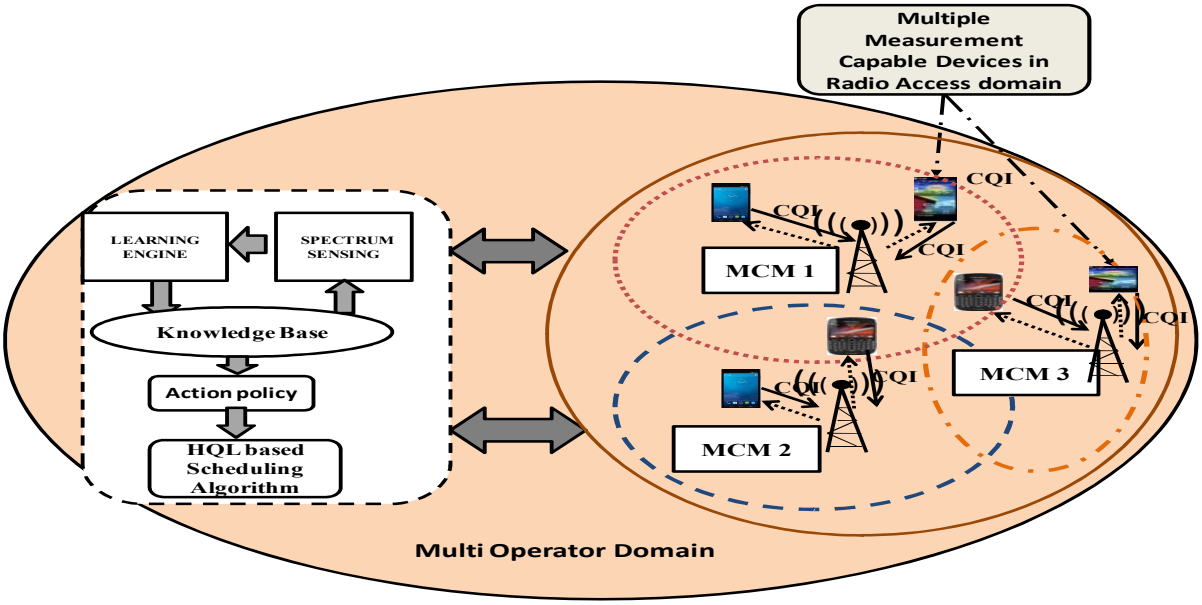


Figure 3. Radio resource management with learning in cognitive environment

3. SYSTEM ARCHITECTURE

The radio resource blocks are seen as time frequency grid. Time units correspond to Transmit Time Intervals (TTIs) where each TTI is divided into two time slots. The frequency and time units are represented as Resource Blocks (RBs). The input to the learning algorithm is based on the Channel Quality Index (CQI) reported on physical uplink/downlink shared channel, the usage pattern of primary user, and the presence of other secondary users (Fig.3). The Measurement Collection Module (MCM) collects information such as CQI from the multiple measurement capable User Equipment (UEs) devices. Q-Learning which is one of the reinforcement learning algorithms addresses the problem of single agent using MDP. By incorporating cognitive concept, the multi agent scenario in LTE networks can be formulated using our HQL algorithm which aims to maximize the available spectrum by utilizing the unoccupied resource blocks. We consider a cooperative game among the agents (operators) where each operator seeks to choose and use its resource blocks through learning and coordination.

4. MODELING OF Q-LEARNING IN CR

The basic reinforcement learning concept is exemplified in Fig.4, where the cognitive eNodeB updates its strategy according to its experience with different actions in the prevailing environment. The agent directly learns about its optimal strategy without knowing the reward function or state transition function, such an approach is referred as model free reinforcement learning of which Q-Learning is one example. Single agent Q-Learning can be extended to Multi agent Q-Learning in the presence of multiple eNodeBs.

4.1. Single Agent Q-Learning

The environment which the agent interacts with the other agents is naturally formulated as a finite state Markov decision process. A Markov Decision Process (MDP) is a tuple of $< X, A, U, T >$ where X is a discrete state space $x \in X$, A is a discrete action space e $\in A$, $u_n: X * A \rightarrow U$ is a reward function of the agent and $T : X * A \rightarrow \delta(X)$ is the transition function where $\delta(X)$ is set of probability distribution over state space X.

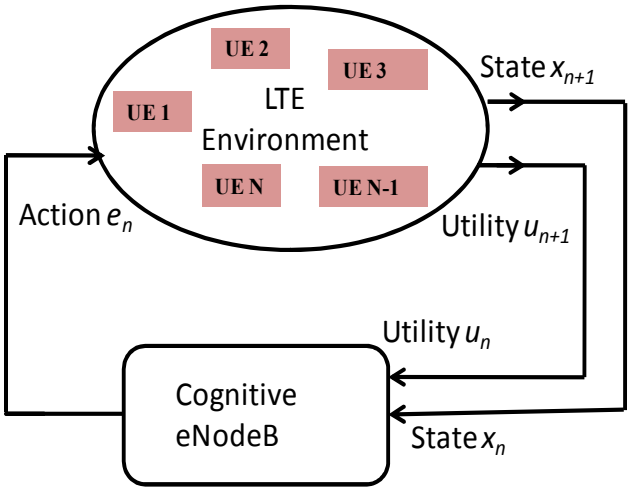


Figure 4. Cognitive reinforcement Learning

In MDP, an agent attains reward value u_n whose mean value $u_{x_n}(e_n)$ depends only on the state x, e. The state of the environment changes probabilistically to next state $x_{n+1} = x' \in X$ according to the law,

$$\text{Prob}\,[x_{n+1} = x' \,|(x_n, e_n)] \;=\; T_{x_n}\, x'\,(e_n) \qquad (1)$$

The agent's goal is to find a strategy σ so as to attain high reward value. The value of state x_n is,

$$V(x,\sigma) = \sum_{n=0}^{\infty} \rho^n\, E\,(u_n \mid (\sigma, x_0 = x)) \qquad (2)$$

Where x_0 is the initial state, u_n is the reward value at time n, $0 \le \rho \le 1$ is the discount factor, $E(\ )$ represents the expected reward value by following a strategy σ , $V(x,\sigma)$ is the value of state x under strategy σ.

The standard solution to (2) is through iterative search method that searches for fixed point of Bellman equation [17]

$$V^*(x) \equiv V^{\sigma*}(x)$$

$$= \max_{e\in A(x)} \{\, u_x(e) + \rho \sum_{x'} T_{xx'}(e)\, V^{\sigma^*}(x')\,\} \qquad (3)$$

Where $T_{xx'}$ is the transition probability to next state x' after performing an action e in state x , $u_x(e)$ is the reward value and σ^* is the optimal strategy. An agent finds an optimal strategy σ^* without initially knowing the reward and state transition probability values. For an optimal strategy σ^*, Q-value is

$$Q^*(x,e) = u_x(e) + \rho \sum_{x'} T_{xx'}(e)\, V^{\sigma^*}(x') \qquad (4)$$

Where $Q^*(x,e)$ is the total discounted reward obtained by selecting an optimal strategy in state x.

If $Q^*(x,e)$ is known then optimal strategy σ^* can be found by identifying an action that maximizes $Q^*(x,e)$ under state x. Instead of searching for optimal value $V(x,\sigma^*)$, the problem is reduced to finding function $Q^*(x,e)$ given by

$$V(x,\sigma^*) = \max_{e\in A(x)} \{Q^*(x,e)\} \qquad (5)$$

Bellman optimality equation expresses the fact that value of state x under an optimal strategy must be equal to the expected reward for performing action in that state can be represented as

$$V^*(x) = \max_{e\in \mathrm{A}\,(x)} Q^{\sigma*}(x,e)$$

This is further expressed as

$$V^*(x) \;=\; \max_{e\in A(x)} \sum_{x'} T^e_{xx'} + \rho V^{\sigma*}(x')$$

$$= \sum_{x'} T^e_{xx'} + \rho \max_b Q^*(x', b) \qquad (6)$$

Where b is the action performed in the next state x'. Q-Learning provides with simple updating procedure in which an agent arbitrarily starts with initial values of $Q(x,e)$ and updates Q-values for time n+1 as follows:

$$Q_{n+1}(x_n, e_n) = (1-\beta) Q_n(x_n, e_n) + \beta \left[u_n + \rho \max_b Q_n\,(x', b)\right] \qquad (7)$$

Where β is the learning rate, $0 \le \beta \le 1$. It is established from (7) that the sequence converges to $Q^*(x,e)$ under the assumption that states and actions have been visited indefinite times often and learning rate satisfies certain constraints [18].

4.2. Multi agent Q-learning

Reinforcement Learning is mainly for learning about the single agent environment which is solved by using MDP. In MAS, the behaviour of agents can be studied using Game theory. Mathematically, it can be formulated as m agent stochastic game which is a tuple of $< m, X, A_{1\ldots m}, U_{1\ldots m}, T >$ Where m is the number of agents, X is the set of states, A_i is the set of actions available to agent i (i=1..m) , U_i: $X.A_i$ is the set of reward values. This looks very similar to single agent MDP except selecting the actions, next state and reward depending on the joint action of the agents.

The competition among agents can be formulated using game theoretic approach. Given state x, agents choose actions $e_1 \ldots . e_n \in A_i$ and changes to next state x' based on fixed transition probability satisfying the constraint.

$$\sum_{x'\in X} p(x'|x, e_1 \ldots . e_n) = 1 \qquad (8)$$

The goal of each agent in a stochastic game is to maximize the discounted sum of rewards. Let σ_i be the strategy of agent i. For a given initial state x agent i tries to maximize the reward. Now (2) can be re-written as

$$V^i(x, \sigma_1, \sigma_2 \ldots . \sigma_n) = \sum_{n=0}^{\infty} \rho^n E(u_n^i \mid \sigma_1, \sigma_2 \ldots . \sigma_n, x_0 = x) \qquad (9)$$

In a stochastic game $\prod$, a Nash equilibrium point [19] is a tuple of n strategies $\sigma_1^* \ldots \ldots \sigma_n^*$ such that for all $x \in X,\ i = 1 \ldots n$. Now (9) can be re-formulated as

$$V^i(x, \sigma_1^* \ldots \ldots \sigma_n^*) \ge V^i(x, \sigma_1^* \ldots \sigma_{i-1}^*, \sigma_i^*, \sigma_{i+1}^* \ldots \sigma_n^*) \forall \sigma_i \epsilon \prod_i \qquad (10)$$

Where $\prod_i$ is the set of strategies available to agent i.

4.2.1 Nash Q-Learning

The Q-function for agent i in multi agent scenario can be obtained by considering joint actions rather than individual actions. Agent $i's$ Nash Q-function [19] is defined over $(x, e_1 \ldots . e_n)$, as the sum of Agent $i's$ current reward plus

its future rewards when all the agents follow a joint Nash equilibrium strategy. Mathematically, it is expressed as

$$Q_*^i(x, e_1 \ldots . e_n) = u_i(x, e_1 \ldots . e_n) + \rho \sum_{x' \in X} p(x' | x, e_1 \ldots . e_n) V^i(x', \sigma_1^* \ldots \sigma_n^*) \quad (11)$$

Where $\sigma_1^* \ldots \sigma_n^*$ is the joint Nash equilibrium strategy, $u_i(x, e_1 \ldots . e_n)$ is agent *i's* total discounted reward over infinite periods given that agent follows the equilibrium strategies.

The Nash Q-Learning equation for agent '*i*' at time n+1 is given by

$$Q_{n+1}^i(x, e_1 \ldots . e_n) = (1 - \beta)\ Q_n^i(x, e_1 \ldots . e_n) + \beta\left[u_n^i + \rho\ Nash\ Q_n^i(x')\right] \quad (12)$$

Where $Nash\ Q_n^i(x') = \sigma_1(x') \ldots \sigma_n(x').\ Q_n^i(x')$.

5. FORMULATION OF RESOURCE ALLOCATION IN CR BASED LTE NETWORKS

We consider a game theoretic approach in which each eNodeB act as an agent and try to find out the best resource blocks from the egoistic cognitive environment to arrive at the optimal strategy. The key component for this environment is the reward function. In this section, we first represent the average throughput calculation for an epoch.

Let x_n represent TTI, α as time window, and $R_j(x_n)$ is the data rate of user *j*. The average throughput $T_j(x_n)$ for each user *j* is given by,

$$T_j(x_{n+1}) = \begin{cases} \left(1 - \frac{1}{\alpha}\right) T_j(x_n) + \frac{1}{\alpha} R_j(x_n), j \in j' \\ \left(1 - \frac{1}{\alpha}\right) T_j(x_n), \qquad j \notin j' \end{cases} \quad (13)$$

Where j' is the scheduled user set. It is essential to identify the environment state, action and reward.

5.1 State, Action and Reward

The environment state is defined as x_n which is considered as state of RBs in each TTI. The action is to find out the free RBs. It can be represented as

$$e_i = \begin{cases} 1, & if\ Idle\ RBs \\ 0, & otherwise \end{cases} \quad (14)$$

The reward $u_i(x_i, e_i, e_{-i})$ of cognitive eNodeB in state x_n is the immediate return due to the exploitation e_i, when all other eNodeBs choose e_{-i}. i.e,

$$u_i(x_i, e_i, e_{-i}) = \begin{cases} +1\ if\ RB\ is\ idle\ in\ x_n \\ +2\ if\ RB\ is\ idle\ in\ x_n, x_{n+1} \\ 0\ else \end{cases} \quad (15)$$

5.2. Harmonized Q-Learning (HQL)

To adapt Q-Learning in multi agent context, we propose harmonized Q-Learning algorithm where the problem is analyzed in two distinct ways: learning and coordination. The Q-learning is identified to tackle the problem of learning that discovers optimal Q function for each agent. The problem of coordination is undertaken using Nash Q-Learning. The knowledge of optimal Q-function is not enough to ensure that all the agents in an environment adopts jointly optimal policy.

Our proposed work ensures that all the agents converge to a joint optimal policy in every relevant state of the game. The resource blocks are allocated for each user in the presence of multiple optimal strategies with the help of two approaches: (i) Simultaneous play mode and (ii) Alternate play mode.

5.2.1. Simultaneous play mode

In this mode, the agents simultaneously share their available RBs equally through joint coordination and learning among other agents.

Definition 1

A joint strategy $(\sigma_1 \ldots . . \sigma_n)$ constitutes Nash equilibrium for the stage game for

$$\sigma_i \sigma_{-i} N_i \geq \widehat{\sigma}_i \sigma_{-i} N_i\ , \forall \sigma_i \in \widehat{\sigma}_i(e_i)$$

Proof: The learning agent indexed by *i* learns about its Q-values by forming arbitrary guess at time 0. One simple guess would be $Q_i^0(x, e_i) = 0, \forall x \in X, e \in e_i, i = 1..n$. In each TTI agent observes the current state and action and after observing its own reward it needs to know other agents reward and new state x'. It then computes Nash equilibrium $\sigma_1(x_1) \ldots . \sigma_n(x_n)$ for the game and update its Q-values.

Lemma: Two joint strategies are [19]:

A. $(\sigma_*^1 \ldots \ldots \sigma_*^n)$ is a Nash equilibrium point in a discounted stochastic game with equilibrium reward $(V^1(\sigma_*^1 \ldots \ldots \sigma_*^n) \ldots V^m(\sigma_*^1 \ldots \ldots \sigma_*^n)$where $V^k(\sigma_*^1 \ldots \ldots \sigma_*^n) = (V^k(x^1, \sigma_*^1 \ldots \ldots \sigma_*^n), \ldots V^k(x^m, \sigma_*^1 \ldots \ldots \sigma_*^n)\ , k = 1, . . m$

B. For each $x \in X$, the tuple $(\sigma_1^*(x) \ldots \sigma_n^*(x))$ constitutes a Nash equilibrium point in the stage game $(Q_*^1(x) \ldots \ldots Q_*^n(x))$ with Nash equilibrium reward $(V^1(x, \sigma_*^1 \ldots \ldots \sigma_*^n), \ldots V^n(x, \sigma_*^1 \ldots \ldots \sigma_*^n))$.

This lemma links agent $i's$ optimal value V^i for the entire stochastic game $Q_*^1(x) \ldots \ldots Q_*^n(x)$. In other words, $V^i(x) = \sigma_1(x) . . \sigma_n(x) Q_*^n(x)$. By using (11) the Q-function for our approach can be obtained by considering other agents strategy. Equation (11) can be re-formulated as

$$Q_*^n(x_i, e_i) = E\left[u_i\left(x_i, e_i, \sigma_*^{-i}(x_i)\right)\right] + \rho \sum_{x' \in X} T_{xx'}\left(e_i, \sigma_*^{-i}(x_i)\right) \max_{b_i \in A_i} Q_*^i(x', b_i)\ , i = 1..n \quad (16)$$

HQL tries to find the optimal Q-value Q_*^i using the information$< e_i, x_i, x', \sigma_i, \sigma_{-i} >$available from the environment at time $n\ and\ n+1$. The Q-value is updated by combining the old value and the new expected reward. The proposed algorithm not only needs its eNodeB's own information but also considers the strategies of other eNodeBs.

5.2.2 Alternate play mode

In a fully aggressive environment with alternate play mode, an agent gets an opportunity of utilizing the free resource blocks based on priority and demand. For multi agent N_i , the set of strategies in (10) is $\prod_i \in \sigma_i(e_i)$, the Nash-Q function can be given as

$$Nash(x, Q_1 \dots Q_n) = \max_{\sigma_i(e_i)} \min_{\Sigma_{i=1}^{n} \prod_i - \sigma_i(e_i)} \sum_{\sigma_i(e_i)} \sigma(e_i) Q_i(x, e_i), \forall i = 1..n \quad (17)$$

6. RESULTS AND DISCUSSIONS

The simulation was carried out on the LTE System Level Simulator [20]. We consider three Operators A, B and C in LTE environment. The total resource available to any one operator consists of its own plus other free resource from other operators. To create environment for cognitive radio in the simulation setup, let Operators A and B have 5 MHz bandwidth each (25 RBs) whereas Operator C has 10 MHz bandwidth (50 RBs). At a particular instance, let Operator C has some un-used RBs (say 20%) that can be shared by operator A and B using the Harmonized Q Learning (HQL). The main simulation parameters are listed in Table 1.

We assume that each Operator's eNodeB sector has the channel state information available for every TTI and hence it knows which RBs are free belonging to other operators. Q-values are calculated for every RB in Operator C's Band based on the Q-Learning algorithm. Out of those available free RBs, best RBs are chosen based on the high Q-Values. Free RBs are given rewards based on the CQI and its probability of availability.

Each operator's eNodeB maintains a Q-table for the available RBs and HQL algorithm is implemented in all the cognitive eNodeBs. In simultaneous play mode, all the eNodeBs share a strategy such that the available RBs are shared among them equally in every TTI. Based on their Q-table, each operator will identify the best available RBs to use them but some RBs would be best for more than one operator and such kind of RBs are shared among them equally.

In alternate play mode, common free RBs among operators are shared in a way that fulfills the demand of operators in that region. This is the more practical way of implementing HQL algorithm and we have simulated both simultaneous and alternate modes.

Table1. Simulation Parameters

Parameter	Values
Frequency band	2.14GHz
TTI length	1 ms
Sub carriers per RB	12
Sub carrier spacing	15 KHz
AMC levels	QPSK,16-QAM, 64QAM
Macroscopic path Loss	TS36942, Urban
Minimum Coupling loss	70
Transmit Mode	Closed Loop Special Multiplexing (CLSM)
FFT points	2048
Antenna azimuth offset	30
Shadowing	Log-normal distribution
Channel model	Winner model
Scheduler	Proportional fair
Number of eNodeB Sectors	19x3 = 57
UEs per Sector	10
Learning Rate (β)	0.8
Discount Rate (ρ)	0.7

Fig.5 shows the eNodeB and UE positions. There are totally 19 eNodeBs (dark red dots). The shaded portion (7 eNodeBs) in the figure indicates the eNodeB Sectors under active consideration. Blue dots represent the active UEs whereas light red dots represent passive UEs.

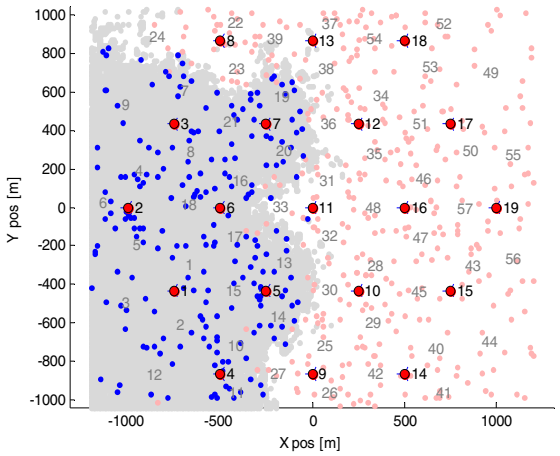


Figure 5. eNodeB and UE location in simulation setup

Fig.6 is a scatter-plot shown for each UE mapping between the UE wideband SINR and average UE spectral efficiency. Similarly Fig.7 is another scatter-plot mapping for the average throughput considering each UE mapping between the UE wideband SINR. In both Fig.6 and Fig.7, binned (over wideband SINR) mean throughput mapping is shown in red.

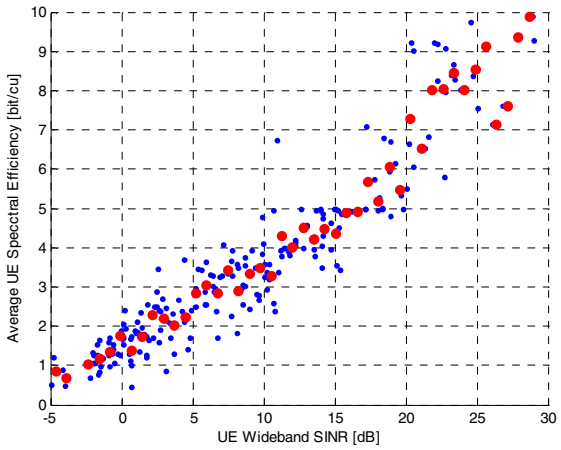


Figure 6. UE wideband SINRvs. average UE spectral efficiency

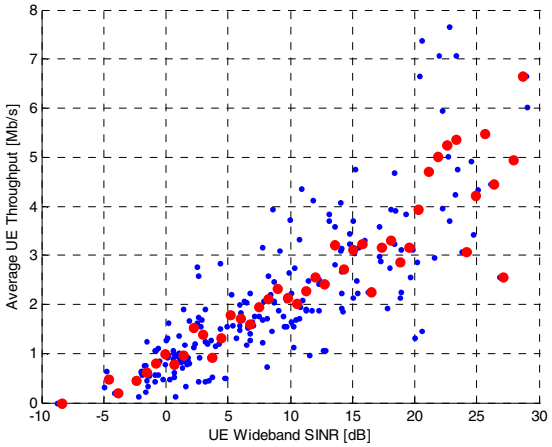


Figure 7. UE wideband SINR vs. average throughput.

The Empirical Cumulative Distribution Function (ECDF) is plotted in Fig.8, Fig.9 and Fig.10 for the UE wideband SINR, average spectral efficiency and average UE throughput respectively.

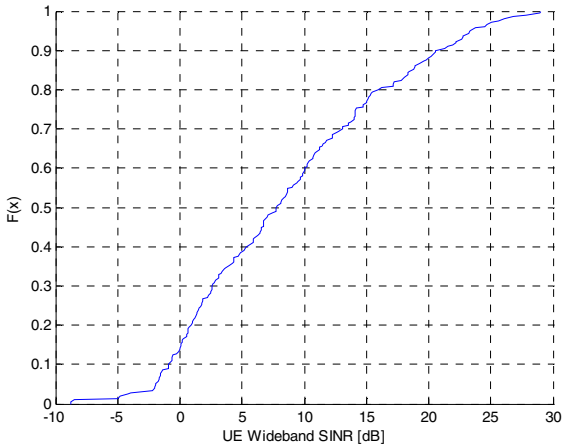


Figure 8. ECDF of UE wideband SINR

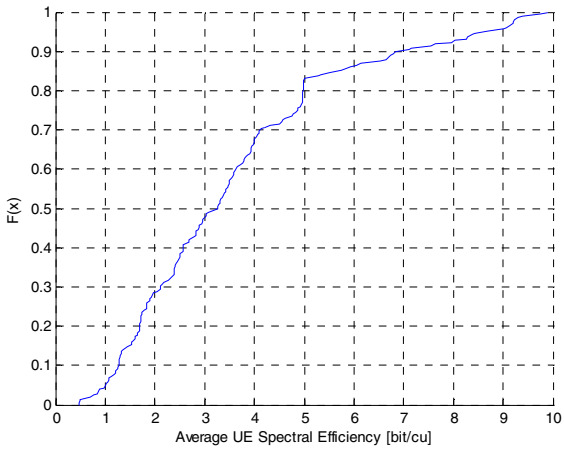


Figure 9. ECDF of average UE spectral efficiency

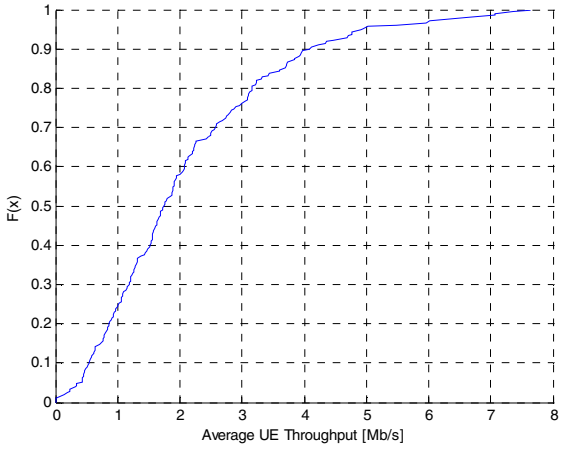


Figure 10. ECDF of average UE throughput

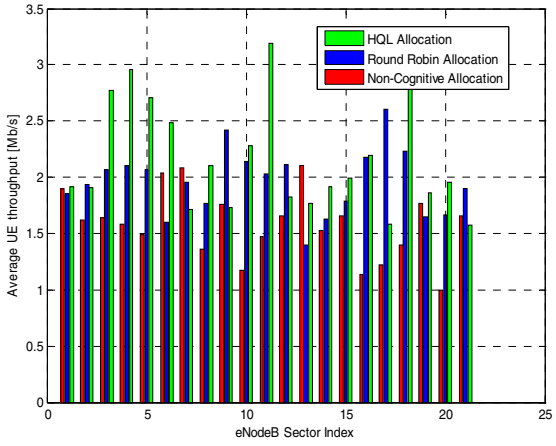


Figure 11. Average UE throughput observed in various eNodeB sectors

Average UE throughput observed at eNodeB sectors during simulation in multi operator scenario for three different packet scheduling schemes: HQL, Round Robin, and Non-Cognitive scheduling is presented in Fig.11. In Non-Cognitive allocation, free RBs are not considered in

resource pool. Alternate play mode of HQL algorithm involves sharing the common best RBs among operators based on the demand basis. It is evident from Fig.11 that HQL based resource allocation provides better throughput compared to other allocations in most of the occasions during simulation. The reason for better throughput in HQL is because of the application of intelligent learning and coordination which allocates best RBs considering the other operator's strategy and traffic demands.

7. CONCLUSION

We proposed a harmonized Q-Learning algorithm for radio resource management in LTE based CR networks. This approach employs Nash Q-Learning which considers the strategy of other agents and tries to find out the optimal solution. Q-Learning was modeled in single and multi-agent scenario. We also formulated resource allocation problem in LTE based CR networks. The system level simulation in LTE platform for the proposed HQL algorithm provided higher throughput while meeting the real-time resource requirement in multi-operator system. Implementation of HQL in LTE-Advanced simulation environment could be the future work.

ACKNOWLEDGEMENT

We thank the Institute of Telecommunication, Vienna University of Technology, Vienna for releasing their LTE simulator in free download version. We also thank the Society for Applied Microwave Electronics Engineering and Research (SAMEER), Chennai for providing us real-time data on spectrum usage.

REFERENCES

[1] JayakrishnanUnnikrishnan, and Venugopal V. Veeravalli "Algorithms for Dynamic Spectrum Access with Learning for Cognitive Radio", IEEE Transaction on Signal Processing, Vol No.2, PP. 750-760, Feb. 2010.

[2] ITU-R M.2242, "Cognitive radio systems specific for International Mobile Telecommunications systems", M Series Report ITU-R (11/2011).

[3] 3GPP TR 36.913, V.9.0.0 "Technical Specification Group Radio Access Network; Requirements for further advancements for Evolved Universal Terrestrial Radio Access (LTE-Advanced) (Release 9)", December 2009.

[4] Junfeng XIAO1, Feng YE, Tingjian TIAN, and Rose Qingyang HU, "CR Enabled TD-LTE within TV White Space: System Level Performance Analysis", Proceedings of IEEE Globecom 2011.

[5] YuhuaXu, Jinlong Wang, Qihui Wu, Alagan Anpalagan, and Yu-Dong Yao, "Opportunistic Spectrum Access in Unknown Dynamic Environment: A Game-Theoretic Stochastic Learning Solution", IEEE Transaction on Wireless Communication, Vol 11, No.4, PP. 1380-1390, April 2012.

[6] Dhananjay Kumar, Kanagaraj N.N "Radio Resource management in OFDMA-CRN considering primary user activity and detection scenario", Proceedings of 4th ITU-T Kaleidoscope Event: The Fully Networked Human? Innovations for Future Networks and Services, Cape Town, South Africa, 2-14 Dec 2011.

[7] Dhananjay Kumar, S. Mahalaxmi, J. Sharad Kumar, and R. Ramya, "Adaptive Resource Allocation for Real-time services in OFDMA Based Cognitive Radio Systems," Proceedings of ITU-T Kaleidoscopic Event: Beyond the Internet? - Innovations for Future Networks and Services, Pune, India, 13-15 December 2010.

[8] IoanSorinComsa, Mehmet Aydin, Sijing Zhang, Pierre Kuonen, Jean-Frederic Wagen ,"Multi Objective Resource Reinforcement learning", International Journal of Distributed Systems and Technologies, 3(2), PP 39-57,April-June 2012.

[9] Husheng Li, "Multi–agent Q-learning for Competitive Spectrum Access in Cognitive Radio Systems" Proceedings of fifth IEEE Workshop on Networking Technologies for Software Defined Radio Networks, 21 June, 2010.

[10] Zhe Chen and Robert C.Qiu, "Q-learning Based Bidding Algorithm for Spectrum Auction in Cognitive Radio", Proceddings of IEEE South East Conference (SECON-2011), 17-20 March2011.

[11] Mo Li, Youyun Xu, and Junquan Hu, "Q-learning Based Sensing Task Selection Scheme for Cognitive Radio Networks", IEEE International Conference on Wireless Communication and Signal Processing,13-15 November, 2009.

[12] Aggelos Saatsakis, Kostas Tsagkaris, Dirk von-Hugo, Matthias Siebert, Manfred Rosenberger, and Panagiotis Demestichas, "Cognitive Radio Resource Management for Improving the Efficiency of LTE Network Segments in the Wireless B3G World", IEEE Symposium on New Frontiers in Dynamic Spectrum Access Networks, October 2008.

[13] Xinsheng Zhao, Zhiyi Guo, Qiang Guo, "A Cgnitive based Spectrum Sharing Scheme for LTE – Advanced Systems, Proceedings of International Congress on Ultra Modern Telecommunications and Control systems and Workshops (ICUMT), 2010.

[14] Mehdi Bennis and Dusit Niyato, " A Q-learning Based Approach to Interference Avoidance in Self-Organized Femtocell Networks" , Proceedings of IEEE Globecom Workshop on Femtocell Networks, 2010.

[15] Sudharman K .Jayaweera, Mario Bkassiny ,keith A. Avery, " Asymmetric Cooperative Communications based Spectrum Leasing via Auctions in Cognitive Radio Networks", IEEE Transactions on Wireless Communications, Vol.10, No.8. August 2011.

[16] Xianfu Chen, Zhifeng Zhao and Honnanag Zhang, "Stochastic Power Adaptation with Multi agent Reinforcement Learning for Cognitive Wireless Mesh Networks", IEEE Transactions on Mobile Computing, Vol. x, No. x, 2012.

[17] Richard S. Sutton and Andrew G. Barto, "Reinforcement Learning: An introduction", MIT press, Cambridge, England, 1998.

[18] Christopher J.C.H. Watkins, and Peter Dayan, "Q-Learning A technical Note", Kluwer Academic Publishers, Boston, Vol No.8, PP 279-292, Year 1992.

[19] Junling Hu, and Michael P. Wellman, "Nash-Q-learning for General Stochastic Games, Journal of Machine Learning Research Vol No. 4, 2003.

[20] J.C.Ikuno, M. Wrulich, and M.Ruppo, "System level simulation of LTE networks," in Proc. 2010 IEEE 71st Vehicular Technology Conference, Taipei, Taiwan, May 2010. [Online] Available: http://publik.tuwien.ac.at/files/PubDat_184908.pdf.

SESSION 5

SUPPORTING FUTURE APPLICATIONS

HYBRIDCAST: A NEW MEDIA EXPERIENCE BY INTEGRATION OF BROADCASTING AND BROADBAND

Hisayuki Ohmata, Masaru Takechi, Shigeaki Mitsuya, Kazuhiro Otsuki, Akitsugu Baba, Kinji Matsumura, Keigo Majima, and Shunji Sunasaki

Science & Technology Research Laboratories
NHK (Japan Broadcasting Corporation), Japan

ABSTRACT

Broadcasting has a role for the public service. Providing the same information to a large number of people at the same time has benefitted modern society in many ways, including presenting the forefront of lifestyle trends, offering dependable media during disasters and cost-effective transmissions. On the other hand, services over the Internet satisfy the individual's needs as seen in customization for each, interactive communication and user-generated media. NHK is developing "Hybridcast", a service platform integrating broadcasting with the Internet. This platform can enhance broadcasting programs and provide other various services by the best mix of features of both media. This paper describes the system and examples of service on Hybridcast for the general public including minority viewers. The next-generation media for a sustainable society will emerge from Hybridcast which is expected to be launched in 2013.

Keywords— Hybridcast, Interactive Broadcast Broadband System, Multi-screen Service, HTML5

1. INTRODUCTION

The digitization of broadcasting and the rapid spread of broadband environment have led to an information infrastructure that provides high-quality digital broadcasting and a variety of Internet services. A new era will soon dawn when broadcasting and the Internet will work together seamlessly to provide new services. NHK is developing "Hybridcast"[1] for a new era expected to launch in 2013.

Broadcasting guarantees a quality of service (QoS) and an efficient delivery of content to a large number of viewers, whereas Internet systems afford the flexibility to deliver personalized content that meets an individual viewer's preferences. NHK has been contributing to both technical and service developments for combining broadcast and Internet for a long time, taking into account their complementary characteristics, i.e., the efficient delivery of high-quality content on a large scale and the delivery of services tailored to a user's preferences, respectively.

Hybridcast is one of the hybrid broadcasting systems to combine broadcasting and network functionalities seamlessly. A hybrid broadcasting system accepts wide range of services. In general, the services can be categorized into following two types;

a) Broadcast-related services
   These are services strongly tied with broadcasting programs or content. For example, the services are intended to add more information to broadcasting content to enhance it. Though interactive TV services and data broadcasting have provided similar features, hybrid broadcasting systems offer much wider flexibility because network connectivity allows handling much more data than broadcasting channels and offering what individuals need, such as video-on-demand.

b) Independent services
   Like services on a smartphone such as Social Network Services (SNS) or games, independent services are not related to broadcast programs or content.

In hybrid broadcasting systems, the flexibility of the systems allows, for broadcast-related services, to provide additional content elements to make the broadcasting programs or content understandable better for every kind of people including the elderly, people with disabilities and minorities. The flexibility also allows important independent services to take care of every kind of people such as e-health.

Hybridcast is the service platform to enable such services and to bring new TV experience for all viewers. This paper describes the examples of services, its system concept and its architecture.

2. SERVICES ON HYBRIDCAST

Some example services on Hybridcast have been developed to verify system functionalities. Accessibility improvement, provision of better presentation of the event from many aspects, and protection of life and properties of people are taken into account for the examples. In this chapter, such broadcast-related example services are described.

2.1. Multilingual Closed-captioning Service

According to Japanese digital broadcasting standard, closed caption can be provided for up to two languages within a broadcast channel. Two languages may not be enough for some people including minority or foreign travelers. Broadcasting service will target majority in language firstly because broadcasting system is a system to deliver the same information to mass viewers very efficiently. If a viewer would like to watch closed caption in a language not

delivered over the broadcast channel, the Internet can be used to deliver closed caption data in the preferred language. Hybridcast has a broadcast-broadband synchronization mechanism which enables the synchronization of various types of streams from multiple sources. The closed caption server feeds time-stamped caption data in the requested language and they are presented synchronously with the broadcast program.

Figure 1 shows an example image of this service. In Figure 1, seven languages are available. A viewer selects the preferred language by using a remote controller, and closed caption in selected language is on the screen. Under some conditions, the service provider may offer "user generated closed caption" which can extend the language selection.

By this service, closed caption service even in minor languages can be available without buying a dedicated receiver so that the people can get information through TV set in understandable or preferred language.

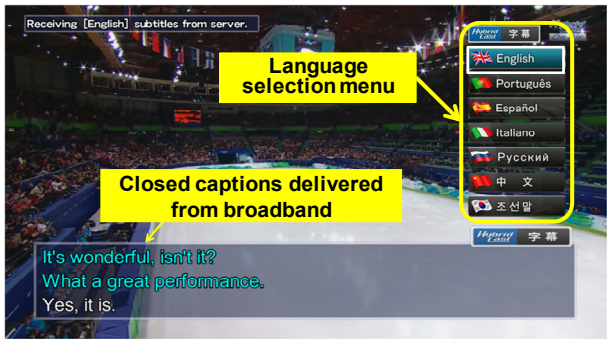


Figure 1. Multilingual closed-captioning service

2.2. Sign Language Animation Service

Closed caption is provided with many TV program in Japan. Although this is a useful service for the deaf, there is a need to provide another type of the service to them. For some people, textual representation is something like a foreign language and the so-called mother-tongue is the sign language. In case of an emergency, provision of sign language interpretation to broadcasting program is a critical issue for such people. Reflecting the situation, there is a strong demand from the deaf community for sign language interpretation to broadcasting programs. However, few TV programs are broadcast with a sign language service for the deaf. The reason of being is the insufficient number of sign language interpreters. And probably they will not be available for an emergency broadcasting program at midnight. Considering that the availability of sign language interpreter will not be improve soon, it is important for the deaf community to make sign language interpretation available at all times even in alternative ways.

For this purpose, NHK is developing a technology to generate sign language Computer Graphics (CG) animation by using TVML (TV program Making Language)[2]. The combination of this technology and Hybridcast can provide sign language animation services.

It is important for broadcasting service compatibility of sign language interpretation for people who want and who don't want the service. Sign language interpretation should be switchable at the receiver side, and video image of sign language interpretation should not be overlaid at the broadcast stations. So, video image of sign language should be delivered independently from the broadcasting video image. However, normally broadcast channel is fully occupied by regular broadcasting data and there is no space in broadcast channel to provide switchable animation video images.

When using Hybridcast, sign language animation video images can be delivered over the Internet. The Hybridcast application for CG sign language running on a receiver requests additional video images of sign language interpretation, and by using the broadcast–broadband synchronization mechanism, a sign language animation is presented synchronously with broadcast program (Figure 2). Because TVML technology can generate sign language animation from textual scripts, emergency information can be provided in sign language even at a time when human sign language interpreters are not available. This will help the deaf to evacuate or take an appropriate action in case of an emergency.

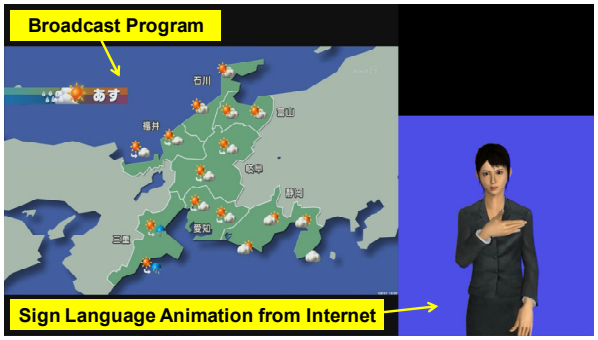


Figure 2. Sign language animation service

2.3. Multiangle Camera Service

Many of the TV programs use multiple cameras during production and each camera shoots different objects or the same object from different angles. It is quite normal for broadcasters to select one of the images as a part of the program production process. On the other hand, viewers may sometimes want to watch the image from another camera. Sports or music program may be typical programs for such needs. For example, during watching a jazz live program, a viewer is interested in each player, on piano, drum, bass, sax and so on. Watching the object on a TV set from multiple angles helps viewers to understand the ongoing event better. Hybridcast with the application for multi-angle satisfies the needs.

This service provides videos of multiple cameras over the Internet. Basic mechanism for synchronization of two video images is the same as for the two services above. The Hybridcast application running on a receiver allocates multiple video images and, in some cases, the selected image can be enlarged by the application. Figure 3 shows an example of a screen image for the service.

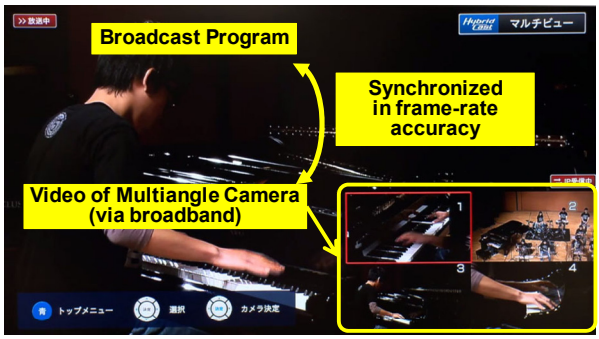


Figure 3. Multiangle camera service

2.4. Language Study Service

Providing an educational broadcasting program is one of the important elements of a public service. A language study program is a typical example of this. Hybridcast can provide a language study service with a secondary screen device (Figure 4), which contributes an efficient and effective education. An application on a tablet is used to ask questions about conversation in the program. Viewers can answer the questions using touch controls.

Synchronization of the application on the tablet is achieved by using a trigger signal within a broadcast channel and multiple-device linkage function in the Hybridcast receiver. A broadcaster sends the trigger signal when asking a question. Once the application on a TV set receives the trigger, it tells the application on the tablet to acquire data of the question over the Internet and to start asking it.

Such a mechanism and user interface of a tablet may introduce viewers to new experience of language learning. Especially, touch interface makes learning fun and viewers may not get tired quickly.

Use of the secondary screen device as an interactive text book can be applied to other subjects of education. It will allow reviewing the matters of wrong answers later very easily, and help effective learning.

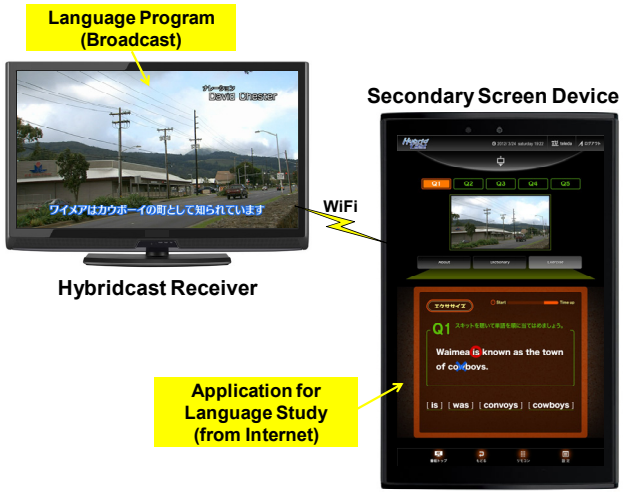


Figure 4. Language study service

2.5. Social TV Service

One of the highlights of TV viewing is to spend time and discuss with family or friends while watching TV programs, especially in live broadcasts. It is quite natural to do so when a family shares the single TV set and watches the same program together. By integrating with an SNS, Hybridcast enables people who are not in the same location to share the feelings as if they were nearby. It is so-called Social TV[4].

Figure 5 shows a chat application example. It lists friends who are watching the same program on top-right of the screen by icon. On the bottom-right, chat messages are displayed.

The Hybridcast application manages the login status of SNS and watching status of the program. The application obtains the ID of program from the receiver. The application then sends the ID of program to the server for the SNS. The friends or family who are logged in and have registered the same ID of program become chat members. Thus, Social TV service can create a virtual space for TV experience and overcomes physical distance among friends or family.

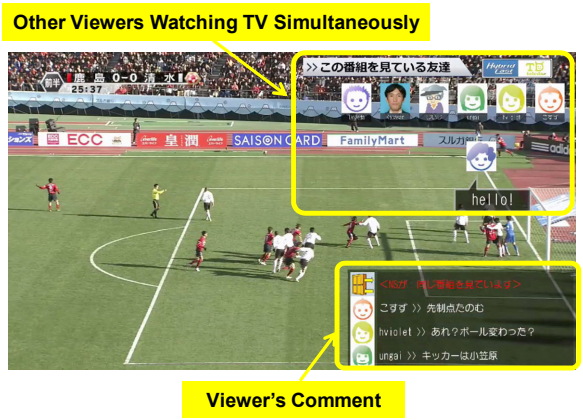


Figure 5. Social TV service

2.6. Prioritized Presentation of Important Information

One of the important roles of broadcasting service is providing emergency information to viewers accurately and quickly as much as possible. This is one of the primary functions of broadcast to communities in case of an emergency. In Japan, the government deployed alert systems against large earthquakes and tsunamis. When broadcasters receive emergency information from the authorities, they have to provide this information to viewers as quickly as possible[5][6]. These systems greatly helped to escape from the disaster when the Great East Japan Earthquake occurred in March, 2011.

This characteristic of broadcast has to be maintained in Hybridcast as well. However, in Hybridcast, two kinds of media share on one TV screen: broadcast video and applications from the Internet, which may be independent each other. If emergency information is broadcast and overlaid information onto the broadcast video hides some pieces of information provided by the broadcasters, there is a risk that viewers may miss the important information.

Especially, earthquake alert is broadcast a few seconds before the quake hits. There is no time to manipulate something by the viewers. The time remaining may be comparable to watching the screen and understanding what is going to happen.
A presentation control in Hybridcast will be engaged to reduce such risks[7]. Figure 6 shows prioritized presentation of broadcast program which conveys emergency information. A receiver continuously monitors broadcasting signal to detect emergency signal information embedded in it. As soon as such a signal is detected, the receiver controls to enlarge the broadcast video to full-screen and removes the application which is running from view. After the signal is turned off, the former layout is resumed.

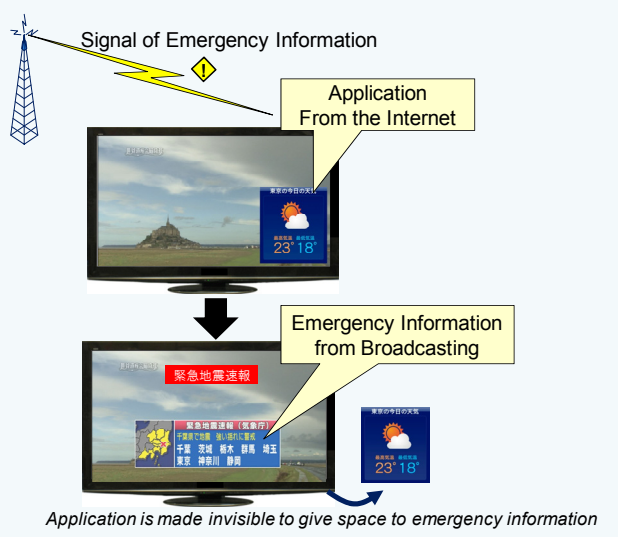


Figure 6. Use case of the presentation control

3. PLATFORM DESIGN

3.1. System Concept

These services are just the tip of the iceberg and so many services that we cannot count will emerge on Hybridcast. Provision of flexibility and service expandability to build a variety of services is the most important characteristics for the Hybridcast platform. Application-oriented approach and introduction of various players, not only broadcasters but also third-party service providers, into the overall system are keys to offering a diversity of services .

3.2. System Requirements

As a first step for system development, the system requirements are determined based on an analysis of the use cases. Most of the requirements derived as a result of the analysis are consistent with those defined in Recommendation ITU-T J.205, "Requirements for an application control framework using integrated broadcast and broadband digital television". Among the requirements of Hybridcast system, five of the most fundamental requirements are described below.

3.2.1. Compatibility with existing broadcasting system

Hybridcast system should be compatible with existing broadcasting systems. In Japan, total shipments for DTV (Digital Television) receiver are over one hundred million. It is impractical to introduce new incompatible broadcasting systems dedicated to Hybridcast. In addition, coexistence with a data broadcasting service is required. Data broadcasting services are widely used and popular. In some cases, it is necessary to start Hybridcast service from a data broadcasting service, and vice versa.

3.2.2. Application management

A viewer can experience various services through the Hybridcast application. This means methods to start and stop the application is essential to the Hybridcast system. For example, when a viewer wants to participate in a quiz show using a Hybridcast application, the application should start at the time of the broadcast program automatically, and stop at the end of the program automatically. On the other hand, there is an application which is independent from a broadcast program such as weather forecast application. This kind of applications should start under the instruction by the viewers at any time, and stop similarly. The application management mechanism in Hybridcast should satisfy these different start and stop scenarios.

3.2.3. Broadcast resource access

The close combination of broadcast and application enables the creation of a new TV experience. Hybridcast applications should be able to access broadcast resources such as video, audio and metadata and to use them TV functions such as tuning to new channel. However, access to those resources should not affect the content integrity of the broadcast.

3.2.4. Receivers

The receiver is one of the key equipment to build the services on the Hybridcast platform. The receiver finally makes presentation, provides interactivity for viewers, handles various control signals, and offer functionalities to the application. If the receiver provides multiple video decoders, it will allow the service to combine multiple video images on the same screen.
Various smart devices such as tablets and smartphones using user-friendly interface have become popular. On Hybridcast, these devices should be available as a remote controller and a screen for displaying information. To use such devices as 'companion devices', it is essential for a receiver to equip link and communication mechanism to them.

3.2.5. Security

Introduction of the third-party service providers will make the range of services wider. Service expandability is a fundamental factor of the system concept of Hybridcast as

described in chapter 3.1, and Hybridcast welcomes third-party service providers. However, security level of third-party application or services may vary for the each application or service. From this viewpoint, security feature of Hybridcast is more important than for existing broadcasting systems.
The security mechanism in Hybridcast platform should protect interests of stakeholders including broadcasters and viewers. For viewers, unintended access to personal information by the application is one of the risks. For broadcasters, unauthorized overlaying of graphics or text onto broadcast video by the application breaks content integrity. The security mechanism to take control of access to the information or behavior of the application in proper manner will bring safety and comfort to stakeholders, so that more service providers will offer the services, or more viewers enjoy the services.

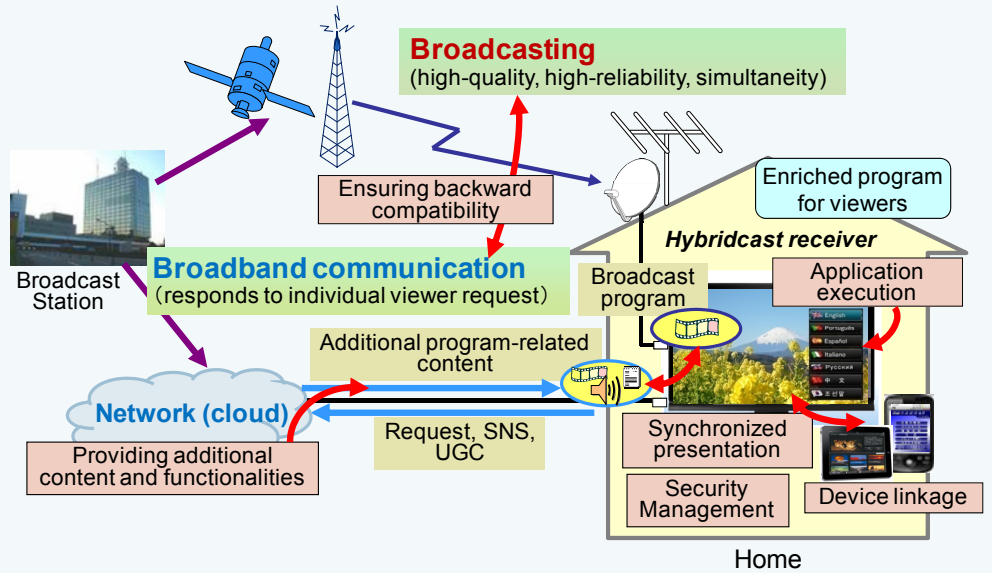


Figure 7. Conceptual diagram of Hybridcast

4. SYSTEM ARCHITECTURE

A conceptual diagram of the Hybridcast is shown in Figure 7. The platform consists of the existing broadcasting system, additional servers in the cloud, and Hybridcast receivers. The receiver functionalities include application execution, the handling of broadcasting resources, multi-device linkage, content synchronization and security management.

4.1. System Overview

To build such a platform, we designed a simple system architecture. The Hybridcast system utilizes the existing broadcast system as much as possible. Figure 8 shows the network part of Hybridcast's basic system architecture. The system consists of three blocks: broadcaster servers, service provider servers, and receivers. The broadcaster servers provide broadcast content and content-related information, which only the broadcasters hold, to the service provider servers. The service provider servers provide applications, content, and relevant information to the services or end users. Hybridcast receivers execute applications to realize various services. Such an application-oriented approach enables a rather simple receiver-side implementation.
Hybridcast applications running on a Hybridcast receiver process the content and relevant information obtained from the service provider servers. The applications on a receiver can access to required information from broadcast and the network, and call functions. Service providers are not only broadcasters but also third-party service providers.

4.2. APIs – Interfaces between Three Entities –

One of the main features of Hybridcast platform design is a structure of the three entities: broadcaster, service provider and receiver (Figure 8). Each entity provides a specific function to other entities via APIs (Application Programming Interfaces). In other words, among these three entities, API is a "grew" to exchange information and request processing information or media content.
APIs in broadcaster side makes it easy to access to them by service providers. Authorized service providers, including third-party, will access broadcast data through the APIs, creating their services, and offer them to the viewers. APIs in broadcaster side can be defined by broadcasters themselves; Common APIs at broadcasters for services providers will reduce complexity and required investment to provide actual services by service providers. The common APIs will be effective when many service providers offer broadcast-related services. Opportunity for the third-party service providers to offer their services on the platform is expected to expand the range of services as well as those by broadcasters. Hybridcast may not only encourage developing new business models for public, but also support to provide various services for minority.
APIs in a receiver have to be standardized to offer built-in functionalities of a receiver to all Hybridcast applications. Standardization of receiver APIs and a format of application signaling is carried out at IPTV FORUM JAPAN for the Hybridcast system.

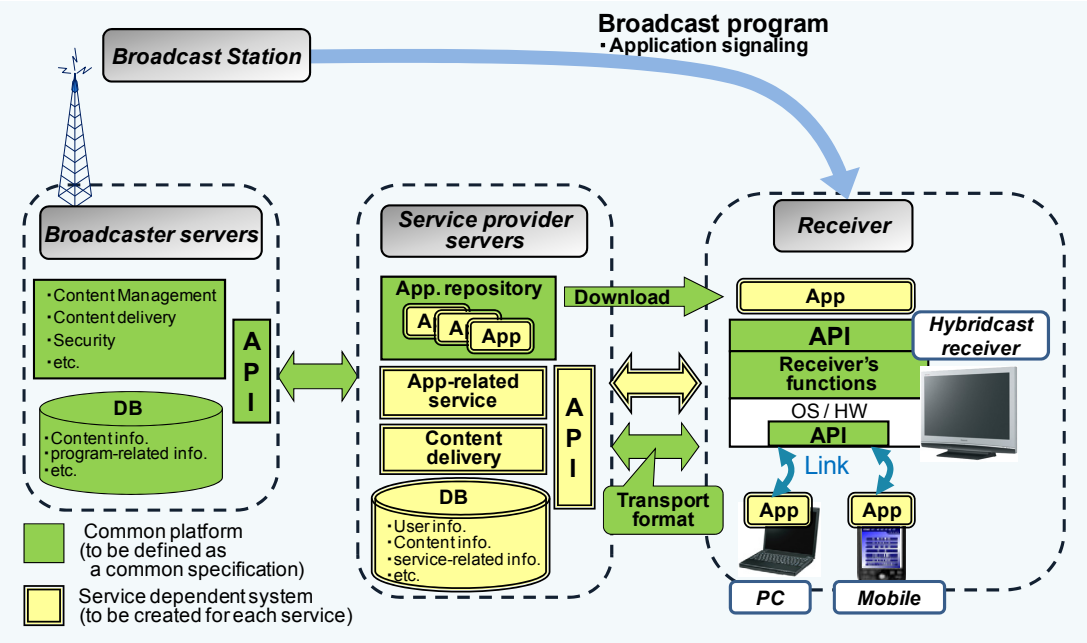


Figure 8. System overview

5. APPLICATION MODEL

Using Hybridcast, viewers can experience various types of services through Hybridcast applications. Considering services offered in Hybridcast, there are several types of applications which execute on the receiver.

Hybridcast applications are categorized as either broadcast-related or independent. For example, the application used for a quiz show should work together with the broadcast program. On the other hand, a weather forecast application is used independently from a broadcast program. As seen in these examples, there are several different application lifecycles, i.e. when to start and when to stop the application, depending on the application type. The lifecycle of a broadcast-related application should be controlled by a broadcast signal to associate with the program, while that of independent applications should be controlled by the viewer. A receiver provides an application launcher that creates an application catalogue for choice of independent applications. If a viewer wants to use such an application, all he or she needs to do is to start the launcher and choose the desired application.

6. HYBRIDCAST RECEIVER

The utilization and management of the application are directly linked to the user experience, because Hybridcast is application-oriented system. In this chapter, the basic architecture of the receiver, application management the mechanism and unique functions of Hybridcast are described.

6.1. Architecture

HTML5[8] is chosen as application environment in Hybridcast. There are three reasons to use HTML5.

(1) Easy implementation. A number of TVs and portable devices provide a Web browser. So, basic functionalities to handle HTML5 are naturally built-in these devices.
(2) High functionality and high efficiency. Using CSS3 (Cascading Style Sheets, level3) and Ajax (Asynchronous JavaScript + XML), applications with a rich interface and functionality can be easily developed.
(3) There are significant support to use HTML5 from many broadcasters and TV manufacturers.

By these reasons, HTML5 is adequate for Hybridcast on consumer electronics devices such as TV sets.

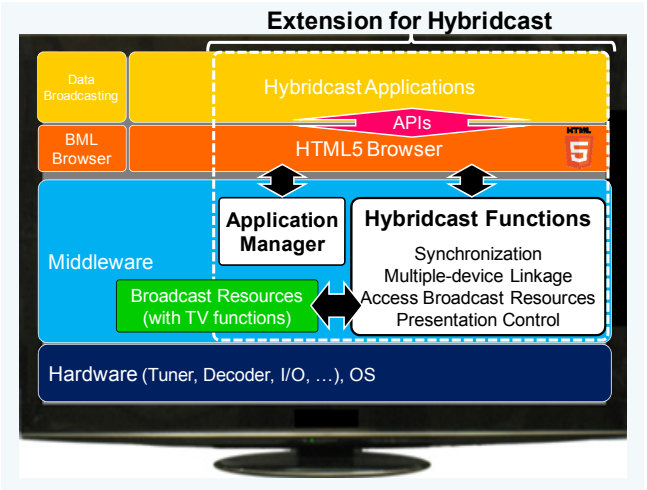


Figure 9. Architecture of a Hybridcast receiver

Figure 9 shows the layer architecture of the Hybridcast receiver. The receiver consists of four layers: the application, browser, middleware, and hardware. A typical Japanese DTV receiver supports data broadcasting services using BML (Broadcast Markup Language). A Hybridcast

receiver is equipped with a BML browser in parallel to a HTML5 browser. The middleware layer handles broadcast resources. Application manager is introduced for controlling applications on an HTML5 browser. Various functions are also implemented in middleware for broadcast-broadband combination and presentation control for emergency broadcasting. In case of emergency, the HTML5 browser is controlled by presentation control function to stop exhibition of the applications. Standardized APIs are prepared between an HTML5 browser and a Hybridcast application so that Hybridcast applications can use these functions easily.

6.2. Application Manager

The application manager controls the start and stop of applications. To allow program-related applications to work together accurately with a broadcasting program, the manager consistently monitors the control signal in broadcast channel. On the other hand, the manager also starts and stops independent applications at any time by viewer's manipulation. When starting an application, the manager instructs the HTML5 browser to load the application from the location represented in URL in the control signal. When stopping an application, the manager also instructs the HTML5 browser to terminate it.

6.3. Security Control

Hybridcast is an open platform where third-party service providers offer their services. As discussed in 3.2.5, open platforms have potential security risks such as application tampering and leakage of privacy[9]. Improper access to broadcast resource is another potential security risk, which may break copyright. It is important to make the system compliant to fixed rules to protect rights of each entity.
In Hybridcast receiver, many functional blocks provide their own security features. These features work together to satisfy the system requirements for security. According to execution context of the application, each block is designed to consult other relevant blocks for eligibility of access to security sensitive resources such as broadcast signal or non-volatile memory potentially containing user information.

6.4. Unique Functions of Hybridcast

6.4.1. Broadcast-broadband synchronization

A broadcast–broadband synchronization function allows an application to present timed content elements synchronously[10]. Figure 10 shows the typical mechanism of this function. This function synchronously presents timed components delivered from different channels, such as video, audio, and closed captioning. A broadcast signal can be considered as a stable and accurately timed component. In contrast, the packet-arrival time in a broadband network normally fluctuates. One solution to overcome the effect of this fluctuation is to provide a buffer in the receiver and transmit the broadband content earlier than the broadcast program. Content delivered through broadband shall provide time-stamp information so that the receiver can determine a synchronization point of the broadcast stream. By comparing time stamps of both streams, the receiver can adjust the time of presentation properly.
For a live broadcast program, larger buffers in a receiver are required since the additional data to be transmitted through broadband cannot be prepared ahead of time. The size of the buffer should be long enough to compensate for a lag in the broadband-delivered content. Actual buffer size will be determined by investigation of network latency.

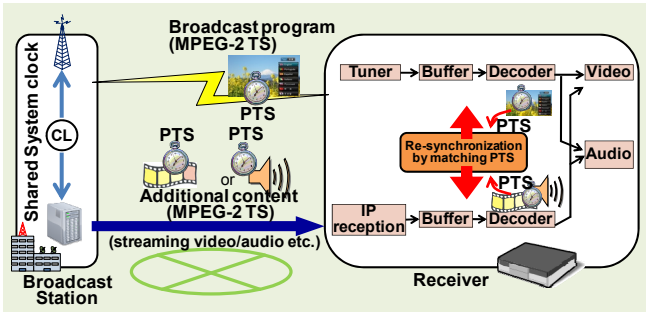


Figure 10. Synchronization mechanism

6.4.2. Multiple-device linkage

As the number of functionalities of broadcasting services and TV sets is increasing, remote controllers become more complicated. Instead of enhancing the functions or buttons on existing remote controller, use of alternative devices with user-friendly interface is one of the practical solutions for advanced services. Touch devices such as tablets or smartphones are typical devices for the usage and have emerged and spread rapidly. These devices have not only user-friendly interface, but also enhanced functionality of presentation of information, which contributes to the improvement of accessibility to all the viewers. Hybridcast provides multiple-device linkage function for this purpose.
Figure 11 shows how the applications on the secondary screen devices interact with the application on TV set through WiFi or a home network. The underlying mechanism of API's on a TV set and the secondary screen devices controls communication across the devices and allows communication between the applications. For example, when a viewer changes the channel using an application on secondary screen device, the application sends the message to the application on TV set. The application on the TV calls the API to tune to the new channel.

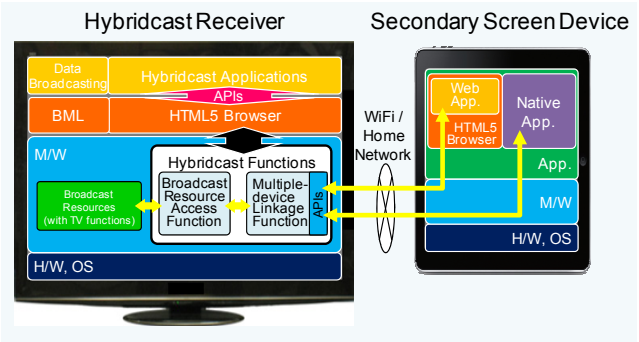


Figure 11. Multiple-device linkage

7. STANDARDIZATION STATUS

Hybridcast, as a service platform, is under standardization at the IPTV FORUM JAPAN. Broadcasters, TV manufactures, telecommunication carriers and service providers are joining a discussion about a specification of the platform toward the launch of Hybridcast in 2013. Among the elements shown in Figure 8, the standardization activity focuses on the receiver and application signaling because these areas are the most important parts for actual deployment of Hybridcast. In fact, the Hybridcast standard defines signaling format, middleware functionalities in a receiver, application format and additional APIs, like standard of HbbTV[11] which is an European hybrid broadcasting system.

Standardization road map of Hybridcast consists of two phases. The specification of the first phase will be finalized in the first half of 2013. The first phase focuses the functionalities for broadcast-related applications and services. Independent application capability and advanced feature such as frame-by-frame synchronization of Hybridcast are subject for standardization in the second phase. However, discussions of the both phases are carried out in parallel, which allows coherent overall system design. The first phase specification allows the service providers to use rich graphics presentation on a TV based on the HTML5 and other related web technology standards, use the secondary screen devices, and synchronize the application behaviors with broadcast program in trigger signal based manner. The standard does not specify actual implementation; the receiver manufacturers are responsible to design and implement the defined functionalities which allow the some of the advanced broadcast centric services, as described in chapter 2 except those in 2.2, 2.3, and 2.6 The second phase specification includes more advanced features such as independent application, an application control mechanism with an application launcher, a security mechanism for independent application, and frame-by-frame synchronization. The second phase specification will allow introduction of the full-fledge Hybridcast services. The finalization of the second phase specification is not yet determined but expected in the near future.

Toward the actual implementation and deployment of the phase one Hybridcast services, some broadcasters including NHK and some manufacturers are working together to establish operational guidelines and test signals of the receiver as well as creating the standard itself.

Knowledge and technologies of Hybridcast have been contributed to create Recommendation ITU-T J.205 and its subsequent series of Recommendations for reference architecture and system specification, Recommendations and Reports on hybrid broadcasting systems at ITU-R Study Group 6, and World Wide Web Consortium (W3C).

8. CONCLUSION

The Hybridcast is a service platform that uses a broadcast-broadband combination to enhance TV programs with variety of services. This paper describes overview of the system and examples of service. In some examples, it is also described that ability of Hybridcast reaches not only for public but also for minority and for family and friends.

NHK is now leading standardization of Hybridcast platform. It is expected that Hybridcast will be launched in 2013 and that it takes a role as a next-generation media for a sustainable society.

REFERENCES

[1] A. Baba, K. Matsumura, S. Mitsuya, M. Takechi, H. Fujisawa, H. Hamada, S. Sunasaki, and H. Katoh, "Seamless, Synchronous, and Supportive: Welcome to Hybridcast: An Advanced Hybrid Broadcast and Broadband System," *IEEE Consumer Electronics Magazine*, vol. 1, no. 2, pp. 43–52, 2012

[2] H. Kaneko, N. Hamaguchi, M. Doke, and S. Inoue. "Sign language animation using TVML," *Proc. of the 9th ACM SIGGRAPH Conference on VRCAI '10*, pp. 289–292, 2010

[3] N. Hamaguchi, H. Kaneko, M. Doke, S. Inoue, and I. Kumazawa, "Live Text-to-Video System Using Realtime Server-side Rendering," *Proc of IWAIT2011*, 2011

[4] M. Miyazaki, S. Nishimura, N. Hamaguchi, R. Sawai, and H. Fujisawa, "Social TV System for Public Broadcasting Services: Analysis of User Behavior in Large-Scale Field Trial -," *Proc of NAB2012*, pp. 154–160, 2012

[5] K. Shogen, Y. Ito, H. Hamazumi, and M. Taguchi, "Implementation of Emergency Warning Broadcasting System in the Asia Pacific Region," *ITU/ESCAP Disaster Communications Workshop*, 2006

[6] M. Hoshiba, O. Kamigaichi, M. Saito, S. Tsukada, and N. Hamada, "Earthquake Early Warning Starts Nationwide in Japan," *Eos, Transactions American Geophysical Union*, vol. 89, pp. 73–74, 2008

[7] K. Otsuki, H. Ohmata, A. Fujii, K. Majima, and T. Inoue, "A Method of Controlling Presentation for Applications in Hybridcast," *Proc. of IEEE ICCE2012*, pp. 323–324, 2012

[8] HTML5, http://www.w3.org/TR/html5/

[9] D. Barrera and P. Van Oorschot, "Secure Software Installation on Smartphones". *IEEE Security and Privacy*, vol.9, no.3, pp.42–48, 2011

[10] K. Matsumura, M. J. Evans, Y. Shishikui, and A. McParland, "Personalization of Broadcast Programs using Synchronized Internet Content," *Proc. of IEEE ICCE2010*, 4.1-5, 2010

[11] Hybrid Broadcast Broadband TV, ETSI TS 102 796 V1.1.1, 2010

STANDARD-BASED PUBLISH-SUBSCRIBE SERVICE ENABLER FOR SOCIAL APPLICATIONS AND AUGMENTED REALITY SERVICES

Oscar Rodríguez Rocha
Boris Moltchanov

Politecnico di Torino, oscar.rodriguezrocha@polito.it
Telecom Italia, boris.moltchanov@telecomitalia.it

ABSTRACT

A Publish/Subscribe mechanism based on the Open Mobile Alliance's (OMA) Next Generation Services Interface (NGSI) open standard, allows interfacing the information available from many publishers with heterogeneous customers. Pervasive devices (including mobile smartphones) publish a huge amount of real world information, which afterwards is accessed through web browsers and applications. The adoption of an open standard interface between information publishers and consumers allows to reduce the gap in the technologies used on both sides, therefore, include new actors into the services, increase the service offers and increment the world-wide and cross-domain usage of services based on the Publish/Subscribe paradigm. Major European Industrial Entities supported by the EU Research Program are deriving a cross-domain Future Internet open standard technology to be adopted and used in any application domain by any customer for any needs. The reference open standard chosen is OMA's NGSI. The open standard based technological binding created in the FI-WARE EU funded project and provided with an open reference implementation performed by Telecom Italia is demonstrated through examples of Augmented-Reality and social-impacting services that improve the quality of life for people (including those decease affected).

Keywords— publish-subscribe, reference standard, service enabler, augmented reality, e-learning, social impact, augmented reality

1. INTRODUCTION

The work presented in this paper, describes the adoption of a mature and world-recognized industrial open standard enclosed by an open technological framework (Chapter 2), which is implemented as an open reference Service Enabler that can be included and used in a wide range of applications and services involving cross-domain applications (Chapter 3). This reference implementation is based on the open specification of the technological framework defined as an open architecture and its source code is available as open-source to be used for further integration, adaptation or improvement.

Moreover, it is shown how this technological framework could be easily embedded into a Cloud Computing paradigm (Chapter 4). Finally, a reference implementation performed by a Telecom Italia is demonstrated with real prototyped services as an integrated common technology, potentially available for any type of service or application (Chapter 5).

The roadmap of this service enabler is given as a future work within the EU funded initiative (Chapter 6), with conclusions focusing on the relevance of this work for a modern future internet and information enabled Society (Chapter 7).

2. OMA NGSI STANDARD FOR PUBLISH/SUBSCRIBE SERVICE ENABLER

In 2010, the EU Commission founded the Future Internet Public Private Partner initiative [1][1]. It resulted into a number of Use Case Projects (UCPs) and a Future Internet Core Platform (FI-WARE project [2]) embracing all the Generic Enablers (GE) commonly used by any UCP. One of the most required GEs identified within FI-WARE, is the Publish/Subscribe GE. We have chosen the OMA's NGSI open standard [3] after a careful analysis of the existing industrial open standards and also by taking into consideration the already existing solutions provided by the FI-WARE partners. During the selection process, the priority has been given to the practically implemented, existing and used solutions, which are rather simple and able to work with heterogeneous devices in different application domains. OMA's NGSI open standard allows to retrieve any type of information, including context data and events (represented as schematically shown in the Figure 1), from their respective providers in different modes: on-request and subscription-based.

The information retrieval is performed with the aid of a broker node (as shown in the Figure 2).

- Additionally, the standard allows the creation of a federation of brokers to avoid bottleneck problems and to yield scalability and flexibility to the final solution.

[1]Thanks to EU PPP Initiative for funding.

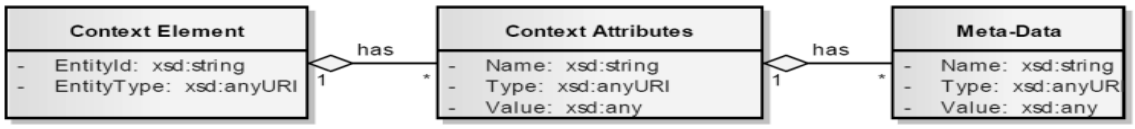


Figure 1. OMA's NGSI data representation model

Nevertheless, no specific technological binding has been created within the OMA, which is not aimed at this purpose, but perhaps to allow different bindings to be adopted by the interested industries.

Particularly, the FI-WARE project has decided to create a FI-WARE's NGSI RESTful binding based on XML standards and XSD schemas of the data resources parameters and interrogation methods [4]. This decision and the derived technology, enable to handle any type of data in a RESTful [5] manner by simply making requests to the data as RESTful resources. Consequently, the final technological solution will follow the Web service design model [6], which is widely used for the creation of many internet-based services and applications on the web.

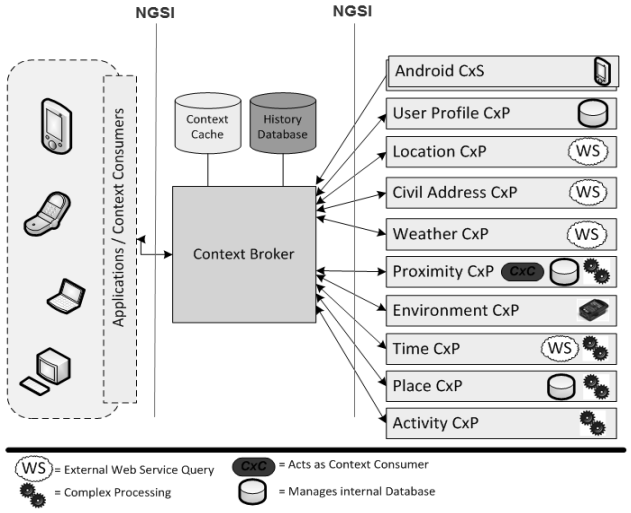


Figure 2. NGSI used with Broker

3. OPEN SPECIFICATION, ARCHITECTURE AND OPEN-SOURCE REFERENCE IMPLEMENTATION

We have derived within FI-WARE project the open specification of the Publish/Subscribe Generic Enabler (GE) [7] based on the Context Management platform (brought into the project by Telecom Italia) and on the OMA's NGSI architecture and API specifications. Two sets of interfaces (ContextML/CQL [8] and FI-WARE's NGSI [9]) together with the open FI-WARE's Publish/Subscribe GE architecture [10] are available on the FI-WARE web site for public use. Currently, we in Telecom Italia developing the reference GE implementation with both types of the interfaces (which also will be available as open-source): RESTlike ContextML/CQL and Restful FI-WARE's NGSI. However, anyone can create its own implementation taking the Open Specifications from the FI-WARE web site.

The main difference between the two interfaces is that the ContextML/CQL implementation (based on the simple exchange of XML-based documents through HTTP requests) has been already in use for long time by Telecom Italia for context management and context-aware applications, thus, it has been tested and is stable. On the other hand, the FI-WARE's NGSI interface is still under development (its first release is already available in the FI-WARE's project test-bed exposed via FI-WARE's GEs Catalogue [12]). It is possible to publish the contextual information of the data producers to the Publish/Subscribe broker, which additionally makes it available to be retrieved by context and data consumers (that might be any applications and/or services).

The openness of the overall architecture, the public available GEs' specifications and their reference implementations have a great value for the international and European society as well as they allows anybody to use a world-wide standard-based implementation of the Public/Subscribe GE, which is a part of the Future Internet Core Platform. Hence, it is expected that any data generated by context producers will be fully interoperable with context consumers (applications and services) supporting the same GE interfaces on a worldwide arena.

Moreover, through the plug-and-play architecture of the FI-WARE's NGSI that allows registering and removing the context and data producers, the Publish/Subscribe GE implementation is fully scalable and flexible. Any new data and context providers are able to autonomously connect and register into the Publish/Subscribe brokers, while the data and context consumers shall not care about where and which data or context information is available.

Finally, comprehensive and extendable query and subscription mechanisms for data and context are supported by both GE interfaces[2]. The subscription allows any data or context consuming entity to be notified by the Publish/Subscribe GE when the subscribed data or context is available.

4. PUBLISH/SUBSCRIBE ENABLER EMBEDDED IN A CLOUD PLATFORM

The Publish/Subscribe GE is based on the broker architecture shown in the Figure 2; and we have embedded it into the Cloud platform designed and developed by the 4CaaSt[3] project [11].

[2] FI-WARE's NGSI interface will support the query functionality in the Release 2 of the reference implementation. Release 1 available for the moment supports on-request data retrieval mode only.

[3] The research leading to these results has partially received funding from the 4CaaSt project [11] from the European Union's

The main goal of the 4CaaSt project is the creation of an automatic platform to allow the deployment and execution of a service or application where publishing and/or subscribing to data and context would be a part of the platform and thus could be used as any other service enabler (data storage, network capabilities, etc.) within the cloud. In order to embed the Publish/Subscribe and context-awareness as a service enabler, we have also integrated the broker designed in the Figure 2 into the 4CaaSt cloud provisioning, monitoring and charging subsystems. The Publish/Subscribe and context-awareness are now part of the 4CaaSt cloud platform and have been implemented as a native service enabler, that is always on and running accordingly to its own dedicated service blueprint deployed an controlled by the 4CaaSt platform.

5. AUGMENTED REALITY AND SOCIAL RELEVANT APPLICATIONS

In this section, two reference service prototypes developed by Telecom Italia embedding our Publish/Subscribe context broker are described in order to show the potential and advantages of the FI-WARE implementation and 4CaaSt integration of the OMA's NGSI as NGSI service enabler facilitating the context-aware service creation and execution. All these service prototypes and trials are integrating the Publish/Subscribe context broker making context information available during the service execution for showing in the customer's screen or process the service's logic.

5.1. Augmented Reality

In order to improve the daily life of mobile users (users with a mobile smartphone equipped with a camera), an Augmented Reality (AR) service has been prototyped. It brings real-time associated information directly into the objects that the user is watching through the screen of his mobile smartphone (an augmented view). Information is graphically shown automatically through layers, and it is based on the mobile user's location, preferences and social relationships.

Some examples of provided information are: nearby monuments' description, nearby buddies, content generated by other users and point-of-interest (POI) details.

A common example application of this service is the case of a mobile user looking for a place to have a lunch. The AR service can provide the user with useful information to find a place by considering her or his personal culinary tastes and preferences of the moment. By interacting with the application, the user can access additional information of each recommended place, including some details like: the address and phone number of the place, official information related to the point of interest (such as the website), reviews made by other users or experts, rankings, comments and pictures take. Additionally, it is possible to "check-in" a place or to see which of the friends of the user have already checked in, as well as the restaurant's menu, price list and ongoing offers and promotions.

The architectural components of this service are:

- Augmented Reality Content Server (ARCS), which contains geo-tagged information fetched through different data sources (such as points of interest descriptions or monuments information), mobile users' preferences and social information (such as friends and friends-of-friends). It also handles user-generated content with its related geographical position, for further use of the client application;
- Client Application, targeted for modern mobile devices. It renders the graphical augmented view by gathering data from the ARCS (based on the geographical position of the mobile user and her or his preferences) and attaching it graphically to the current reality view. It also provides content generation capabilities, it makes possible to upload contextualized user-generated content to the ARCS, which will handle it and manage it for further use. On Figure 3, a view from the client application is shown.

This service is mainly targeted to increase the usage of the data channel provided by the mobile operator and moreover could be either sold as an application pay-per-use or used to increase the appealing of the operator over the competitors.

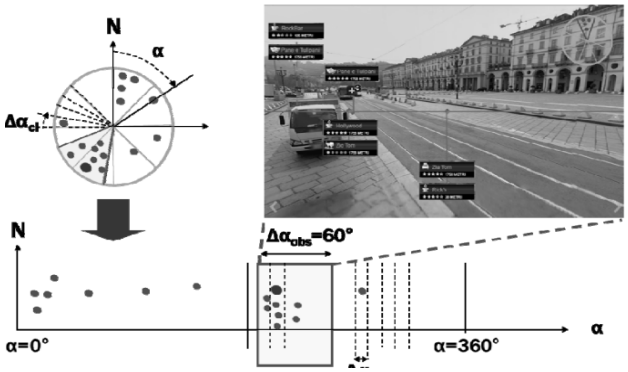


Figure 3 Augmented Reality view from client application

5.2. Social Reading

This service has been created with the aim of providing a more interesting and appealing experience of reading books; it focuses on creating a social community around the reader, in particular, those who read electronic books (eBooks) using modern mobile devices.

Such community is made up of service users (eBook readers). Groups of readers can be formed by existing relationships extracted from popular social networks and/or by common reading interests.

While reading an eBook through the system's client application, the user has the ability to make annotations or comments about a piece of text or paragraph, which can be also shared with the whole community, with specific groups or selected users (in compliance with the privacy options

Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 258862.

chosen by the user). Sharing with popular social networks like Twitter and Facebook is also possible.

A key feature provided by the service, is the automatic semantic enrichment, which adds associated information from Semantic Web data sources to the notes or comments generated by the users. This process is performed by the Semantic Annotator integrated into the system, which analyzes them as plain text entries in order to recognize relevant entities such as places, POIs, names and concepts, which are used to perform queries to different Semantic Web sources. The results of these queries are analyzed with a simple algorithm that determines the most probable entity; finally, interesting related content can be shown to the user graphically inside the client application to enrich her or his reading experience.

Non-semantic enrichment is also possible, as the graphical interface of the application also provides options to the reader to attach multimedia content such as images, videos and audio either from common web sources or from his existing files on the mobile device.

Strong emphasis has been made on the implementation of accessibility features to bring eBook-reading experiences to people with limited capacities, such as:

- Blindness: The system can aid blind or partially sighted persons by means of its integrated text to speech (TTS) engine. Since the availability of eBooks is greater than audio books, our system provides more reading possibilities. A set of different voices is also available;
- Vision problems: the font and size of the eBooks can be easily adjusted (increased or decreased) and the background color of each page can be selected to the one that fits better;
- Dyslexia: currently, some features being implemented to make the application Dyslexic-friendly.

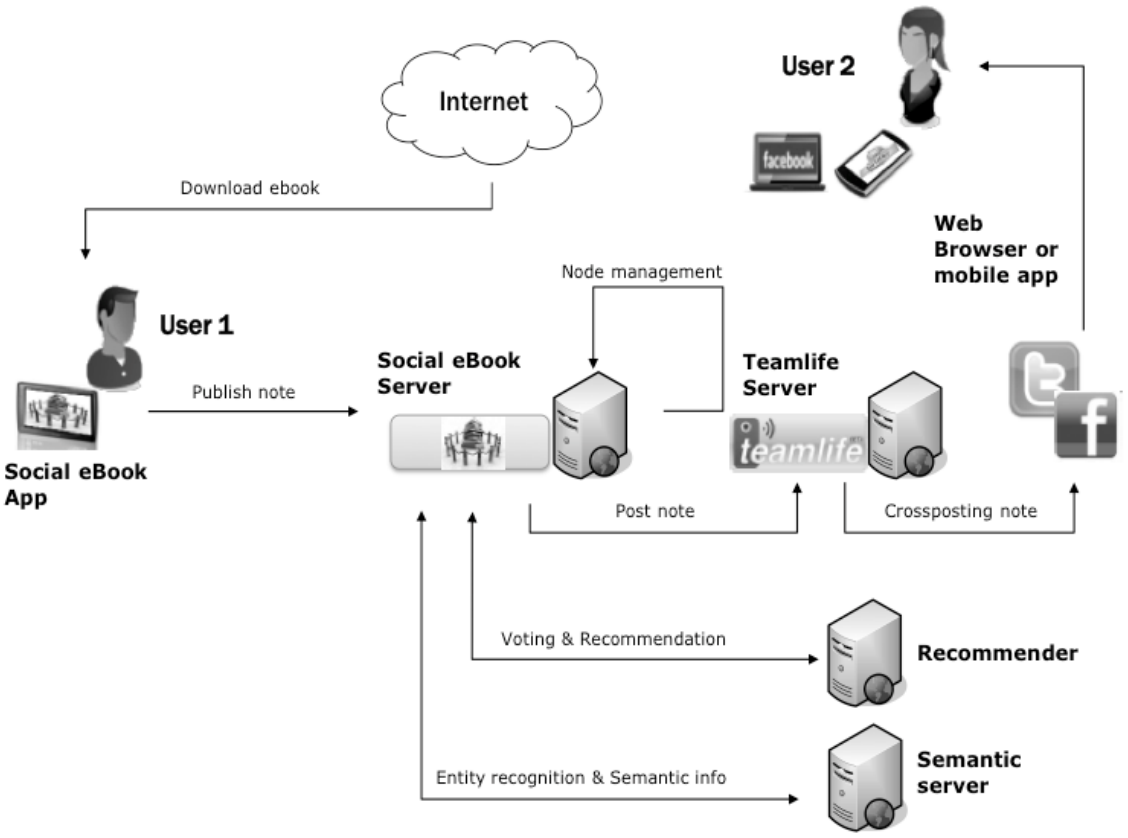


Figure 4 Social Reading platform architecture

This service can also be used in the education sector, especially to promote the digital education and to avoid the hard-printed books and improving the communication between teachers and students.

The high-level view of the service's architecture (shown in the figure 4) is composed by:

- Social eBook client application

It's a native client application with an interface for eBook reading (at the moment only ePub format is supported) and interacting with the social community; currently, most popular mobile devices are supported.

- Social eBook Server

It stores the eBooks available on the platform and at the same time, makes them available for download to the client applications. It additionally manages the notes and comments generated by the users.

- Teamlife Server

It is used to communicate of the Social eBook Server with the most popular social networks, thus, enabling the "share to" feature in notes and comments.

- Recommender

Interrogated by the Social eBook Server and considering the user's preferences, it returns a list of suggested eBook titles.

- Semantic Annotator

Performs the semantic enrichment process of user-generated notes and comments, as described previously.

At the moment, this service is being exploited by Telecom Italia mainly as a social effecting initiative aimed to support limited capacity people and improve the efficiency of the education process in schools. However economic introits are also possible through eBooks distribution supporting this social comments exchange features.

A set of trials of this service has been conducted with a selected group of schools in Italy. The main goal was to collect real usage data and feedbacks to improve the application and its current features.

6. FUTURE WORK

The Publish/Subscribe GE issued in its first release already supports the ContextML/CQL RESTlike full mode and FI-WARE NGSI RESTful limited mode communications. This GE will support the full mode of NGSI communications (which does not yet mean a full NGSI support) by the end of 2012 and full NGSI support mode is planned for 2013.

However, this implementation, architecture and API specification are already publicly available on the FI-WARE web site. Once the NGSI implementation will be accomplished, most attention will be given to the integration of the Publish/Subscribe GE with the FI Core Platform supporting systems such as monitoring, provisioning and charging capabilities in the FI-WARE Cloud integrating with the work performed within 4CaaSt project. A great attention will be also dedicated to the integration of this GE into the FI-WARE security framework. Finally, starting from the very first release of the Publish/Subscribe GE by the year 2012, integrations of this GE with other important and relevant GEs within FI-WARE will be performed (such as Big Data, Complex Event Processor, Multimedia Analysis, etc.). At the same time the work of usage and integration of the Publish/Subscribe GE with a number of UCP including OUTSMART, ENVIROFY and SmartCity, has been already started.

The technical binding created and implemented in Publish/Subscribe context broker exposing FI-WARE NGSI could be submitted to OMA for a further accomplishment and improvement of the OMA NGSI Enabler specification. Other standard de-facto, as PubSubHubbub[4] will be considered when we will start to create a federation model of the Publish/Subscribe context brokers handling different type of the context information distinguished per owner, per source, application domain and per functional principle. Moreover any other standardized and best-practice solutions such as XMLL and SIP presences, OMA Location, and others will be evaluated and, if requested by service scenarios or use cases (UCPs), will be integrated with the broker, as context providers or sources, supporting the NGSI communication. We're already working currently on the Semantic enhancement of the Publish/Subscribe context broker following the OWL standard, SPARQL communication pattern and RDF data representation model.

7. CONCLUSIONS

In this paper we have described a real-life big effort done by industrial entities to bring their assets for the common usage of the worldwide open community.

A Publish/Subscribe GE has been presented as an example of an solution openly defined and based on an open standard demonstrated with a couple of service prototypes created and provided by Telecom Italia. The services are impacting in both the user appealing and the social usefulness perspectives. This effort has been made possible through the support of the European Research Program funding Future Internet Public Private Partnership Program, including the FI-WARE project in which this activity has been performed. Moreover, the results of the EU funded project, 4CaaSt, have been extensively used for the integration of this GE into the cloud technologies and for making it available as a native service for a further usage or service composition and execution.

Although for the moment no performance measurement experiments have been performed in this phase targeting creation of the prototype complex service creation enabling and executing system, current trials didn't show any performance bottlenecks or latency with limited number of customers and moderate platform usage.

However this work is not a final job and there are still a lot of development, implementation and integration activities regarding this GE that will be carried out by the end of 2012 and continuously during year 2013.

[4] https://code.google.com/p/pubsubhubbub/

REFERENCES

[1] FI-PPP web-site: http://www.fi-ppp.eu/.

[2] FI-WARE web-site http://www.fi-ware.eu/.

[3] OMA NGSI Open Specification web-site: http://www.openmobilealliance.org/Technical/release_program/ngsi_v1_0.aspx.

[4] FI-WARE's NGSI XSD RESTful binding: http://forge.fi-ware.eu/plugins/mediawiki/wiki/fiware/index.php/FI-WARE_NGSI_Open_RESTful_API_Specification_%28PRELIMINARY%29.

[5] REST specification de-facto: http://www.ics.uci.edu/~fielding/pubs/dissertation/rest_arch_style.htm.

[6] Web service design model description: http://en.wikipedia.org/wiki/Web_service.

[7] Publish/subscribe broker in FI-WARE http://forge.fi-ware.eu/plugins/mediawiki/wiki/fiware/index.php/Data/Context_Management#Publish.2FSubscribe_Broker.

[8] Moltchanov B., "Context Representation Formalism and Its Integration into Context as a Service in Clouds", ITU Kaleidoscope 2013, Cape Town, South Africa, December 2011.

[9] FI-WARE's NGSI Open RESTful API Specification http://forge.fi-ware.eu/plugins/mediawiki/wiki/fiware/index.php/FI-WARE_NGSI_Open_RESTful_API_Specification_%28PRELIMINARY%29.

[10] Publish/subscribe GE Architecture and Open Specifications web-site http://forge.fi-ware.eu/plugins/mediawiki/wiki/fiware/index.php/FIWARE.ArchitectureDescription.Data.PubSub.[11] 4CaaSt Project web-site http://4caast.morfeo-project.org/.

[12] FI-WARE GE Catalogue, Test-bed web site http://catalogue.fi-ware.eu/enablers.

QOXPHERE: A NEW QOS FRAMEWORK FOR FUTURE NETWORKS

Eva Ibarrola1, Eduardo Saiz1, Jin Xiao 2, Luis Zabala1, Leire Cristobo1

1 Faculty of Engineering in Bilbao, University of the Basque Country, Spain
2 Division of ITCE, Pohang University of Science and Technology, Korea

ABSTRACT

The telecommunications sector has experienced significant changes over the past few years. The advent and rise of new applications and services, together with a competitive market, has led to a complex scenario in which quality of service (QoS) plays a major role. Under this condition, novel QoS regulation and standardization initiatives are required. During the last few years new terms and concepts, such as Quality of Experience (QoE) or QoS Perceived (QoP), have been included in the updated and new QoS-related standards as to better integrate the user's point of view, as opposed to only network performance parameters. The influence of the user's satisfaction on the Quality of Business (QoBiz) has also been given increased attention in the regulation and standardization bodies recently. The result is a loose collection of metrics and models that are not standardized and do not integrate all aspects of quality. Such integration is necessary to assure the successful development of this sector. This paper presents a new and integrated QoS model (QoXphere) that is spherical, adaptive and multi-layered.

Keywords— *QoS, QoP, QoE, Quality of Business, Satisfaction, QoX*

1. INTRODUCTION

The drastic increase of networked applications and services together with the deregulation of the telecommunications market has led to a complex and competitive business scenario. End users are demanding higher levels of Quality of Service (QoS) thanks to the rising interests in real-time and multimedia applications. The providers therefore need to adopt and implement new QoS management policies in order to ensure user satisfaction and avoid subscriber churns. User satisfaction is then a pivotal metric to measure and manage.

The ITU-T E.800 Recommendation [1] defines quality of service as "*the collective effect of service performance which determines the degree of satisfaction of a user of the service*". Based on this new user-centric QoS definition, new concepts and metrics have been recently defined.

One of the QoS-related terms most widely mentioned lately is Quality of Experience (QoE). The ITU-T Recommendation ITU-T P.10 Amd.2 [2] defines Quality of Experience (QoE) as "*the overall acceptability of an application or service, as perceived subjectively by the end-user*" and remarks that the overall acceptability may be influenced by user expectations and context, and that quality of experience includes the complete end-to-end system effects (client, terminal, network, services infrastructure, etc.). Based on this definition, managing and measuring QoE becomes a complex task and, therefore, the ITU-T G.1011 Recommendation [3] provides a reference guide to QoE assessment methodologies.

As referred in the ITU-T G.1000 Recommendation [4], QoS should also embrace other user-centric terms and concepts such as QoS Perceived by user (QoP) [4] or QoS Required by users (QoSR) [5]. These concepts are closely related to the user satisfaction with the service. Quality of Business (QoBiz) [6-8] is another example of QoS-related concepts that have seen general use both in the scientific and the standardization realms.

Despite the fact that these contributions have helped steering the global management approach towards user centric QoS space, it is difficult to navigate the relations among all the different concepts and terms mentioned above.

Furthermore, the standardization of an integrated QoS model is especially challenging when considering the specific scenario of an Internet service provider, due to the evolving service landscape brought on by constant technological improvements and the inherent heterogeneity of the diverse services offerings.

In this paper, we present a novel QoS model (QoXphere) defined based on a new user-centric and business-oriented QoS dimension. The new model proposes a new spherical, adaptive and multi-layer model that takes into account most of the different concepts and aspects defined in the current QoS regulation and standards. The aim of QoXphere is to contribute to the definition of a new standardized QoS model that will help to advance the ongoing trend of user-centric service management designs and to complement the existing standardization efforts by providing a unified QoS management terms and concepts.

The remainder of the paper is organized as follows: Section 2 summarizes related work both in standardization bodies and scientific area. Section 3 describes the new QoXphere model. In section 4, the dynamical behavior of the QoXphere model is presented. Section 5 describes the platform developed to validate the QoXphere model and, finally, section 6 contains some conclusions and final remarks.

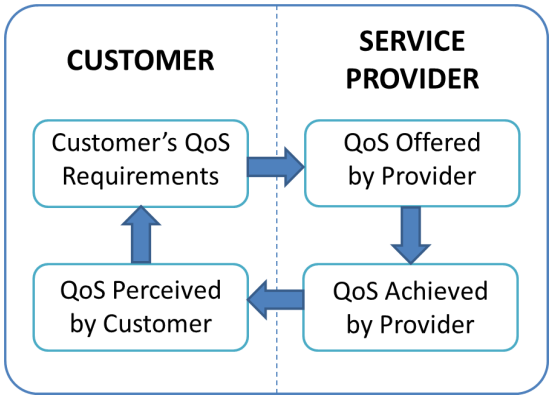


Figure 1. The four QoS viewpoints of G.1000 Rec.

2. BACKGROUND

2.1. QoS and Standards

In recent years, the meaning of the concept of QoS in the standardization environment has undergone profound changes. From Network Performance (NP) parameters to those related to the user's experience and perception together with cost, a new approach to quality of service has been developed, strengthening the idea that the isolated study of each term cannot lead to an effective QoS management. In the new telecommunication environment, a more complex analysis is needed to cover the interests of all the different actors involved in the service provision: providers, users and regulators. Therefore, the standardization and regulation bodies have realized that there is a need of new regulatory frameworks to encompass the evolution of the QoS in order to integrate the user's point of view.

In this sense, ITU-T has already updated some of the most important recommendations related to QoS, such as the E.800 Recommendation: "Definitions of terms related to Quality of Service" [1], comprising both the user's and the provider's point of view. In the new updated E.800 recommendation, the QoS framework established in G.1000 recommendation [4] is adopted. This framework considers the four different points of view of QoS as described in figure 1. This recommendation also states that "*for any framework of QoS to be truly useful and practical enough to be used across the industry, it must be meaningful from these four viewpoints*". This statement implies that QoS assessment will necessarily involve a combination of many objective and subjective aspects that, properly linked, may be used as an indicator of user's perception.

Still this framework is not specific enough to be developed in real Internet network scenarios. In addition, other QoS aspects that ITU-T advises to analyze [5], such as user satisfaction (figure 2) or the Quality of Business (QoBiz) [6] are not considered within the G.1000 framework.

It is also remarkable that no general QoS terminology has been adopted by all the standardization bodies [11] and different terms to refer to the same concept can be found in different standards.

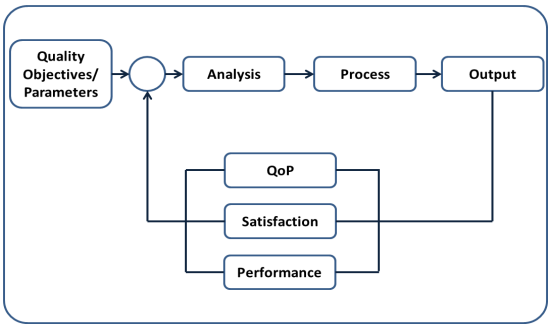


Figure 2. E.802: Process of managing a quality policy

Therefore, there is a need for a global and general QoS framework to unify criteria in terms of concepts and terminology and to cover all the different aspects that should be taken into account for a practical QoS management in the Internet services environment.

2.2. QoS and Research

Considering the new scope of quality of service, several scientific researchers [11, 12] agree on the need for a stratification of the different aspects of QoS in three different layers: intrinsic QoS, perceived QoS and assessed QoS. Figure 3 represents this stratification and the relation of this vision to the approaches taken by different standardization bodies.

In the specific area of IP networks, one of the pioneering researches has been carried out by Aad Van Moorsel [13]. Moorsel proposes that to effectively manage QoS in Internet services, three different aspects should be taken into account: QoBiz, QoE and a particular definition of QoS which includes the quality of service of the system and the applications. Along with this, Moorsel defines a relationship framework for the management of the three aspects (figure 4).

This very same terminology is used by Kilkki [14]. The QoS model defined by Kilkki shows clear similarities with Moorsel's, but also includes business objectives such as ARPU and churn, and divides QoE into two different concepts, differentiating between user QoE and client QoE.

Finally, regarding the socio-economic aspects involved in QoBiz, the work of Peter Reichl [15] offers an interesting discussion of different charging schemes and their implication in the transition from the original QoS to the new user centered QoS dimension regarding user satisfaction. Also, Murali Muniyandi [8] analyzes an optimal strategy aimed at satisfying user's expected service levels, while at the same time improving revenue and outputs of the service providers. The three-stage model proposed by Murali prioritizes and schedules service requests so as to maintain the levels of QoE and QoBiz.

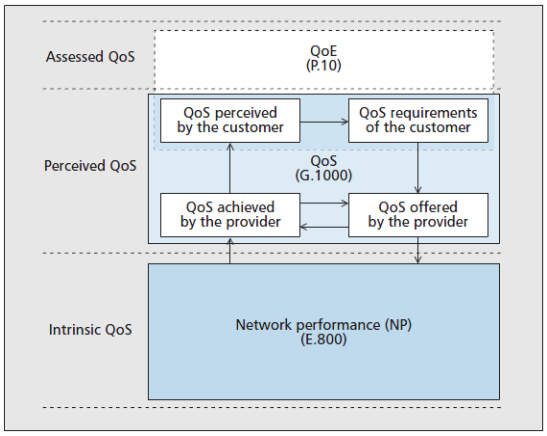


Figure 3. The general QoS model, and ITU/ETSI and IETF approaches (Gozdecki)

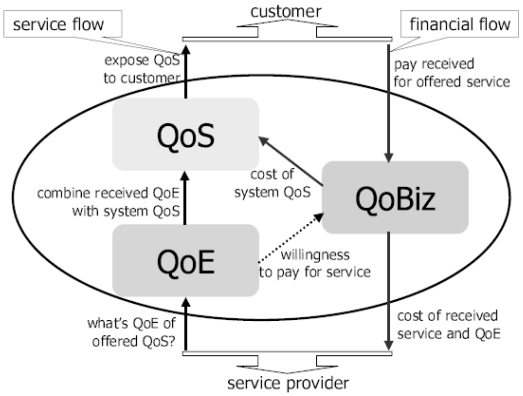


Figure 4. QoS, QoE & QoBiz relationship framework (Van Moorsel)

2.3. QoX: A new concept

Despite all the research and standardization work done in the QoS area, there was still a need for a new concept that would embrace the different QoS-related aspects: QoBiz, QoE, QoS (as contemplated in the context of ITU-T G.1000 Recommendation) and intrinsic QoS (network and applications related).

Recent work has been studying this issue [15, 16] and a new term has been coined to designate the new QoS dimension: QoX. Nevertheless, none of these studies have faced the difficult task of defining a new QoS framework to link all these QoS aspects (QoX) in the search for user satisfaction and fidelity. We have taken up this challenge and a new QoS framework has been developed.

The new framework is backed up by a multi-layer spherical architecture (QoXphere) that is based on a QoS management model for the analysis and evaluation of the QoX as a whole. The starting point for the definition of the proposed QoS framework is the QoS management model presented by Ibarrola in [17, 18].

This model is built on the context of the ITU-T G.1000 framework and, therefore, ensures its compliance with standards. In fact the implementation methodology of this model is based on the ITU-T E.802 recommendation [5] and, as a result, all satisfaction, QoP and performance aspects are considered. The new QoX framework also takes into account the quality of business (QoBiz), as referred in the eTOM framework in Rec. M.3050.1 [7].

3. QOXPHERE: A NEW FRAMEWORK FOR QOS

QoXphere defines a spherical architecture (figure 5) for the global analysis and evaluation of all the different aspects of the QoS (QoX) organized in four different layers:

- Intrinsic QoS
- Perceived QoS
- Assessed QoS
- Business QoS

Table 1 summarizes all these concepts and related terminology.

Table 1. QoXphere layers and related concepts and terminology

QoS Layer	Feature	Acr.	Definition
Intrinsic QoS			**Reflects the service features stemming from the technical aspects**
	Network Performance	NP	Ability of a network to provide the functions related to communications between users
	Grade of Service	GoS	Categorization of services with respect to requirements that can be verified through NP
	Class of Service	CoS	Any of the network-oriented designations that can distinguish between various services
	Quality of Resilience	QoR	Describes network survivability. It mainly concerns recovery time and availability
Perceived QoS			**Reflects the customer's experience of using a particular service**
	QoS Required	QoSR	A statement of QoS requirements by a customer/user or segment/s of customer/user
	QoS Offered	QoSO	A statement of the level of quality planned and therefore offered to the customer
	QoS Delivered	QoSD	A statement of the level of QoS achieved or delivered to the customer
	QoS Perceived	QoP	A statement expressing the level of quality that customers believe they have perceived
Assessed QoS			**Reflects the user's satisfaction and decision of remaining with the provider or not**
	QoS Experienced	QoE	The overall acceptability of an application or service, as perceived by the user
	Satisfaction	SAT	Global customer's satisfaction with the service
	User's selection	UPS	Provider Selection made by the user/customer
	Expectation	EXP	User expectations concerning the quality of the service
Business QoS			**Reflects the provider business stage (pertains with cost, revenue, investment…)**
	Attrition rate	Churn	Measure of customers moving out of a collective over a specific period of time
	Revenue	ARPU	Average Revenue Per User (services provided/the number of users buying the services)
	QoS of Business	QoBiz	QoS metric that quantifies the business return of a service provider (profit/revenue)
	Advertisement	ADV	Providers media and publicity policy

The aforementioned four layers are distributed in a macroscopic approach to the sphere and each of them contains the specific QoS aspects that need to be considered in each layer (figure 6). Moreover, each layer has an immediate effect onto the adjacent, through the interrelations between different QoS aspects of different layers. These interrelations, together with the intra-relations between the QoS aspects considered in each of the layers, provide the interconnection of all the aspects to be analyzed.

Figure 5. QoXphere model

From the lower level to the top, this section will try to bring light to the microscopic architecture of each layer.

3.1. Intrinsic QoS

The lower layer of QoXphere architecture analyzes the objective QoS parameters evaluation at the Network Performance level, as defined in ITU-T Y.1540 and Y.1541 Rec. [19, 20]. This analysis has a direct impact on the Quality of Service Delivered (QoSD) by provider, as defined in the G.1000 recommendation and treated in layer 2 of the QoXphere.

As mentioned in [11], Grade of Service (GoS) can be used to categorize services with respect to high-level requirements. In future converged networks (e.g. GMPLS-based optical networks) GoS provisioning is a real challenge since it is not easy to determine the grade of service to support a certain level of quality of service. Based on a given set of QoS requirements, a GoS is defined on an end-to-end basis as defined in the QoS offered (QoSO) and the SLA. For each major service category a Class of Service (CoS) is established.

The Class of Service (CoS), is defined in ITU-T E.417 [21] as "*any of the network-oriented designations or features that can distinguish between various services or application-layer users, of lower-layer telecommunications capabilities for the purpose of more effectively accommodating the network performance needs of specific services*". Any variation in the Quality of Service Offered (QoSO) by the provider implies the redefinition of the defined CoS and, therefore, the capability of the network should be adapted to each new situation, as does also the survivability against failures.

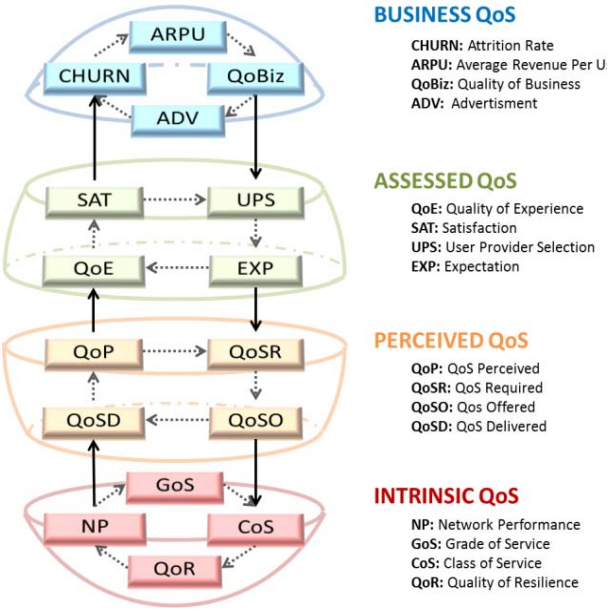


Figure 6. QoXphere: Layer specification

This is controlled through the Quality of Resilience (QoR), which may induce variations in the objective parameters being measured by NP.

3.2. Perceived QoS

The second layer of the QoXphere is based on the four viewpoints of ITU-T G.1000 recommendation. This layer can be considered the core of the sphere, since the QoS model is defined and developed in accordance with it.

According to the specifications of the methodological framework for the QoS model, the analysis must start with the definition of the user's requirements (QoSR).This information is provided through the definition of a set of Key Quality Indicators (KQI) useful to establish the QoSR, the QoS Offered by the provider (QoSO) and the final QoS Delivered to the user (QoSD).

The Key Performance Indicators (KPI) related to each of KQI will be defined and used to measure and determine the quality of the provider's network and services operation through the measurement platforms [22]. The KPIs offer information about a monitored resource and KQIs are used to estimate the end-to-end QoS as perceived by the user. Therefore, the set of the KPIs that contributes to each KQI must be defined [23, 24]. The KQI/KPI identification and the definition of their relationships is one of the major goals of the QoS model. In addition, the Quality of Service Perceived by user (QoP) is a key aspect to be analyzed in the QoS model, since its value provides the required feedback to the upper layer. Once the KQIs and KPIs have

been determined, they must be monitored. The results of the measures lead to indicators value analysis.

The QoP is estimated taking into account the gap between the user's perception and expectation. These data are inferred from the user's experience information provided to this layer by means of specific in-depth surveys (layer 3).

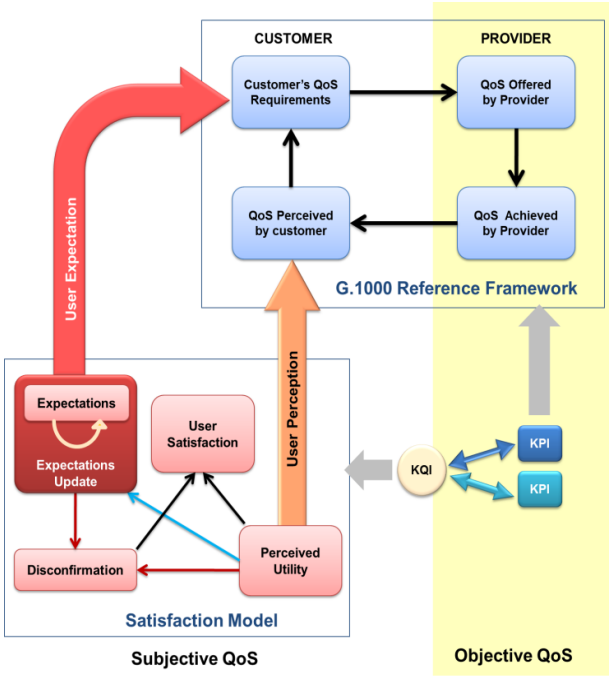


Figure 7. QoXphere Layer 3: Satisfaction Model

3.3. Assessed QoS

The performance of a service has a positive or negative effect on user satisfaction depending on the user's experience, which is based, in its turn, on QoP. In our model, the Quality of Experience (QoE) is fed by the QoP estimated in layer 2, and has a direct impact on satisfaction (SAT). The user satisfaction with the service has direct influence on the loyalty of the user and, thus, on the jeopardy of contracting other provider (Layer 3: UPS, Layer 4: Churn).

If the provider wants to ensure the loyalty of the customer, securing the user's satisfaction is critical. For this purpose, a satisfaction model is proposed.

Based on the research by Anderson and Sullivan [25] and Xiao's CSAT model [26], the proposed model has been adapted to suit the QoXphere multidimensional model (figure 7), bearing in mind the G.1000 reference framework and the QoS model presented in layer 2.

3.4. Business QoS

The higher layer of QoXphere model is oriented to guarantee the provider's profitability. Thanks to the feedback from the other three layers, the actions that service providers should take in this layer are determined.

Service providers should focus on achieving an effective use of the resources to maximize benefits (QoBiz), taking into account the Average Revenue Per User (ARPU) as well as the user fidelity in terms of the user's intention of swapping to a different provider (churn).

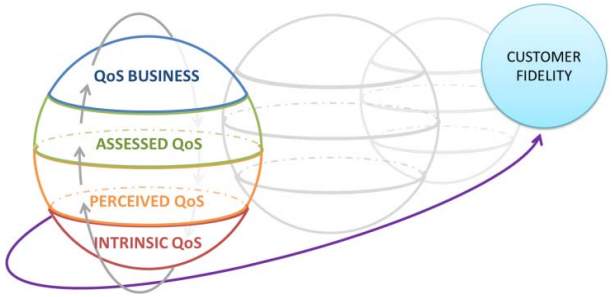


Figure 8. QoXphere: Rotation and Translation

In order to guarantee the user's loyalty, the user requirements should be fulfilled not only by means of objective QoS parameters (NP) and restoration mechanisms (QoR) but also by looking into many other subjective and contextual aspects that may have significant effects on user satisfaction (e.g.: advertisement may influence provider's reputation).

Subjective elements such as user's requirements and expectations do have influence on the perceived quality (QoP); and contextual aspects, such as the provider's public image, advertising, market dynamism, fees and cost, also affect the final quality experienced by the user (QoE).

4. QOXPHERE ROTATION AND TRASLATION

The most innovative part of the QoXphere model is that, besides the layer specification, the whole sphere defines similar movements to those described by a planet: QoXphere rotation and translation derive their meaning from the convergence of all the different aspects of QoS.

The rotational movement allows obtaining feedback from the different layers. For example, a certain objective network performance parameter can have an influence on the end user's perception (QoP). Being constantly exposed to perception as well as to other factors, such as satisfaction and expectation, leads to a particular quality of experience by the user, which, eventually can affect the loyalty of the user.

The providers must ensure their benefits offering a fully satisfactory service and, therefore, they should keep track of the information that lower layers provide, and react with new advertising campaigns, network improvements, and new services deployments.

All these actions lead to new reactions by the users depending on the layer they have been introduced in. In the case of network improvement, this leads to the reevaluation of objective parameters. The new QoSD, along with the other campaigns held by the provider, modify the QoS perceived by the user, thus accomplishing a full rotation lap.

The translational movement takes place in the constant iteration of the rotation movement with the purpose of leading the four layers of quality towards convergence in order to guarantee both provider profitability and user satisfaction (figure 8).

5. QOXPHERE PLATFORM

Since one of the major aims of this QoS model is to contribute to new QoS standards, a novel QoXphere system platform has been developed to validate the proposed model. The implementation of this platform considers the following infrastructures (Figure 9):

5.1. QoSmeter

QoSmeter is a neutral QoS-measurement infrastructure [27] meant to measure objective network parameters through a wide variety of tests that allow users to determine the degree of compliance of their SLA. The information gathered is also provided to the users of the system: customers, providers and regulators.

5.2. LabQoS

LabQoS [28] is a further development of the QoSmeter subsystem. It has been developed for testing and simulating experimental scenarios. QoSmeter measures different QoS parameters directly linked to the Internet connection and LabQoS works on both controlled and simulated scenarios providing specific results for different network configurations.

Both LabQoS and QoSmeter focus on the Intrinsic QoS layer, and mainly on the network performance (NP) monitoring where most of the tests are held.

5.3. ObavaQoS

The OBservatory for the Analysis and VAlidation of QoS (ObavaQoS) has been developed for the analysis of the Perceived QoS layer aspects of QoXphere. This subsystem has been defined to automatically determine the KQI and KPI indicators, and estimate, through the QoS model, the QoS perceived by users in terms of the G.1000 Rec. framework.

5.4. ENQoS

ENQoS is a general purpose, on-demand survey management and configuration subsystem that, in the context of QoXphere, centers on the analysis of the QoE and the Assessed QoS layer. This platform permits the preparation of specific surveys that are presented to users on a regular basis to evaluate the evolution of the experienced QoS.

Finally, to conclude with the description of the QoXphere, it must be said that this platform is still under development. Once the integration of the model is completed, the QoXphere quality model will be ready for its validation and consideration in the standardization process of regulation and standardization bodies.

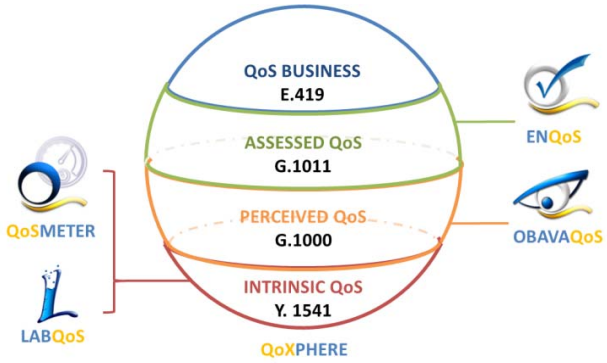


Figure 9. NQaS projects accommodation into QoXphere model

6. CONCLUSIONS

In this paper we have presented a global QoS model (QoXphere), defined on the basis of the ITU-T G.1000 and E.802 recommendations, with the general purpose of unifying criteria and embracing all the general QoS aspects as discussed in most of the standardization forums in recent years.

The QoXphere model is based on a novel multi-layer model that enables the analysis of all the QoS features (QoX). The QoXphere architecture consists of four different layers defined to analyze the intrinsic QoS, the perceived QoS, the assessed QoS and the business QoS. The user's perception and satisfaction is modeled through the dynamic and adaptive behavior of the sphere's rotational and translational movements towards convergence, ensuring user loyalty and, consequently, provider's profitability.

The QoXphere model is still in an early stage of development. Once the QoXphere platform becomes ready for validation, its reliability and suitability will be proven in a wide variety of scenarios for different services and requirements, always in compliance with reference standards and never forgetting the aim of providing a new dimension to quality of service in future networks.

REFERENCES

[1] ITU-T, "E.800: Definitions of terms related to Quality of Service", 2008.

[2] ITU-T, "P.10/G.100 (2006) Amendment 2 (07/08): New definitions for inclusion in Recommendation ITU-T P.10/G.100", 2008.

[3] ITU-T, "G.1011: Reference guide to quality of experience assessment methodologies", 2011.

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

[5] ITU-T, “E.802: Framework and methodologies for the determination and application of QoS parameters”, 2007.

[6] ITU-T, “E.419: Business oriented key performance indicators for management of networks and services”, 2006.

[7] ITU-T, “M.3050.1: Enhanced Telecom Operations Map (eTOM) – The business process framework”, 2007.

[8] M. Muniyandi, S. Krishnaswamy, and B. Srinivasan, "Improving the quality of business and quality of experience in Web services through prioritising and scheduling," International Journal of Business Process Integration and Management, vol. 2, pp. 156-171, 2007.

[9] ETSI, "EG 202 765-4: QoS and network performance metrics and measurement methods; Part 4: Indicators for supervision of Multiplay services," 2009.

[10] ETSI, "TS 102 250-1: Speech Processing, Transmission and Quality Aspects (STQ);QoS aspects for popular services in GSM and 3G networks; Part 1: Identification of Quality of Service criteria," 2011.

[11] J. Gozdecki, A. Jajszczyk, and R. Stankiewicz, "Quality of service terminology in IP networks," IEEE Communications Magazine, vol. 41, pp. 153-159, 2003.

[12] W. C. Hardy, QoS: Measurement and Evaluation of Telecommunications Quality of Service: John Wiley Sons, Inc., 2001.

[13] A. V. Moorsel, "Metrics for the Internet Age: Quality of Experience and Quality of Business," in 5th Performability Workshop, Erlangen, Germany, 2001.

[14] K. Kilkki, "Quality of Experience in Communications Ecosystem," Journal of Universal Computer Science, vol. 14, pp. 615-624, 2008.

[15] P. Reichl, "From charging for Quality of Service to charging for Quality of Experience," Annals of Telecommunications, vol. 65, pp. 189-199, 2010.

[16] R. Stankiewicz, P. Cholda, and A. Jajszczyk, "QoX: What is it really?," Communications Magazine, IEEE, vol. 49, pp. 148-158, 2010.

[17] E. Ibarrola, F. Liberal, A. Ferro, and J. Xiao, "Quality of Service management for ISPs: a model and implementation methodology based on the ITU-T Recommendation E.802 framework," IEEE Communications Magazine, vol. 48, pp. 146-153, 2010.

[18] E. Ibarrola, X. Jin, F. Liberal, and A. Ferro, "Internet QoS regulation in future networks: a user-centric approach," IEEE Communications Magazine, vol. 49, pp. 148-155, 2011.

[19] ITU-T, “Y.1541: Network performance objectives for IP-based services”, 2006.

[20] ITU-T, “Y.1540: Internet protocol data communication service – IP packet transfer and availability performance parameters”, 2007.

[21] ITU-T, “E.417: Framework for the network management of IP-based networks”, 2005.

[22] ITU-T, "G.1030: Estimating end-to-end performance in IP networks for data applications," 2005.

[23] TMF, “SLA Handbook Solution Suite, V.2.0”, 2005.

[24] ETSI, “EG 202 009-1: Quality of telecom services; Part1: Methodology for identification of parameters relevant to the Users”, 2007.

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

[26] J. Xiao and R. Boutaba, "Assessing network service profitability: modeling from market science perspective," Networking, IEEE/ACM Transactions on, vol. 15, pp. 1307-1320, 2007.

[27] R. Partearroyo, J. L. Jodra, J. O. Fajardo, A. Ferro Vazquez, and B. Blanco, "QoSmeter: Generic quality of service measurement infrastructure," in IFIP Networking 2006, workshop 'Towards the QoS Internet' (To-QoS'2006), Coimbra, Portugal, 2006.

[28] L. Zabala, A. Ferro, C. Perfecto del Amo, E. Ibarrola and J. L. Jodra, "LabQoS: A platform for network test environments," in ITU-T Kaleidoscope 2011. The fully networked human? − Innovations for future networks and services, Cape Town, South Africa, 2011.

TELEBIOMETRIC INFORMATION SECURITY AND SAFETY MANAGEMENT

Phillip H. Griffin, CISM

Booz | Allen | Hamilton, Linthicum, Maryland USA

ABSTRACT

Organizations that rely on human-oriented technologies such as telebiometrics should protect and manage the safety and security of their physical and information assets. Data that documents the safe and secure operation of telebiometric system devices should be collected and captured in an information security and safety event journal. Event journal data provides an audit trail that should be protected using digital signatures, encryption and other safeguards. A system heartbeat record should document and monitor the safety, performance, and availability of telebiometric system devices and alert system administrators to security and safety events and changes. Heartbeat data should provide metrics that inform the continuous improvement of a telebiometric information security and safety management program. A signcryption cryptographic message wrapper should protect event journal, biometric reference template, and other telebiometric information to promote user security and respect for user privacy rights.

Keywords— ASN.1, signcryption, telebiometrics

1. INTRODUCTION

Organizations that rely on telebiometric technology should protect and manage the safety and security of their telebiometric assets [3]. Physical security and personnel security are important telebiometric considerations, and two of the pillars of information security management. The safe operation and performance of telebiometric systems are closely related to availability, a cornerstone of information security. Telebiometric systems safety management and performance monitoring should be integrated into the organization's overall information security management program.

This management program should be based on safety and security policies designed to achieve the objectives of the organization. A risk-based approach should be used to select and impose proper controls and to monitor their effectiveness. A periodic heartbeat message sent from each node of a telebiometric system to a central management collection point should document system compliance to safety and security policies in a secure journal, and help guide operations.

The international biometric information management and security standard, ISO 19092 recommends that compliance of a biometric system "should be periodically validated according to the organizations [*sic*] policy, practices and procedures" [4]. ISO 19092 defines a set of secure event journal records that "should be used in the capture of the validation material" [4]. However, the standard does not define a telebiometric system heartbeat record for monitoring and managing the safety, security and performance of biometric devices in a telebiometric system.

2. DATA SECURITY

Telebiometric information systems are vulnerable to loss of data integrity, origin authenticity, and confidentiality when their biometric data are transferred on "telecommunications network or via wireless communication devices", such as smartphone and tablet computers, using "wireless LAN or Bluetooth" [15]. The raw biometric data in a biometric sample is vulnerable to being "altered or intercepted by an attacker and used for illegal purposes" when being sent "to the signal processing component" [15]. Biometric data is also subject to attack when transmitted for "storage in registration or to the comparison component in authentication" [15].

When biometric devices are used for identification or verification, even when protected by liveness detection, "live-scanned data can be intercepted" during transmission and "replaced by forged biometric data" [15]. If biometric reference templates must be transferred, the confidentiality of their biometric data must be ensured. When templates are stored in a centralized template management system, their authenticity and integrity, as well as the confidentiality of their biometric data, must be protected from purposeful or accidental modification and from attack by trusted insiders.

ITU-T X.1086 proposes "countermeasures to ensure data integrity, mutual authentication, and confidentiality" to protect telebiometric information and users against threats, such as "hijacking, modification and illegal access" [15], and ITU-T X.1084 specifies a biometric authentication protocol and telebiometric system model profiles [14]. The X.1086 standard requires that personally identifiable biometric data, such as the "faces, fingerprints, irises, and voices" of users, be protected by a confidentiality safeguard and treated as the "providers' private information" [15]. ISO 19092 also requires confidentiality for biometric data and recommends that cryptographic safeguards ensure the integrity and authenticity of biometric objects [4].

Digital signatures based on the certificates in a Public Key Infrastructure (PKI) can be coupled with encryption

techniques to protect the confidentiality, integrity, and authenticity of biometric data and associated information. The SignedData cryptographic message used to sign electronic mail can provide data integrity and origin authenticity for an entire biometric object. Once created, the SignedData message can be encapsulated in an EncryptedData message to provide confidentiality for the entire signed object [17].

However, the signature followed by encryption approach employed in electronic mail systems lacks the processing efficiency demanded by modern telebiometric applications. The electronic mail approach also provides insufficient granularity. The entire biometric object must be encrypted, while only a few selected elements in biometric objects might require confidentiality protection. A more efficient, granular alternative for encrypting selected fields in signed telebiometric data is described in this paper.

This granular approach uses a new cryptographic message type, SigncryptedData [3]. This secure message allows a sender to sign and encrypt selected fields in a biometric object, and then to sign the entire object. SigncryptedData can provide confidentiality only where it is needed, and data integrity and origin authenticity over the entire object in a single cryptographic message.

3. PUBLIC SAFETY

3.1. Telebiometrics

Biometric recognition is a key form of automated identification and authentication based on the ability to distinguish individuals by their physiological or behavioral traits. Networked biometric systems are "increasingly used in a wide range of applications" [1] that enhance the quality of human life, including healthcare, law enforcement, border control, and financial services. These applications are enabled by "advanced pattern recognition algorithms applied through powerful ICT" [1] that merge remote biometric sensors and telecommunications.

To interact with biometric recognition applications, a human being must come into physical contact with "telecommunication systems and biometric devices" [13]. During this contact, telebiometric data from one or more sensors is "recorded by a measurement instrument" to collect biometric samples [13]. When "the human body meets electronic or photonic or chemical or material devices capturing biometric" data, the safe operation of these devices must be assured [13].

ITU-T Recommendation X.1081 defines a "framework for identifying and specifying safety aspects of telebiometrics" [13]. This international standard "provides a structure for categorizing the interaction of human beings with telecommunication terminals" [13]. The X.1081 framework can be used to derive "safe limits for the operation of telecommunication systems and biometric devices" [13].

X.1081 defines taxonomy of "all possible human-device interactions" [13]. This taxonomy provides a set of ASN.1 (Abstract Syntax Notation One) [8] information object identifiers whose values can be included in a biometric system heartbeat message that documents system safety and other operational characteristics. These values represent the safety posture of an active telebiometric system at a given point in time. Safety posture can be specified using "quantities and units of measurement based on the ISO/IEC 80000-series of standards" [13]. These values can be compared against device manufacturer recommendations, to ensure safe operation over the life of a device.

3.2. Information Management

Safety posture information that documents the operation of telebiometric system devices should be captured in a secure system event journal. Journal records should provide compliance validation material [4], such as evidence that equipment operation falls within the recommended safety levels for which the equipment manufacturer accepts liability. System safety posture can be used operationally to determine trends in device behavior over time. These trends may indicate that replacement, adjustment, repair, or other corrective action is needed to ensure public safety, system performance, and availability.

Compliance of a telebiometric system to the availability, performance, and safety objectives of the organization should be periodically validated. Secure system event journals should provide validation material that can be used to determine system compliance to organization policies [4]. Independent third parties should use the journal to validate system safety compliance and publish formal reports to ensure public trust in the ethical management and safe operation of telebiometric systems. Metrics gathered from system event journals should be used to inform management decision making, document the safety posture of the organization's biometric devices, and for continuous improvement of the organization's information security and safety management program

4. FUTURE STANDARDIZATION

4.1. Focus

ITU-T Study Group 17 (SG17) is widely known for its expertise in the development of information and communications technology (ICT) standards. Unlike organizations limited to a single technology domain, such as biometrics or security, SG17 can bridge multiple domains, bringing them together in standards with a cross industry focus that benefit multiple communities. Through its communications process and liaison activities, SG17 engages experts from across the world to standardize solutions that have a global impact. It is well positioned to play a central role in the development of standards that enhance the safety, security, and privacy of individuals and make sustainable communities possible.

SG17 includes experts in biometrics, information security, public key infrastructure, schema definition languages, and telecommunications technologies. The following proposals for standardization will require expertise from all of these disciplines. In particular, the involvement of the abstract

syntax, information security, telebiometrics, and Directory experts of SG17 will be needed to ensure the development of high quality solutions that are safe and secure. This combination of expertise and cooperative engagement makes SG17 uniquely suited to lead in the development of the following proposed standards that cross industry boundaries and technology domains.

4.2. Heartbeat message

Telecommunications-enabled biometric devices can support real time remote management and monitoring by system administrators. These devices can send periodic system heartbeat records to alert system administrators of security and safety events, such as changes in device settings or geographic location. Over time, heartbeat records can provide evidence of the safe and secure operation of a telebiometric system.

In aggregate, heartbeat record data provides a measure of the performance and availability of the devices in a system. Coupled with security event information, this data can be used to present a dashboard view of the security and safety posture of the system. When compared against policy requirements, metrics derived from heartbeat record data can indicate whether operations are achieving the policy objectives of the organization.

The safe and secure operation of telebiometric systems can be affected by "real-world factors such as 1) Human factors, 2) External environmental conditions, 3) System related issues" [9]. Heartbeat records logged in a secure telebiometric event journal provide an audit trail that document the safety and security posture of system devices.

Metrics collected from journal records can inform the continuous improvement of a telebiometric information security and safety management program.

An ASN.1 schema for a telebiometric system heartbeat message should be standardized by ITU-T. This message should be recorded in a secure telebiometric system event journal. Data values that measure device operational safety should be included in the heartbeat message. These values should be associated with those identified in the X.1081 taxonomy [13]. Security and performance information on the quality of a remote verification process of a biometric device should be defined as an optional heartbeat message field that carries an ISO/IEC 24761 report [6]. The report would not be used in this context for making an informed access control decision, but to provide operational values useful for system administration and for information security and safety management.

4.3. Telebiometric event journal

ITU-T Study Group 17 should develop a standard telebiometric event journal. This effort should provide an extensible ASN.1 schema [12] for journal records that makes widespread information exchange possible. A standardized schema would facilitate the development of interoperable applications from multiple vendors. Extensibility would support sustainable, flexible systems that can evolve over time, and allow any adopting community with a need to extend the schema.

This ASN.1 schema should support both compact binary and XML journal formats. Providing two ways to represent journal records would allow peer applications to support either or both formats. Application designers could choose the format best suited for use in a given context. They might use XML locally, then store or exchange telebiometric information efficiently using a compact binary format as depicted in Figure 1.

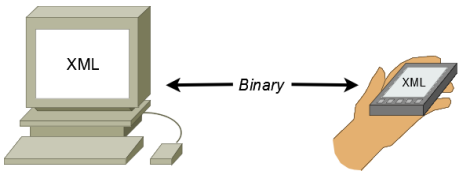


Figure 1. XML and binary information exchange

Binary event journal messages are appropriate for use in environments constrained by mobility, limited battery life, or bandwidth (e.g., wireless communications using hand held and personal devices). Compact binary journals are needed when there are high volumes of transactions (e.g., mobile internet commerce) or limited storage capacity (e.g., common access (CAC), personal identity verification (PIV), and other smart cards). Telebiometric system devices that must transfer data over radio waves or congested communications links or with devices whose period of use may be limited by battery life can benefit from using compact binary formats, and still leverage XML markup when needed.

A standard for a telebiometric event journal should build upon the event journal defined in the ISO 19092 biometric information management and security standard. This standard predates the emergence of cloud computing, smart phones and tablets, and the convergence of wireless telecommunications and biometric technologies. A new ITU-T standard could build on the security requirements of ISO 19092 to improve the security, privacy and safety of modern telecommunications users. Extensions could include support for system availability and performance monitoring and the safe operation of telebiometric devices.

Telebiometric event journals and records should be signed objects. Without the protection of a digital signature, the integrity of telebiometric information cannot be assured. Digital signatures can provide integrity assurance that information has not been altered since being signed. Signature verification can also provide evidence of data modifications that might otherwise go undetected. With PKI-based X.509 certificates [11], the origin of a signer can be determined. This determination can provide assurance that the information source can be trusted and allow sources that are not trusted to be detected.

Multimodal biometrics applications depend on reliable data collected from multiple sensor locations and vendors. International travel that depends on biometric enabled passports has led to an increasing need for sharing biometric and other personal information across legal and regulatory boundaries. Persons charged with maintaining our safety and security require reliable telebiometric information to make informed decisions. This reliance on technology makes origin authenticity and data integrity crucial for ensuring our communities are sustainable.

4.4. Cryptographic schema

Telebiometric information should be protected using appropriate cryptographic safeguards and other security measures. System managers, regulators, and the public should be confident of the integrity and authenticity of the telebiometric information used by decision makers to ensure public safety and security. Information assets that should be protected include the biometric data of system users, personally identifiable subscriber information, and event journal data used to improve information security and safety management programs. These objects should be signed using a digital signature associated with a PKI using an ASN.1 SignedData cryptographic message.

SignedData is an extensible cryptographic message that provides data integrity and origin authenticity services using a PKI-based digital signature. SignedData is used in many network security protocols. One variation of the message is used to distribute X.509 certificates and certificate revocation lists (CRLs). SignedData is one of a set of cryptographic key management messages referred to as Cryptographic Message Syntax (CMS) [17].

SignedData is widely deployed in network security protocols, including Secure Sockets Layer (SSL) and Transport Layer Security (TLS). SignedData is used to distribute X.509 certificates in the Organization for the Advancement of Structured Information Standards (OASIS) Web Services Security (WSS) X.509 Security Token. SignedData is included in tool kits on a number of popular operating systems, including several versions of Windows and Unix servers.

Several widely deployed CMS standards define a SignedData message. These include the RSA Public Key Cryptography Standard (PKCS) #7, the Secure Electronic Mail (S/MIME) CMS standard defined by the Internet Engineering Task Force (IETF), and the X9.73 Cryptographic Message Syntax: ASN.1 and XML standard [17]. However, there is no international CMS standard that uses valid ASN.1 syntax to define its message schema.

The data security of a number of international biometric standards depends on the cryptographic schema defined in the IETF CMS standard. IETF CMS SignedData is used in the "International Civil Aviation Organization (ICAO) standard for machine-readable travel documents (MRTD) including electronic passports" [1], the ISO/IEC 24761 Authentication context for biometrics, and the ISO/IEC 19785-4 Common Biometric Exchange Format Framework (CBEFF) – Security block format specifications.

Two informational IETF CMS publications contain valid schema definitions. The message schema specified in the normative IETF 3852 CMS standard [16] does not contain valid ASN.1 syntax. Though the standard lists the current ASN.1 standards in its reference section, these versions of the ASN.1 standards are not used. Instead, the IETF 3852 CMS schema is based on X.208, the deprecated 1988 version of ASN.1 that was withdrawn as a standard in 2002 [10]. The known defects in X.208 were never corrected before it was abandoned. These defects were corrected in the current ASN.1 standards [8].

New types to support national languages were never added to X.208. The Distinguished and XML Encoding Rules were never defined for use with the X.208 syntax. The IETF CMS schema attempts to mix X.208 syntax with syntax from post-1988 versions of ASN.1 in the same ASN.1 module, which is not allowed in the ASN.1 standards. Based on this ambiguous syntax, tools that implement the ASN.1 standards are not able to generate applications that conform to any version of the ASN.1 standards. The reliance on invalid cryptographic syntax for data security in the ICAO, ACBio, and CBEFF standards does not enhance the ability of these important biometric standards to provide information assurance and security.

ITU-T should create an international Cryptographic Message Syntax standard. This work should be developed by the security, PKI, and schema language experts in Study Group 17 (SG17). A new CMS standard should be created as a generic application of ASN.1 in the X.890-series of recommendations and standardized jointly with ISO and IEC. The new CMS standard should contain valid ASN.1 syntax and its schema should support all of the compact binary and XML encoding rules.

The X9.73 CMS standard contains valid syntax whose binary encodings are valid IETF CMS binary values. To the greatest degree possible, XML encoded values of the X9.73 CMS standard should be valid encodings in the new ITU-T CMS standard. The ITU-T standard should follow the approach of the X9.73 standard and collect in a single set of ASN.1 modules the schema for all of the key management techniques defined in CMS. These techniques include key agreement, password-based encryption, and constructive key management.

4.5. Signcryption message

Signcryption is a relatively new cryptographic primitive standardized in ISO/IEC 29150 [7]. Signcryption uses a special algorithm that blends together signature and encryption schemes to perform digital signature and asymmetric encryption functions simultaneously. This hybrid cryptographic technique provides confidentiality, data integrity, and origin authenticity in a single, efficient operation.

Efficient cryptographic protection methods are needed in telebiometric systems if they are to empower human users and manage their safety and security risk. Such methods help to ensure that telebiometric systems provide the privacy and security citizens need to build sustainable

communities, and the reliable management information system providers need to ensure high quality, safe, reliable service.

Signcryption offers a smaller message size and faster processing speed compared to *sign-then-encrypt* signature followed by encryption techniques [2]. Unlike safeguards that rely on symmetric keys, the reliance of signcryption on asymmetric cryptography makes non-repudiation possible. These features make signcryption ideal for protecting telebiometric system information, such as ISO/IEC 19785 templates, ISO/IEC 24761 reports, and ISO 19092 journals.

In the paper *Protecting Biometrics Using Signcryption* presented at the 2012 ID360 Global Forum on Identity [3], an ASN.1 schema for a signcryption message is defined. The paper proposes that a new SigncryptedData type be added to the X9.73 CMS [17] standard to extend CMS functionality. The proposed schema is based on the familiar SignedData type used to protect "electronic mail, biometric enabled watch lists, biometric reference templates, and [*sic*] biometric elements in the Electronic Biometric Transmission Specification (EBTS)" [3] transactions used by law enforcement.

The SigncryptedData type allows biometric information objects to be signed, and for selected elements within these objects to be encrypted. In the *signcrypted-components* mode, one of three proposed modes of operation, one or more elements of an information object are signcrypted. The resulting object is then cryptographically bound to one or more attributes under a digital signature. These signed attributes must include a manifest of all of the elements in the information object that have been signcrypted.

This field-level encryption capability coupled with a digital signature makes the SigncryptedData type ideal for use in managing the safety and security of telebiometric systems, and for protecting the sensitive elements found in biometric templates, verification reports, and event journals. Using SigncryptedData, a biometric reference template can be signed, and the biometric data component within the template can be signcrypted using the same cryptographic keys. While there is security risk that must be managed when using cryptographic keys for more than one purpose, this solution meets the security requirements of ISO 19092 by ensuring the user's biometric data remains confidential within a reference template having origin authenticity and data integrity protection.

Rather than including SigncryptedData in the optional signature block defined in the ISO/IEC 19785 CBEFF standard, stronger protection of biometric templates can be achieved when SigncryptedData is used as a message wrapper that encapsulates the entire template. A message wrapper approach allows a trivial attack on reference templates to be detected using signature verification. In environments in which the optional signature block is not required to be present, it is possible for a low skill attacker to remove the entire signature block to thwart the signature safeguard's effectiveness. Telebiometric systems that serve global communities of mobile users should ensure the effectiveness of security safeguards in all usage contexts.

When a biometric verification process is performed at a remote location, identification and access management (IdAM) systems must make authentication decisions using devices in an uncontrolled environment. Relying parties may lack administrative control over the remote biometric devices used to authenticate users that access their systems. Remote biometric systems owned or operated by others may not be subject to the security and privacy policies of the resource owners.

The ISO/IEC 24761 (ACBio) standard [6] provides relying parties security and performance information on the quality of a remote biometric verification process. ACBio transfers a biometric verification process report to the relying party. A digital signature associated with a SignedData message protects this report. However, in the current version of ACBio, this SignedData message is based on an IETF CMS schema, which does not conform to any version of the ASN.1 standards.

Verifying the signature and validating the certification path from a report signer to a trust anchor provides the report recipient data integrity and origin authenticity assurance. The report itself gives a relying party assurance that the match decision returned by a remote biometric verification system can be trusted. ACBio provides a means for “falsified reference templates, forged raw data” and “unreliable biometric devices” [6] to be detected. ACBio reports allow the security risk associated with remote biometric verification to be managed. The SignedData message provides integrity and authenticity protection, but does not protect the confidentiality of ACBio information.

ACBio respects human values by ensuring that its reports do not contain personally identifiable information, such as biometric data from a biometric reference template or the biometric sample of an identity claimant. However, the SignedData message wrapper does not prevent ACBio device identification and operational information from being collected and viewed by an eavesdropper or by a trusted insider. Transport layer encryption could be used to provide point-to-point protection, but that approach does not provide an end-to-end confidentiality solution.

ACBio reports contain biometric device identification, match control configuration settings, and match processing information that might benefit an attacker. This information could be aggregated over time to help plan attacks or to identify a weakness in a biometric system. Replacing the SignedData wrapper with a SigncryptedData [3] wrapper would add confidentiality services to the data integrity and origin authenticity services provided by SignedData. This would extend the usefulness of ACBio to law enforcement, defense and intelligence environments, where access to system operational information may be restricted by security classification level or on a need-to-know basis.

The SigncryptedData message could extend ACBio with support for additional signed attributes added by any user with a need. These attributes might include system heartbeat information, operational safety and performance reports, security classification markings or security and privacy policy. For stationary equipment, a geolocation

attribute could be used to monitor and detect unexpected relocation of a system device.

5. CONCLUSION

ITU-T should create a standard telebiometric event journal whose records are defined in an ASN.1 schema. Records should document biometric system security events following the events defined in ISO 19092 [4]. Event journals should be signed objects. When necessary, event journal records should also be signed, and selected fields should be encrypted to protect the privacy of human beings. A system heartbeat journal record should be defined along with one or more records containing useful metrics collected from journal entries. A field in this heartbeat record should allow optional inclusion of an ACBio system verification report.

ITU-T should create a new Cryptographic Message Syntax (CMS) security standard whose messages are defined using valid ASN.1 syntax. This new generic application of ASN.1 should be standardized jointly with ISO/IEC. ITU-T experts should promote its adoption in international standards that are important for securing telebiometric information, such as ICAO, ISO/IEC 19785 and ISO/IEC 24761. Improving security standards that contain invalid ASN.1 syntax can enhance the safety, security, and privacy of telebiometric system users.

ITU-T should standardize the schema and associated cryptographic processing of a new SigncryptedData type. This new type should be added to an ITU-T CMS standard. All of the signcryption mechanisms and cryptographic algorithm identifiers defined in the ISO/IEC 29150 standard [7] should be supported. Though signcryption is not intended for use in signing X.509 certificates, the Directory standards should be examined to determine whether modifications are needed to support signcryption operations.

REFERENCES

[1] Biometrics and Standards. ITU-T Technology Watch Report #12, December 2009. Retrieved November 21, 2012 from http://www.itu.int/oth/T230100000D/en

[2] Dent, Alexander W. (2004). Hybrid cryptography, Cryptology ePrint Archive Report 2004/210. Retrieved November 21, 2012, from http://www.signcryption.org/publications/pdffiles/Dent-survey-eprint-04-210.pdf

[3] Griffin, Phillip H., *Protecting Biometrics Using Signcryption*. Proceedings of ID360: The Global Forum on Identity, the Center for Identity, University of Texas at Austin, 2012. Retrieved November 21, 2012, from http://phillipgriffin.com/innovation.htm#ID360

[4] ISO 19092:2008 Financial services – Biometrics – Security framework.

[5] ISO/IEC 19785-1, Common Biometric Exchange Formats Framework – Part 1: Data element specification

[6] ISO/IEC 24761 (2009), *Authentication context for biometrics*.

[7] ISO/IEC 29150 (2011), *Signcryption*.

[8] Larmouth, John, *ASN.1 complete*. San Francisco: Morgan Kaufmann Publishers, 2000. Retrieved November 21, 2012, from http://www.oss.com/asn1/resources/books-whitepapers-pubs/larmouth-asn1-book.pdf

[9] Pour, Babak Goudarzi, There's A Metric for That': How 'Big Data' Impacts Biometrics Market and Industry, June 16, 2012. Retrieved November 21, 2012, from http://biouptime.com/2012/06/16/the-business-impact-of-big-data-on-biometrics/

[10] Recommendation ITU-T X.208 (1988). *Specification of Abstract Syntax Notation One (ASN.1)*.

[11] Recommendation ITU-T X.509 (2008). The Directory: Public-key and attribute certificate frameworks. Retrieved November 21, 2012, from http://www.itu.int/rec/T-REC-X.509

[12] Recommendation ITU-T X.680 (2008). *Abstract Syntax Notation One (ASN.1): Specification of basic notation*. Retrieved November 21, 2012, from http://www.itu.int/rec/T-REC-X.680-200811-I

[13] Recommendation ITU-T X.1081 (2011). *The telebiometric multimodal model – A framework for the specification of security and safety aspects of telebiometrics*. Retrieved November 21, 2012, from http://www.itu.int/rec/T-REC-X.1081-201110-I

[14] Recommendation ITU-T X.1084 (2008). *The telebiometrics system mechanism - Part 1: General biometric authentication protocol and system model profiles for telecommunications systems*. Retrieved November 21, 2012, from http://www.itu.int/rec/T-REC-X.1084-200805-I

[15] Recommendation ITU-T X.1086 (2008). *Telebiometrics protection procedures – Part 1: A guideline to technical and managerial countermeasures for biometric data*. Retrieved November 21, 2012, from http://www.itu.int/rec/T-REC-X.1086-200811-I

[16] RFC 3852 *Cryptographic Message Syntax* (2004). Internet Engineering Task Force (IETF). Retrieved November 21, 2012, from https://www.ietf.org/rfc/rfc3852.txt

[17] X9.73-2010 *Cryptographic Message Syntax – ASN.1 and XML*. U.S.A.: American National Standards Institute (ANSI).

SESSION 6

STANDARDIZATION ISSUES

OPEN STANDARDS: A SHRINKING PUBLIC SPACE IN THE FUTURE NETWORK ECONOMY?

William H. Melody

Center for Media and IT
Aalborg University Copenhagen

ABSTRACT

The capacity to sustain most communities depends on their access to technical, economic and political resources. The interaction among technologies, markets and government policies that direct technology and market activity generates the resources. Most countries, and many international agencies have adopted information society policies to promote broadband infrastructure and NGN development as a platform for ICT applications. The goals are economic development and universal access, i.e., more inclusive communities. This paper examines how ICT sector market and policy trends are influencing the environment for developing innovative NGN applications and their essential supporting standards.

The interaction between ever more expansive and generous patent and copyright (IPR) awards in the ICT sector with the winner-take-all network characteristics of most ICT and content markets is fostering oligopoly markets based on proprietary standards. This trend toward increasing policy-permitted standards and market exclusivity as the foundation for asset values and industry growth steadily narrows the scope for NGN applications based on open standards. It reduces opportunities for participation in the development of future knowledge communities. A major initiative is proposed to build the evidentiary and analytical support for policy reforms that will reverse this trend.

Keywords— Telecommunication reform, IPR, NGN, open standards,

1. INTRODUCTION

This conference is exploring opportunities for ICT international and open standards processes to play a *catalytic role* in the development of ICT applications that will help make communities more sustainable. The concept of sustainability has many dimensions in this context, but for local, national and international societal communities the most important may be to make them more inclusive.

This has been an important dimension of telecommunication (telecom) development and its standardization activities throughout its history with respect to extending network interconnection, interoperability and access. Inclusiveness is the theme of much of the government policy rhetoric and research that permeates the information society literature today. A significant push forward in the commitment to new standards development for information society products, networks and services can provide essential substance for the implementation of these policy objectives.

The major challenges in building more inclusive communities are to expand the capacity to sustain them, and to reduce unnecessary exclusionary barriers. The next stage of ICT development provides unique opportunities for enlarging the capacity to sustain more inclusive communities because they will be centered on NGN product and service *applications* throughout most industry, government and other institutions. The ICT revolution is moving beyond the ICT sector and beginning to penetrate modes of production and methods of operation in fundamental ways.

This enhanced integration of ICTs is associated with new areas of technological integration that will expand the *diversity of interests* participating in ICT standards development. It will change boundary relationships between *open and proprietary standards*, and between the traditional domains of *open standards development* on the one hand and *patents and copyright* on the other. ICT standards associations and standards development processes will not be exempt from making necessary adaptations to the dynamic changes associated with the next wave of the ICT revolution.

2. APPLICATIONS OPPORTUNITIES: THE 3rd WAVE OF TELECOM REFORM

We are now at the beginning of the third wave of telecom sector reform. Each wave has been driven by changes in industry policy and regulation, new technologies and changing markets. Each wave has brought the telecom sector into convergence with other sectors of the economy that have stimulated innovation. New standards have provided an essential foundation for converting this innovation into new products, networks, services and markets. These standards have been developed from two very different models of promoting innovation that have been for the most part mutually supportive and sometimes synergetic. For the future their relationship is uncertain.

Monopoly proprietary - IPR protected proprietary standards associated with a technology that may or may not be licensed for use by others at whatever fees the patent holder

decides to charge. They may or may not be extended to become de facto industry standards.

Open public – standards accessible to anyone. The extent of openness varies when it is implemented in different circumstances. Industry standard setting organizations (SSO) usually require openness in the development of the standards as well as access to them. ITU-T chairman Malcolm Johnson defines these as" *standards developed in an open and transparent manner based on consensus and in which any intellectual property rights can be acquired either free of charge or on reasonable and non-discriminatory terms, so that anyone anywhere can develop products and services to these standards*" [1].

The first wave of reform involved policy-driven reforms of national telecom monopolies and the liberalization of telecom markets. Firms from other industries, notably electronics and computing, stimulated a technological revolution in the production of telecom equipment ranging from an enormous new diversity of handsets and terminal attachments to central office switching equipment as telecom networks began their gradual conversion to digital standards and data communication began to grow. New standardization agreements were instrumental in enabling terminal interconnection and data communication services in the new liberalized industry structure. The first wave of reform was essentially a restructuring and modernization of the telecom sector, opening the door to entry from technologies and standards from related industries [2].

The second wave of policy-driven reforms supported the increasing convergence of ICTs and the pervasive digital penetration of telecom facility and service networks. Voice, data, TV, pictures and video no longer required separate networks. Any form of telecom service could be supplied using any technology or combination of technologies. Digitalization improved spectrum efficiency dramatically, allowing mobile growth to take-off. The Internet was transferred to the commercial market and its growth soared. Internet protocol began to penetrate telecom networks allowing a further integration of all forms of telecom on the Internet layer, including public voice telephone. It provided a clear demarcation between network infrastructure capacity and network services provision. Triple and quadruple-play network services could be offered by providers other than incumbent telecom operators.

The second wave of telecom reform was essentially the conversion of the telecom sector into a much larger ICT sector with enormously expanded technological and network service capabilities. Once again it was made possible by major contributions from the standards community enabling interoperability among a rapidly increasing number of products, services and networks.

During this second wave, much of the new technological development came from the IT industry which also produced major innovations with computer applications in production processes ranging from robotics to CAD-CAM systems. But demonstrations of significant improvements in economic productivity were difficult to demonstrate. Research documented that the primary beneficiary of the ICT revolution was the productivity of the ICT sector itself, not the economy as a whole [3].

The third wave of telecom/ICT policy reform, now still in a very early stage of its development, is focused on extending the ICT revolution throughout the economy and society as a whole, to improve productivity and significantly increase the capacity to sustain more inclusive communities. Once again this is driven by major government policy initiatives – broadband, NGN access and applications, information economy, knowledge society, etc.

At present, government policies are preoccupied with the development of "ultra-fast" broadband networks intended to provide universal access to NGN networks and services. Digital agenda policies list many potential applications that will help create future information economies and knowledge societies. But at this stage, these are mostly wish lists lacking substance. Substance will begin to be supplied when the new standards for the specific new ICT applications are developed [4].

Implementing the third wave of reform requires integration of ICT equipment, information processing and telecom services into specialized industry and government activities that have very different policies, practices, operating methods and business cultures from those in the ICT sector. This will inevitably involve the development of a mix of generic and industry-specific open standards with "enterprise" proprietary standards in response to the specific needs of specialized activities in specialized operations.

Although the opportunities are great, the challenges are formidable. A high degree of openness, crowd sourcing and experimentation will be needed to acquire the knowledge to reach optimal solutions. The development of standards will be central to this effort, but it will require that standards processes be adapted to new multi-industry requirements, including co-ordination among SSO from different industries and an inherited body of diverse proprietary standards.

3. FORCES SHAPING THE DEVELOPMENT PATH FOR ICT APPLICATIONS

The development path for ICT applications in the NGN era will be shaped by the interactions among *technologies, markets and policies*, as it has since the earliest days of electronic communication. Alexander Graham Bell's 1876 U.S. patent for "Improvements in Telegraphy" (i.e., the telephone) is still regarded by many as the most economically valuable patent ever awarded. National patent policies around the world have been instrumental in shaping the direction of technological development in telecom, computing and ICT. They provided policy protection for the original network monopolies. When the patents expired, industry-specific government policies decided the industry structure thereafter, which in many countries was acceptance of direct market competition for a period, and later industry concentration into national monopoly networks.

As networks grew, the market incentives of the dominant, often monopoly national network operators produced a two-pronged approach to network standards development to facilitate interoperability. The national monopolies supported standards establishing interoperability with one another as this extended the reach of their networks to new international markets, but they opposed them with any competitors in the national markets they dominated. One of the earliest activities of the ITU was to facilitate the development of technical standards to facilitate international telecommunication.

The dynamic interactions among technologies, markets and policies that have shaped, and are shaping the telecom sector development path are illustrated in Figure 1. Technologies have their impact on development by their applications. Markets have their impact on economic growth by the products and services they provide. Government policies have their impact through the implementation of laws and regulations that promote and constrain development in specific ways. The influences of each can be seen at virtually every major historical change to the development path of the telecom sector.

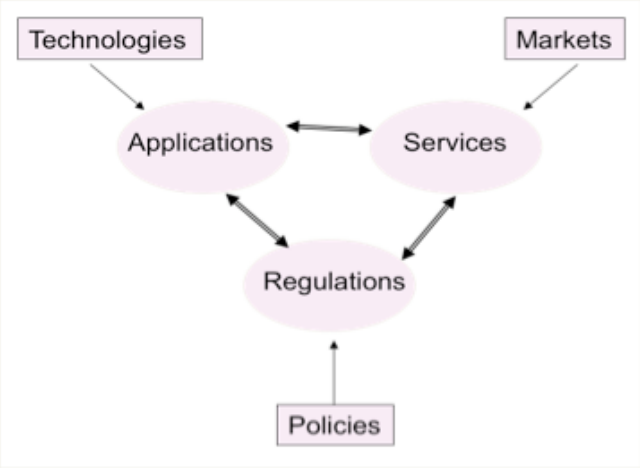


Figure 1. Primary Forces Shaping the Telecom/ICT Development Path

Sometimes new technologies have stimulated positive market and policy responses, as for example in the cases of satellites and computer-communication; sometimes market developments have stimulated policy and technology responses, as for example in the case of mobile communication where exponential growth has prompted major changes to spectrum policies and increased attention to new approaches to mobile standards; sometimes it has been changes in policies and regulations that have opened new opportunities for market and technological development, as for example when the US FCC required the interconnection of all network terminals that met approved standards in 1968, and more recently policymakers unbundled the provision of network services from network capacity, opening opportunities for new directions of technological development and what we know today as the Internet [2].

A review of the historical development of telecom shows that there have been times when great synergy existed among the technology, market and policy domains shaping development, as illustrated by the examples noted above where changes in one domain triggered opportunities in the others and even a virtuous interactive cycle of developments among the three domains. But this has not always been the case. For example, a century of telecom monopoly policies severely narrowed opportunities for technology and market developments imposing severe restrictions on the synergy potential from the interaction of technologies, markets and policies.

In different times and circumstances the synergy from the interaction among technology, markets and policies has ranged from highly positive to highly restricting and constraining. In some cases they seem to have been working at cross-purposes. For example, government regulations and/or monopoly operators have restricted the application of new technologies in extending networks to un-served areas despite the existence of universal access policies. This exists even today in many countries with respect to Wifi technologies and services. [5]

To date, government policy directives and regulations associated with the development of information societies have focused primarily on economic and financial issues relating to telecom network rollout and new network services development. Although there is continuous reference to the importance of an environment to stimulate innovation, there is rarely any consideration of the market conditions and government policies affecting innovation in the IT industries that are developing the new technologies. If new NGN applications are to help achieve the policy objectives of inclusiveness and universal access, there must be strong policy support for innovation based on open standards.

4. STANDARDS, IPR AND MARKETS

Patents providing monopoly proprietary standards protection has worked reasonably well in stimulating innovation and technology development in most industries during the industrial era when multiple and overlapping patents were not widespread and seldom raised irresolvable conflicts. A high degree of interdependence among patented standards had not yet developed and computer software didn't exist. SSO facilitated industry standards development under conditions where the benefits of expanded markets were generally seen by the major industry players as justifying their forthright and co-operative participation in open standards development process, including willingness to classify standards-essential patents for open standards development. Network industry standards development was able to incorporate broader public interest considerations in interconnection and interoperability among network equipment and services, as illustrated by the early establishment of the ITU in 1865 as a sector-specific SSO.

The technological revolution in the converged ICT sector has led to explosive growth in patent and copyright (IPR) applications and awards, and fundamental changes in the

role and significance of IPR in this sector. The dramatic growth of software as a key resource in most digital technologies, and of electronic information products and services also has contributed to a massive expansion of IPR in this sector. Obtaining, defending, negotiating and managing IPR has become as important to the competitive positioning of firms as innovation and the development of improved technologies [6].

The major industry players have become more reluctant to identify and volunteer standards-essential patents in the development of industry standards by SSO, precisely at a time when their forthright participation is required if the NGN and its applications are to achieve their potential. For the future development of NGN applications, the current trend suggests the synergy of the productive past relationship between the different models for standards development is being eroded rapidly and the capabilities for timely industry and public interest standards development significantly reduced. This can severely restrict the opportunities for making catalytic ICT applications contributions to sustainable inclusive communities.

4.1 IPR in the Information Economy

IPR policy and practice is the major force today shaping not only ICT innovation, but also sector development. This has not been the result of decisions by governments to use IPR as a vehicle to achieve particular policy objectives for the sector. Rather it is the result of the major players using, and abusing the IPR processes to protect and enhance their competitive market positions. The essential economic characteristics of ICT networks, combined with the weakening of the patent and copyright processes have provided a market environment where there can be enormous financial rewards for firms highly skilled at manipulating the IPR processes and enormous financial losses for firms that are not.

Networks are characterized by economies of scale and scope. Additional economies can be obtained, and new services introduced when network interconnection and interoperability are provided. These characteristics are well-known and their effects in telecom networks widely researched. ICT convergence has introduced additional network considerations. The competition among technologies in the early stages of new network development produces positive feedback effects favoring the technology with the earliest development of the largest network. This too is reflected both in the size of the network provided and in the demand of users who obtain greater benefits by being part of larger networks. There are also externality effects as the choices of potential buyers of network products and services are influenced as well. In addition, the marginal cost of extending networks in digital products and services are extremely low, approaching zero for information products that can be copied.

Once a network technology establishes a foothold in the market, there are often significant costs for users to switch to a competitive technology, especially if it has a smaller market share. Once a network becomes dominant, the risks and costs of establishing a competitive network often create such high barriers to change that a lock-in effect can provide benefits for an extended period. As a result of these compounding effects, the most typical result of competition in digital network technology development is the early mover to a market-tipping point wins the entire market [7].

Even in rapidly growing markets, for the competitors there is a serious risk that despite rapidly growing markets, they are playing a zero-sum game. The winner takes it all. Thus the focus of network technology competition is strategic promotion of one's own technologies and strategic delay of competitive technologies, forces quite outside the merits of the competing technologies.

As ICT digital networks have grown and expanded in capabilities for product and service provision, they have incorporated the IPR products and standards of other firms, usually by some form of licensing. Most networks have come to be supported by a large and increasing collection of technologies, all supported by IPR and many of which are essential to the effective functioning of the networks. Clarity with respect to IPR licensing is important because a disagreement about IPR rights to a small, but essential technology in the network can jeopardize the entire network. This characteristic of new network technology development provides yet another powerful incentive for industry players to focus as much attention on the pro-active management of IPR as upon the development of new technologies.

The ICT network and IPR characteristics described above have stimulated, and in many cases forced industry players to commit substantial resources to manipulating the IPR processes to advantage, with the result that patent and copyright authorities have been overwhelmed. But this ICT industry manipulation of the IPR process is creating increasing negative feedback that is significantly raising industry costs, distorting competition, providing major unnecessary barriers to entry for smaller firms, making industry and public interest standards development more difficult, and creating increased rather than decreased uncertainty with respect to IPR protection of technologies and standards. [8].

The distortions in IPR processes relating to ICT digital technologies are by now well documented. Defensive IPR awards are obtained not to develop a technology but rather to prevent competitors from doing so, thereby protecting the ground around a firm's existing IPR and establishing a potential claim if another firm develops a related technology. A robust inventory of IPR to negotiate cross-licensing agreements is now necessary to clear a path through IPR "thickets" created by the decisions of all the main players to build a robust inventory of defensive IPR.

A lucrative revenue stream can be developed from IPR used to "ambush" competitors with claims when their new technology development is approaching completion, or when new industry standards are ready for implementation, or it is desired to force a delay in implementation. Lucrative law practices have developed based on trolling patent thickets for patents where violations can be claimed and "hold-up" license fees sought in court. The threat of expensive and time-consuming patent infringement claims

can win large payouts from large firms and bankrupt small ones.

Strategic gaming of the IPR processes begins with manipulation of the criteria that must be met to justify an award of IPR, for example, a demonstration of "novelty and non-obviousness" to obtain a patent. These have been continuously watered down by corporate patent engineers and lawyers who have overwhelmed patent offices. The threshold for demonstrating innovation that justifies an IPR award has been reduced virtually to the use of innovative words in the description of the claim. The problem is magnified by a trend in IPR applications away from specific claims to general and expansive claims, which leads to overlapping IPR awards without clear limits, inviting the testing of those limits through aggressive negotiation and court cases [9].

IPR awards are also for far longer periods than can be justified by the rapid pace of ICT technological improvement– commonly 20 years for patents and a lifetime or longer for copyright. As most new generations of technology build on the older generations, continuous incremental innovation is the most important part of the process. Thus old and obsolete IPR in a firm's inventory may sometimes be used to claim license fees from new technological developments. The vague and general criteria for awarding IPR provide the foundation for the strategic gamesmanship described above, which will not change until they are changed.

Almost every major ICT company is involved in ongoing patent battles. The number of patent lawsuits filed in U.S. district courts increases steadily and reached 3,260 in 2010. In the smartphone industry alone, according to a Stanford University analysis, as much as $20 billion was spent on patent litigation and patent purchases over a recent 2 year period. Patents for software and some kinds of electronics are now so problematic that they contribute to a so-called patent tax that adds as much as 20 percent to companies' R&D costs [10].

The experience of Apple is instructive. After the company was a victim of a patent attack that cost the company $100 million, it decided to use patents as leverage against competitors. Since 2000, the number of Apple's annual patent applications has risen by a factor of ten, receiving 4,100 patents. In 2010 spending by Apple (and Google as well) on patent lawsuits and patent purchases exceeded spending on R&D for new products. In 2011 Apple won a $1 billion patent infringement judgment against Samsung – now under appeal.

Apple has won patents, broadly applicable, for pinching a screen to zoom in, for using magnets to affix a cover to a tablet computer and for the glass staircases in Apple stores. Its patent number 8,086,604 (the Siri patent) has acquired special notoriety. The initial application to the U.S. patent office in 2004 for a voice and text based search engine was rejected as an obvious variation on existing ideas. In following years the application was modified, resubmitted and rejected eight times.

The tenth application was approved, and soon after used as a basis for lawsuits against Samsung, including the recent billion dollar winner. The initial patent application was considered to be "aspirational" as the iphone didn't exist. The patent not only protects Apple's smartphone technologies, it may give Apple ownership of the smartphone market and the now-commonplace technologies associated with it, as well as dominance in the smartphone "apps" market, an early NGN applications development [10].

The number of patent applications filed with the U.S. patent office has increased more than 50 percent over the last decade to more than 540,000 in 2011, the majority in the ICT/electronics sector. I.B.M., the world's largest ICT applications provider, received more patents in 2012 than any other company for the 20th consecutive year. Its 6,478 awards in 2012 were more than twice the number received by any other company except Samsung which received 5,081. Most of IBM's treasure chest of patents relate to IT and ICT application in vertical markets [11].

Microsoft received 2,613 patents in 2012. Apple was awarded 1,236 patents, a 68 percent over 2011. Google received 1,151 patents, a 170% increase over 2011. This evidence indicates that the patent battles in the ICT sector will escalate further and have a major influence on future sector development, including standards development for NGN applications. Any claim that can't be patented can obtain copyright protection.

4.2 Standards and NGN Development

Successful telecom standards development in the past has rested on a mutually recognized balance between proprietary monopoly and open industry standards, with public interest considerations being recognized in open industry standards where they are relevant. Although the boundary determining when proprietary or open industry standards best serves the public interest has never been defined, industry player decisions on when their best interest was served by participating in open industry standards development coincided reasonably well with the assessments of SSO, and with the public interest requirements of government policies for telecom. A major influencing factor was the obvious benefits of interconnection and interoperability in extending markets for all industry players.

However IT industry development has not been based on similar network development, but rather on the sequential development of stand-alone products and services driven by IPR in proprietary standards. Interoperability was not necessarily desired for security reasons, and often avoided to protect market positions. With the interconnection of IT terminals to telecom networks, telecom standards provided a new market for an extension of this IT industry model. As ICT convergence has grown to a more advanced stage and the telecom network converted to digital standards, internet protocols and software technologies, greater integration of the telecom and IT models for standards development has become increasingly necessary. This has brought the standards development models of the two sectors increasingly into conflict. The weaknesses of the

IPR process that have been exploited by the IT industry have spread to the converged ICT sector and the development of NGN technologies.

The aggressive pursuit of IPR in proprietary standards as a lucrative source of licensing revenue and a valuable competitive weapon significantly reduces the interest of IT-based firms in agreeing to give this up by designating their IPR as "standards-essential" for NGN open standards development. Telecom firms, now operating in the same ICT industry sector, are often forced to play the IT sector IPR game. We are already seeing increasing evidence of a reluctance to declare standards-essential patents, as well as legal disputes over claimed infringements of these patents and the level of fair, reasonable and non-discriminatory (RAND) license fees associated with them. Standards-essential patents are being brought into the IPR strategic gamesmanship of aggressive industry players to protect markets and block competitors. The domain of open industry standards is steadily shrinking [12].

Leading SSO have begun to take steps to counter these developments that are narrowing the scope and effectiveness of open industry standards development and reducing industry competitiveness, efficiency and growth opportunities. For example, an ITU, ISO and IEC led initiative has led to the establishment of a Common Patent Policy for standards development, and SSOs have published more precise guidelines regarding the treatment of IPR. ITU and the European Patent Office signed an agreement in 2011 to share information.

More recently, the ITU called a special Patent Roundtable in October 2012 to examine the "patent wars" jeopardizing the effectiveness of FRAND-based patent policies in the context of standards-essential patents with all the stakeholders. These initiatives undoubtedly help, but given the systemic institutionalized nature of the problem, fundamental institutional changes will be needed to reverse the current trend [18].

4.3 The NGN Applications Environment

The NGN provides the platform ICT applications throughout the economy and in government and other organizations - vertical markets. Initiatives are underway in many of them now, for example, e-health, e-learning, smart grids, intelligent transport systems, etc. This requires industry convergence of a different kind, not driven by IT technologies transforming telecom networks, but rather the integration of NGN products and services with specialized firm, industry and organization IT systems that already have been integrated into their respective production and administrative operations. This will require cooperation and collaboration between ICT and application sector SSO in the vertical markets, introducing a more complex environment for standards development that will include a new set of complex issues relating to standards-essential patents in open standards development.

As a result of comprehensive ICT convergence and policy changes liberalizing telecom markets the basic functions of the telecom network have been unbundled into clearly identifiable, but inter-related sub-networks, all of which require compatible efficient standards for effective operation. These unbundled network markets are illustrated in simplified form in Figure 2. Layer 1: Infrastructure Facilities represents the historically familiar territory of standard-setting primarily for hardware technologies where standards-essential patents have been an important element. With digitalization and internet protocols, Layers 2, 3 and 4 are associated primarily with software.

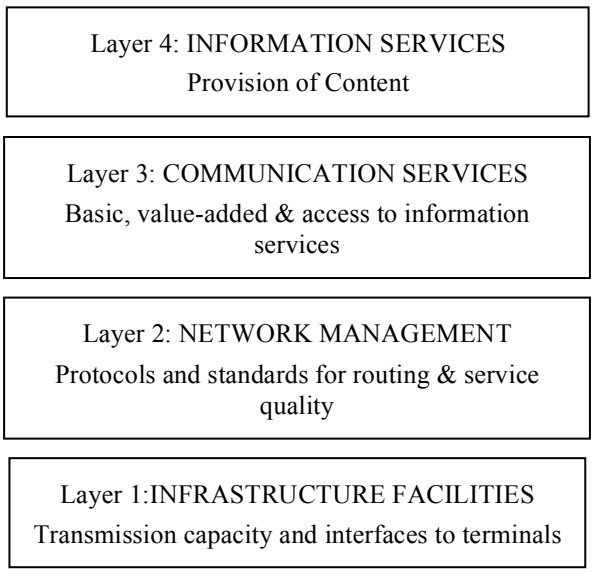


Figure 2. Unbundled Network Markets

Software was protected by copyright during the first quarter century of its development and specifically rejected by patent offices. Protection against copying was provided, but not the use of the same formats and ideas in other software development. A 1981 landmark US Supreme Court decision held that some software could be patented. Since then the interpretation of the US courts and US patent office has been continuously broadened to the point that after a 1998 US court decision even software business methods could be patented [7].

Software patents have been extended to include the structure, sequence and organization of programs, as well as their graphics, sounds and appearance. Apple has been particularly aggressive in seeking and enforcing patent protection for the "look and feel" of its programs and products. Patent offices in other countries have not broadened their interpretations to the extent that the US has, but they have been dragged in the same direction, as US patents can be enforced in US courts for most global market activity.

Software patents are now considered to be the most important resource influencing NGN developments and NGN applications in vertical markets. They provide a wide moat of protection against potential competitors, and their proprietary standards lock-in customers as the cost of changing to a different supplier is prohibitive, providing monopoly power leverage for the initial provider for upgrades and extensions. The reaction to this over-reach of software patents has been the growth of an open standards and open source software community.

Some corporate, government agency and other organizations are employing open source software where possible. Some governments are taking steps to establish policies to ensure that open standards are required in all software applications purchased by government agencies. A notable example is the UK which announced in November 2012 an "Open Standards policy" establishing open standards principles for software interoperability, data and document formats in government IT specifications. [13]. For NGN applications, it is these IT software standards that must be compatible with NGN standards now in development.

5. CONCLUSION

The enhanced integration of NGN in the vertical markets of the application sectors will expand significantly the diversity of interests in standards development. They will raise new issues of changing boundary definitions and overlaps among the traditional domains of different SSOs, as well as the standards and IPR issues they address. Proprietary software IPR is expanding aggressively in the ICT sector threatening to narrow the scope and conditions for open standards development and implementation, and it dominates the IT applications field in the vast majority of vertical markets. Under present conditions, the intersection where open standards in the ICT sector meet open IT software standards in the vertical markets is a small and shrinking public space. The challenges are: a) stemming and reversing the deteriorating trend in the ICT sector; b) stimulating more rapid growth of the open standards and open source communities in IT software generally and in specific applications sectors; c) promoting synergy between the two open standards communities in taking pro-active initiatives to meet these challenges.

Some initiatives that might be considered include.

1) Software patent policies – The necessity for narrowing the criteria, scope and time period for software patent protection has become urgent and the supporting evidence substantial. SSO can present credible evidence and analysis to influence patent authorities and government policymakers.
2) Purchasing policies in the application sectors – Credible evidence and analysis can be presented to governments, industries and professional associations on the benefits of requiring open standards for NGN product and service purchases.
3) Public sector investment - Most countries have announced major policies, including significant public subsidies, for promoting broadband development as the NGN infrastructure. Open standards can be encouraged as a requirement for bidding for projects involving public funds. International agencies funding NGN applications in developing countries can adopt policies requiring open standards for funded projects.
4) ICT and application industry SSO can define specific network application technology areas where competition, sector efficiency and growth, and public interest considerations require open standards.
5) Education and awareness activities with industry-specific regulators and competition authorities on the implications of key standards and IPR issues for implementing their objectives. This could be extended to relevant international agencies such as WIPO and WTO as well.
6) Independent research support – On-going research is needed to strengthen the evidence and analysis supporting credible advocacy for open standards in the above initiatives. Academic research and public interest advocacy can make a major input.
7) ITU-T, working with national and international donor agencies, is uniquely positioned to ensure that developing countries receive maximum benefit from the above initiatives.

To restore a synergetic balance between proprietary interests in IPR and industry competitive and public interests in open standards, the open standards community will need to play a more active role in the policy debates that will define the boundaries of the open standards commons in the future economy. It must become a champion for defending and, where justified, expanding the domain of open standards in NGN applications in an increasingly contested policy environment. This has become the arena for formulating the NGN application standards that will have a major influence both on the "capacity to sustain" and the exclusionary barriers of future communities.

REFERENCES

[1] http://www.itu.int/en/ITUT/ipr/Pages/open.aspx (visited 2012-12-08.)

[2] Melody, W.H. (2011). **"**Liberalization in the telecom sector" in R.W. Künneke and M. Finger eds., International *Handbook of Network Industries: The Liberalization of Infrastructures***.** Edward Elgar, Cheltenham, UK, 103-122.

[3] Gordon, R. J. (2000). "Does the "New Economy" Measure up to the Great Inventions of the Past?", Working Paper No. 7833, New York: NBER.

[4] Melody, W.H. (2013). "Next steps in Europe's Digital Agenda", *info*, 15(2), March, *forthcoming.*

[5] Lemstra, W., J.P.M. Groenewegen and V. Hayes, eds. (2011). *The Innovation Journey of WiFi**:** The Road to Global Success.* Cambridge: Cambridge University Press.

[6] Varian, H.R. and C. Shapiro (1999) *Information Rules: A Strategic Guide to the Network Economy*, Boston, MA: Harvard Business School Press.

[7] Melody, W.H. (2007). "Markets and Policies in New Knowledge Economies", in R. Mansell et al. eds*., Oxford Handbook on Information and Communication Technologies,* Oxford: Oxford University Press, 55-74.

[8] Caplan, P. (2003). Patents and Open Standards, NISO White Paper , Bethesda, MD :NISO Press. http://www.niso.org/press/whitepapers/Patents_Caplan.pdf

[9] DeNardis, L. ed., (2011), Opening Standards: The Global Politics of Interoperability, Cambridge MA.: MIT Press.

[10] Duhigg, C. and S. Lohr (2012). "A System in Disarray: The Patent, Used as a Sword," Part 7: The iEconomy, NY Times 07 Oct.

[11] IFI Claims (2013). "2012 Top 50 Patent Assignees", IFI Claims Patent Services. *http://ificlaims.com/index.php?page=news&type=view&id=ifi-claims%2Fifi-claims-announces_2* (visited 2013-01-15).

[12] Brooks, R.G. and D. Geradin (2011). "Interpreting and Enforcing the Voluntary FRAND Commitment," *International Journal of IT Standards and Standardization Research*, 9(1), 1-23.

[13] *http://patent.gov.uk/about/consultations/conclusions.htm*. (visited on 2013-01-15).

INNOVATION MANAGEMENT OF ELECTRICAL VEHICLE CHARGING INFRASTRUCTURE STANDARDS IN THE SINO-EUROPEAN CONTEXT

Martina Gerst, Gao Xudong

Tsinghua University, School of Economics and Management

ABSTRACT

Energy challenges, changing consumer attitudes and evolving government mobility policies impact today's automotive industry. Mobility in sustainable communities of the 21st Century may to a considerable degree be based on New Electrical Vehicles (NEV) as an important part of electric mobility (e-mobility) concepts. One of the central factors to gain market acceptance is the interoperability of the different NEV sub-systems, particularly the standardization of the charging infrastructure. E-mobility is embedded in a rapidly changing, competitive and complex global environment, highly influenced by competing regional innovation policies. Therefore, this paper highlights some of the tensions in standardization management by Multi National Automotive Enterprises (MNAE) of a charging infrastructure in a Sino-European context.

Keywords— Standardization management, New Electric Vehicles (NEV), charging infrastructure

1. BACKGROUND

Energy challenges, changing consumer attitudes and evolving government mobility policies all impact today's automotive industry. Future mobility may be electrical. New Electrical Vehicles (NEV) as part of electric mobility (e-mobility) concepts will be an integral part of many communities' sustainable energy concepts in the 21st century. As the use of of fossil fuels for internal combustion is decreasing while prices are rising as a result of fuel shortage, e-mobility is gaining global importance particularly with regard to the building of sustainable mobile communities. The timely establishment of emission-free mobility not only stimulates long-term progress, it also allows consumers to retain the comfort to which they have become accustomed. [1]

For sustainable communities, addressing the topic of transportation and its impact on the environment is vital, considering that mobility is essential for economic growth [2]. On a global scale, the unprecedented growth in the global population has led to an expanding middle class, which is demanding increased mobility [3]. Due to this increasing demand, there is an urgent need to reduce greenhouse emissions. Therefore, NEVs can be one pillar to ensure sustainability in the transportation sector including this reduction and improving air quality [4].

The dissemination of e-mobility as an important pillar of sustainable communities is to a large extent based upon standardization of different interfaces, including vehicle engineering, energy supply, and the associated information and communication technologies of NEVs [5]. Standardization in the field of e-mobility is characterized by several features distinguishing it from previous standardization processes. Standards in the electrical engineering/energy technology and in the automotive technology domain have thus far been considered as separate entities [6]; there have only been very few attempts to look at them from a more integrated point of view.

Against this background, it is worthwhile to note that today standard setting in general has fundamentally changed from being a narrow 'technical' issue to a means of alignment of individual interests between the different players [7]. One of the main reasons is the fact that interoperability standards have become a significant factor in international trade [8] with considerable economic importance and serve as valuable enablers of innovation [9]. At a time when globally implemented technologies such as advanced Information and Communication Technologies (ICT) increasingly require compatible and harmonized standards to be fully effective, the role of standards is of increasing policy importance [8].

Actually, the standard setting practice is subject to very different challenges for all stakeholders. Firstly, standard setting is characterized by an on-going and increasing convergence of IT systems and their global implementations, coupled with, and further accelerated by, the Internet. Hence, standard setting processes are embedded in a rapidly changing and complex standardization environment, driven by the growing importance of ICT and the globalization of markets on the one hand [4] and the respective national innovation policies on the other. Secondly, the standardization environment is characterized by tensions triggered by different technology levels as well as by both new and well-established players with different sets of interests leading to different standardization dynamics [10]. Consequently, the outcomes of standards setting often remain uncertain because they are subject to competing arrays of interests including driving and opposing forces [10].

However, companies have understood that for a broad take-up of e-mobility standardization of different interfaces of an NEV is crucial. The challenge here is to continuously coordinate and integrate the diverse activities of stakeholders from different sectors in order to effectively

meet customer demands. An NEV compared to a traditional combustion engine car is a radical innovation that requires a new, cross-sector systems thinking.

In this context, one of the emerging questions is how standardization challenges in the field of NEVs can be addressed with a view towards speedy marketability (market acceptance and market access). NEV technology is less mature regarding the market than ICT and the technology trajectory may lead to different dynamics of standardization. This also includes the question how standardization management influences technological innovations. In particular, the role of stakeholders in different global regions, and their influence on the broader process of emerging technological innovations - e.g. the impact of international and regional standardization management on the development of charging infrastructure as one of the key components of a NEV – needs to be addressed.

The framework of analysis draws upon the Social Shaping of Technology (SST) perspective by explaining ICT innovations as history- and context-specific, actor-focused technological change processes [11]. The SST framework provides the perfect tool to explore the complexity associated with innovation development by incorporating a multifaceted socio-technical perspective.

This paper is based on the findings of two research projects in the area of standardization management funded by the European Union (EU) and the German Government, both recently finished. Besides introducing e-mobility standardization in Europe and China, the paper provides an insight into current approaches in standardization management of MNAEs within the European-Chinese context, with a particular focus on the charging infrastructure of an NEV.

2. E-MOBILITY STANDARDIZATION IN EUROPE AND CHINA

The strategies of NEV standardization and certification differ vastly between Europe and China. Without a general understanding of these main strategies it might be difficult to understand why the two systems so differently impact the standardization management of MNAEs.

2.1. The NEV Standardization strategy of Europe

The European strategy on clean and efficient vehicles has been adopted in May 2010 as part of the European response to the financial/economic crisis of 2008/2009. It provides a public policy framework to support the development of alternative technologies in the automotive sector with the main objective to stimulate the competitiveness of the European automobile industry and to promote sustainable mobility.

The European Strategy contributes to the European debate on measures to promote decarbonisation of transport and low emissions, but also takes the changing preferences of consumers (need for industry adaptation and innovation) into consideration. Green technologies play a central role in the sustainable development in Europe. The Strategy follows a two-track approach, assuring technology neutrality: One is the promotion of technologically advanced and fuel efficient vehicles to be put on the market in the near future, with a focus on the combustion engine (2020 perspective), increased use of sustainable bio-fuels, and gaseous fuels. The second is the European roadmap and the action plan for promoting and facilitating the emergence and proliferation of breakthrough technologies, mainly focusing on Electric Vehicles (plug-in hybrids and fully electric) and Hydrogen-powered vehicle [12].

Germany has been very active in e-mobility standardization and has launched a national e-mobility program in line with the European strategy, called “Nationale Plattform Elektromobilität” [13]. Although the market introduction of NEVs is understood as a challenge for Germany, it also offers market opportunities for energy technology, electrical engineering and automotive sectors where a high level of quality, safety and interoperability already have been achieved.

Figure 1 below shows how German regulation and standardization are integrated in the European and global regulatory and standardization framework. This system has proven to be very effective in ensuring that there are no differences between the national, European and international layer.

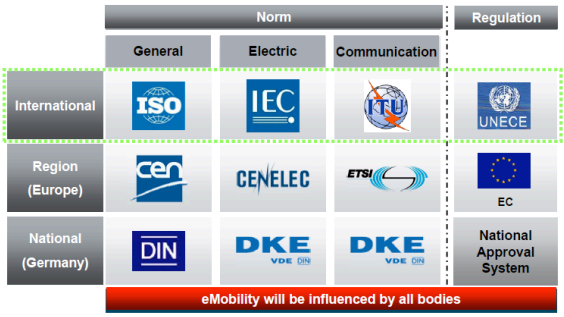


Figure 1: Standardization Organizations and Regulatory Bodies (Source: NPE – German Standardization Roadmap)

Stakeholders’ benefits in contributing to NEV standardization are manifold: First, standards have a pioneering role in preparing the market, particularly where user safety has to be ensured through testing methods and conformity assessment. Second, standards support innovation, thus securing to a certain extent investments made for development and research. Finally, standardization accelerates the further development of the Electric Vehicle sector by providing an enabling framework.

There are currently a number of pilot projects carried out in Germany within the EU’s e-mobility strategy. The main objective of these projects is to gain experience using existing standards and to identify where new standards are needed. Whereas some of these projects focus on standardization, for example how long-term programs may support innovations including standards, others investigate

ICT issues of key technologies used in NEVs in selected pilot regions.

2.2. The NEV Standardization strategy of China

According to the 12th Five-Year Plan, the goal of the Chinese standardization strategy is to establish a scientific, systematic, open, orderly and adjustable renewable energy vehicle standard system which fully meets research, industrialization, commercialization and management requirements and becomes an important technical support for the NEV industry. The idea behind this is to transform a large number of latest achievements and advanced experiences into standards and to get involved in associated international standard activities.

China's standards development attempts to transform itself from a follower to a standard setter. The technical route of standards development will be transformed from research to a combination of joint research and industrialization. The emphasis of the standards development will be on the coordination of a key framework of national standards with both standards developed for the industry sector and company specific standards.

Since the first standards for NEVs have been released in 2001, they have become the basis and the technical support for project application and evaluation, such as the State '863' program launched by the Ministry of Science and Technology (MOST) as well as technically supported NEV technology innovations. The "Renewable energy automotive manufacture and product access management" released by the Ministry of Industry and Information Technology (MIIT) in June 2009, stipulates that NEV must meet the existing conventional test items and specific standards. NEV standards play an import role not only in the NEV industry but also in manufacturing.

In total, 57 standards have been published since the Ninth Five-Years Plan (12 standards for Electric Vehicles, 8 for HEV (Hybrid Electrical Vehicle), 7 for FCEV (Full Cell Electrical Vehicle), 6 for e-motorcycles, 8 for energy storage, 5 for Electric motor and 11 for infrastructure). Currently, 7 standards have passed the examination waiting for approval, 19 new standards are under development and 45 standards are under preliminary research. Compared to the European approach to use and develop standards in an international regulatory and standardization framework, China develops a number of own national NEV standards, impacting directly MNAEs standardization management.

3. MNAE STANDARDIZATION MANAGEMENT IN CHINA

As outlined before, standardization is one of the central aspects for the market acceptance of NEVs. For Europe and China, and the respective industry and governmental organizations, the question will be to which extent the corporate stakeholders, for example MNAEs or power suppliers, are able to cooperate in standardization. From the perspective of an MNAE active in e-mobility standardization management, a number of requirements are important with regard to NEV standards management. Firstly, international corporations rather more aim at international standards than national companies do in order to save R&D cost, to focus on large scale production, and most notably to save market access time. Actually, national and international standardization concepts are competing, as outlined in chapter 2. Although standardization on a national level is much quicker, at least in China, it is considered by international corporations to be inadequate for global markets. Due to the different stakeholders and interests involved, collaboration and coordination amongst all relevant parties turns out to be a time-consuming and costly challenge.

Secondly, MNAEs favor the approach that existing standards have to be used and further developed. There are already a number of standards available in the automotive technology and electrical engineering sectors that could be used and further developed. However, in some countries the tendency to set national standards in order to gain an advantage in the national market may be observed. Nevertheless, some technical solutions need to be defined as interface standards to ensure interoperability (e.g. between vehicles and the network infrastructure). Due to the existing different regional standard setting approaches, companies often have to face not only the ambiguity of standards developed but also have to deal with some very strict local technical specifications. For example, the standardization of battery dimensions is not normally a topic for mandatory standards because the dimensions of batteries are usually different for each model and adequately adapted to the car size.

Thirdly, besides standard setting the conformity assessment is another important topic for MNAEs, particularly with a view on the market access of NEVs. In order to sell a new NEV model in a foreign market it has to go through a so-called homologation process, a government approval system proving conformity required for the registration of new products and services. Homologation is based on regulations and compulsory standards which may refer to standards and technical specifications as means to comply with the regulations. Homologation processes may include some or all of the following elements: approval of documentation, type testing, factory and production inspections, certification, registration and licensing, post-registration surveillance. Homologation rules may include a description of processes, applicable standards, and of the testing methodology to be used. On this account, it is interesting to analyze the standardization management of charging infrastructure in a Sino-European context. Charging infrastructure is a key component of a NEV and interoperability and through a high level of standardization will contribute to gain market acceptance of NEVs.

4. CHARGING INFRASTRUCTURE – STATE OF THE ART IN EUROPE AND CHINA

Regarding standardization, one highly relevant sub-system of e-mobility is the charging infrastructure. A recent study commissioned by the German Government analyzed the standardization 'Status Quo' of NEVs in general and with a

focus on Charging Infrastructure. The study revealed that there are still some gaps in infrastructure standardization management, in standardization and in conformity assessment, particularly with regard to homologation [14].

Thus, interoperability on the one hand between the individual components of NEVs and on the other hand with respect to the communication with infrastructure provided by various operators has to be ensured. The standardization of charging techniques and billing/payment systems has to be user-oriented, uniform, safe and easy-to-operate. Stakeholders involved, particularly the MNAEs, claim that a charging infrastructure has to be internationally standardized in order to succeed in the commercial auto market which is currently dominated by countries such as US, Japan, Germany or China.

Relevant standards cover connectors (e.g. plugs, socket-outlets, couplers, inlets), communication (e.g. communication protocols), safety (e.g. supply, batteries), and charging topology (e.g. charging station, conductive connection). The underlying idea commonly accepted is to provide safety for conductive charging of NEVs being enshrined in the respective IEC standards. Despite this basic understanding there are major differences in charging modes, in the charging infrastructure itself, and in communication systems. Nevertheless, there are now positive signs of a global convergence of standards.

In 2011, all European car manufacturers agreed on standards for charging modes and plugs for NEVs, related both to AC and DC infrastructure. This understanding is compounded by very advanced discussions with the US manufacturers on a common infrastructure, and supported by a dialogue with Japan on a common solution. The European agreement and the discussions with US and Japan shall lead to a unified set of standards fully implemented by 2017. There will be some backward compatibility for cars produced according to previous standards becoming defunct by the date.

However, this common agreement does not yet include the Chinese automotive industry: Whilst discussions have started, China is not part of the international understanding on preferred future charging modes. China developed its own standards for AC and DC connector types, which are based on international standards but contain significant modifications. As a consequence, this could pose a major safety risk for consumers, as it is possible to connect different electrical systems, one based on IEC standards, and one based on the Chinese compulsory standards.

China has been very active in NEV standardization for more than a decade and has developed a framework of national standards to support this emerging industry as shown in chapter 2. The Chinese standards for AC infrastructure seem to be fairly close to the future internationally agreed system; however, this is not the case for the currently applicable DC standards. It is unclear whether China plans to further harmonize the standards for plugs and system architecture/topology with the international system.

Regarding communication it is still too early to discuss a globally aligned system. In Europe, regarding plugs and communication, the objective is to achieve one standard between the charging station and the NEV. Therefore, the German-US Combined Charging System (CCS), a universal charging system for fast loading requiring only one vehicle interface enabling the customer to charge with all sorts of different available charging modes has been developed. European car manufacturers have agreed on a system based on power-line communication using internationally recognized standards such as “Home Plug Green PHY”, whilst the Chinese solution seems to be geared towards separate communication channels, though no final decision has yet been made. Both sides are aware that additional communication channels will be needed for a full integration of batteries of NEVs into future Smart Grids. Interestingly, regarding charging cables currently no generic European or Chinese standard is currently available.

Whilst these differences are relevant and render the infrastructures in Europe and China incompatible, there seems to be a common understanding of testing and homologation processes. At present, all testing is based on existing specifications for the “low voltage” area. Since charging currents generally exceed the limits set for “low voltage” it is understood that such requirements need to be adjusted to the specific needs of Electric Vehicles in the near future. Testing processes both in Europe and in China are a combination of various testing standards, mostly applicable for a power supply of 220/380V only. For power supply exceeding this value there seems to be no standardized testing process – leaving it to the individual testing laboratory to define specifications.

In Europe, market access for low voltage equipment is free, testing by a certified body generally suffices to comply with regulations; this also applies to the charging infrastructure. The situation for charging powers exceeding the low voltage area remains unclear – though current practice is to also leave such equipment firmly in the area of producer self-declaration. Applicable EN/IEC standards define the required testing. In China the charging infrastructure for NEVs is currently not part of the China Compulsory Certification System (CCC). Neither China nor Europe currently have specific homologation requirements for Charging Infrastructure in place: The basic principle is that this infrastructure must be safe for consumers, comply with low voltage regulations and ensure EMC compatibility. Hence, in terms of testing procedures, which are required for homologation of cars in the respective markets, there is still a gap to fill.

In sum, the elements to be standardized include charging poles, wall boxes (or home-chargers), cables and plugs. Currently, there is no existing global set of standards for the charging infrastructure for electric vehicles. Thus, current market access regulations are based on a combination of existing standards for vehicles in general with some specific EV requirements and reliance on automotive manufacturers’ own testing specifications. Since it is unclear which regulations and standards will be applied, the ground for any meaningful and comprehensive certification scheme is not yet in place. For an MNAE this means that if

it intends to sell an NEV in China, there is no type approval available with a specific focus on charging.

5. FUTURE PROSPECTS

A fast market access of NEVs is high on the agenda of MNAEs and governments. However, the speed and success of the intake of the new technology in both regions will depend on a variety of complex and intertwined factors. Amongst others, those factors include: the availability of a reliable and customer friendly charging infrastructure; the capacity of the automotive industry to tackle technical challenges such as cruising range, battery durability and reliability; trust in overall car safety; the ability of industry and governments to agree on a globally harmonized approach towards standards, common interfaces, vehicle topologies and architecture, homologation and certification processes. The uncertainty regarding the existing NEV standards world and the degree of global convergence of standards for infrastructure does not yet allow a substantial discussion on harmonization of testing and homologation processes which is a topic for all countries.

Regarding standardization and its management, one has to notice that standardization has gradually become a strategic instrument, particularly in new areas of technology such as the e-mobility field. For the stakeholders involved, particularly the MNAEs, the task to develop and implement e-mobility standards and to ensure their conformity on an international are challenging. Also, it introduces additional risks because automotive cycles, although already shortened, still require some planning. They have to deal with different standardization approaches comprising international and national standards, policies and regulations. Market access for NEVs seems to be driven by economic interests of the stakeholders rather than by interoperability.

Some of the NEV technology is still not very mature and in the area of charging huge infrastructural investments have to be made in the future. We may expect to see the emergence of new business relationships and business models offering value added services. New service configurations, e.g. in the battery field, require standards to ensure the necessary interoperability for re-charging. Further research will need to be carried out to analyze, for example, how stakeholders such as grid operators or manufacturers of the charging stations will influence NEV standardization management.

The example of the charging infrastructure shows that there is indeed a basic global understanding on some key elements – nevertheless the standardization work is yet to be completed. It is thus pre-mature to compare testing requirements or to develop certification schemes. Moreover, the fact that certification and approval requirements for infrastructure are still incomplete and consequently hamper the market entry of this new technology remains a major concern for many stakeholders. In particular, an agreed harmonized set of standards for compliance with infrastructure requirements is still missing.

Whilst a global alignment of standards for the charging Infrastructure of NEVs is on the way, this cannot be said for testing and homologation processes in Europe or China. It is thus imperative that cooperation on conformity assessment is taken seriously, and that such work has to be undertaken in close cooperation between governments and industry.

In the area of communication systems a consensus for global interoperability is still far away. The MNAEs are focusing on a system based on power-line communication (PLC) by using international standards, whilst the Chinese solution seems to be geared towards separate communication channels. These differences are relevant and make the infrastructure in Europe and China in compatible. Therefore, both MNAEs and Chinese car manufacturers are concerned that such testing rules will become a major obstacle for the commercialization of NEVs if not handled properly; current markets are largely segmented by different rules for infrastructure and related testing processes. Whilst integration of processes within Europe and within China is foreseeable, an alignment of testing processes in Europe and China would need a clear political will on both sides.

REFERENCES

[1] Niesing, B.:„Driven by electricity“, Fraunhofer magazine 1.10, 2010.

[2] SLoCAT@ Rio + 20: “Expanding the Use of Electric Mobility: Options for sustainable urban transport”, UN Department of Economic and Social Affairs (DESA), available at http://www.slocat.net/content/expanding-use-electric-mobility-options-sustainable-urban-transport (visited on 2012-11-28).

[3] ibid.

[4] Cox, W.: Reducing Greenhouse Gases from Personal Mobility: Opportunities and Possibilities, available at http://reason.org/files/reducing_greenhouse_gases_mobility_development.pdf (visited on 2012-11-28).

[5] Fortschrittsbericht der Nationalen Plattform Elektromobilität (Dritter Bericht), 2011.

[6] Federation Internationale de l’Automobile: TOWARDS E-MOBILITY THE CHALLENGES AHEAD, Brussels, 2011.

[7] Williams, R.: “Common ICT Standards or Divided Markets: the EU and China”, China-Eu Standards Project Workshop held in conjunction with the I-ESA Conference, 2009.

[8] Gibson, Christopher S., “Globalization and the Technology Standards Game: Balancing Concerns of Protectionism and Intellectual Property in International Standards.” *Berkeley Technology Law Journal*, Vol. 22, p. 1401; Suffolk University Law School Legal Studies Research Paper No. 07-39.

[9] Jakobs, K.: Blind, K.: “The European ICT Standardization Policy – Recent Developments, Current Situation, and a Brief Outlook.” *Proceedings of the 16th EURAS Conference. EURAS contributions to standardization research*, vol. 5, 2011.

[10] Jakobs, K.: "The European Union's ICT StandardizationPolicy - Changes Ahead!?" In: Jakobs, K. (ed): *New Applications in IT Standards: Developments and Progress*. IGI Global.

[11] R. Williams, D. Edge: "The Social Shaping of Technology: a Review of UK Research Concepts, Findings, Programmes and Centres", *New Technology at the Outset*, M. Dierkes, U. Hoffmann, (eds.), Campus-Verlag & Westview, Wiesbaden, 1992.

[12] EU Presentation, DG Enterprise, Automotive, November 2011, Beijing.

[13] Nationale Plattform Elektromobilität, the German Standardization Roadmap, Version 1.0.1, 2010.

[14] Gerst, M., Ziegler, K. (2012): Feasibility Study for BMWi (German Ministry of Economy and Technology) and CNCA: Sino-German cooperation on conformity assessment for New Energy Vehicles (NEV).

SESSION 7

ENERGY ISSUES

AN ANALYTICAL EVALUATION OF ENERGY CONSUMPTION IN COOPERATIVE COGNITIVE RADIO NETWORKS

Mahdi Pirmoradian, Olayinka Adigun, Christos Politis

Wireless Multimedia and Networking Research Group, Kingston University, London, UK
m.pirmoradian@iiau.ac.ir,{o.adigun, c.politis}@kingston.ac.uk

ABSTRACT

This paper studies the total energy consumption of a cooperative cognitive radio network in coexistence with a stochastic multi-channel licensed network. Energy consumption of each cognition phase at the secondary user is mathematically analyzed and obtained. The numerical results presented show the various interactions between the secondary network size, availability of appropriate spectrum holes and the total energy consumption for different states of the cognitive radio network under discussion.

Keywords— *Cognitive Radio, Energy Consumption, Energy Efficiency.*

1. INTRODUCTION

The volume of data transmitted over mobile networks will be significantly increased by about 800% over the next four years [1], also 7 trillion wireless devices will be serving 7 billion people by 2017, implying 1000 wireless devices per person [2]. This anticipated proliferation of wireless devices and multifarious networks require advanced and intelligent devices to overcome the challenges of spectrum scarcity, power consumption, interoperability and users' demands. Cognitive Radio technology is a well known solution to enhance better spectrum utilization and improved energy efficiency in order to satisfy future users' demands and contribute towards building a sustainable green wireless networks.

Cognitive radio (CR) is a promising paradigm in the next generation of wireless communication networks and is seen as a good enhancer of green communications in wireless networks in many perspectives including: better utilization of scarce spectrum band, leveraging cognitive radio's features to encourage improved energy efficiency in radio communications and making cognitive radio operations more energy efficient [3]. In cognitive radio networks, energy optimization must be considered as an essential key point in network and communication protocol architectures. The amount of energy which is consumed to deliver secondary data via spectrum holes need to be derived and optimized under desired QoS requirement for both the primary and secondary users. This could mean secondary data delivery will be carried out under limited transmission power levels, interference power constraint, precise spectrum hole lifetime and channel capacity requirement. In Green Cognitive Radio Network (GCRN), energy optimization is required during the cognition cycle, which includes: sensing, decision making and acting processes. The output decisions of Green Cognitive Engine (GCE) in a cognitive radio device are achieved through various energy efficiency optimization algorithms using the observed radio environment information and defined local regulatory policies.

Recently, many researches have focused on energy optimization in both wired and wireless telecommunications networks, which has to be achieved while still meeting the Quality of Service (QoS) requirement of various applications and at minimal cost [3], [4], [5]. Advanced low energy consumption integrated circuits have been a key consideration to meet the green criteria in wireless networks. Several techniques have been considered in the field of energy efficient networks such as; the exploitation and the control of power management capabilities (i.e., sleeping and rate adaptation) inside components of network equipment [6], [7], [8]. In this respect, at the physical layer level, advanced programmable hardware (i.e. ADC/DAC, Adaptive Filters/Antenna, and power-efficient Digital Signal Processing (DSP)) is the main goal in academic and industrial efforts. At the link layer, efficient transmission and error detection techniques will effectively decreases retransmission of data packets and eliminate the energy consumption associated with retransmission. Hence, appropriate routing with low-level transmission power significantly decreases energy consumption under desired quality of service level. In the previous works, energy consumption of the Secondary Users (SU) at different states (transmit, collision, idle modes) have not been studied comprehensively in cognitive radio networks.

This paper studies determining the average energy consumption of a secondary user at successful transmission, collision and idle states. The secondary network is allowed to use the licensed channels in an opportunistic manner. The analytical model could make cognitive radio operations more energy efficient. We obtain the average energy consumption at different cognitive functionalities such as observation, transmit, listen and collision period in a stochastic radio environment. This work is an initial study of an optimization technique on energy consumption of a secondary network with respect to the interference constraint at the primary receiver and desired QoS of the

system. The rest of the paper is structured as follows. Section II presents system model and proposed network architecture. Analytical model and energy consumption are obtained in section III. Section IV includes numerical analysis and discusses the achievements, and finally conclusion and future works are presented in section V.

2. SYSTEM MODEL

We consider an open licensed spectrum network consisting of several static wireless nodes communicating with each other using N licensed channels, while a multi-user cognitive radio network is located within the licensed coverage area. The coverage area of the secondary network is small compared to the distance between the secondary users and the primary transmitter such that the effect of primary signal at the secondary user can be ignored (see figure 1). The cognitive users are equipped with spectrum sensors, which periodically perform sensing and report channel states information to a Central Node (CN) via dedicated common control channels. The CN assigns appropriate vacant channel to the secondary users located within the secondary network based on the received channel state statistics.

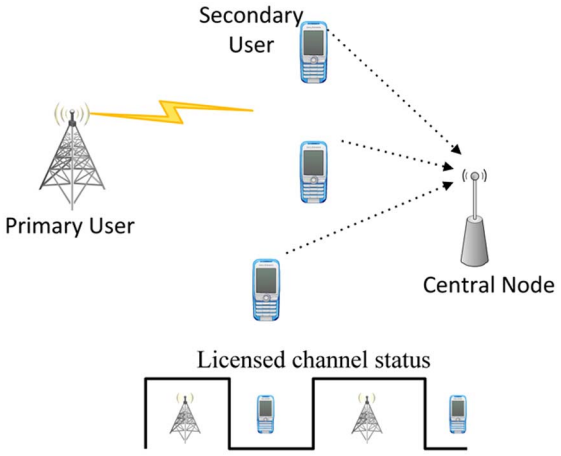


Figure 1. Network topology; secondary users report sensing information to the central node (decision centre)

2.1. Primary User Activity Model

Let's assume that the licensed bands utilization can be modeled as a Poisson process. The channel model can be estimated as a birth death process with different transit probability. The duration between two utilization periods (inter-arrival rate of the PU) are identical independent distribution (i.i.d) random variables following an exponential distribution with constant busy and idle periods during system analysis. Let α denote the probability that the channel transits from state ON (Primary user is active) to state OFF (Primary user is inactive) and β denote the probability that the channel transits from OFF state to ON state. Therefore, the channel availability can de denoted as $P_{OFF} = \frac{\beta}{\beta+\alpha}$, and the probability of occupied channel can be obtained as $P_{ON} = \frac{\alpha}{\beta+\alpha}$.

2.2. Cognitive Cycle and Power Consumption

The secondary user activity cycle is shown in figure 2. The figure shows different power level at different phases (i.e. sensing, reporting, decision, data delivery/transmission, acknowledgement and idle). The secondary user uses receiving power level during the sensing, decision and acknowledge processes, and uses the transmit power level at the reporting and data delivery/transmission processes. Users may be switched on idle mode when there are no vacant channels for data transmission.

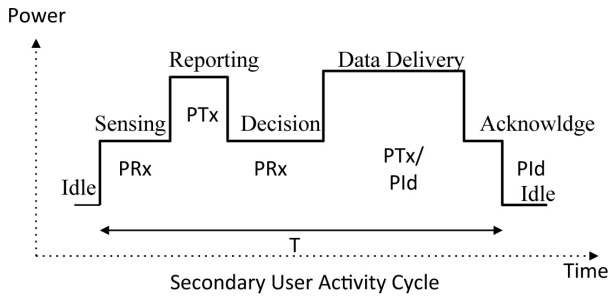
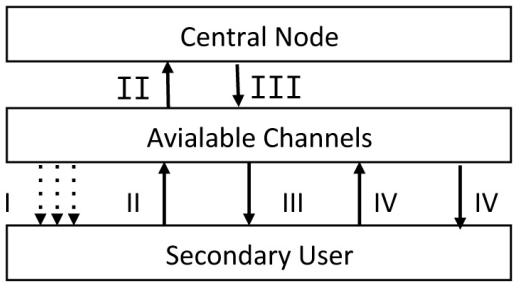


Figure 2a. Secondary user activity and power levels at sensing, reporting and idle states



I: Sensing Avaiable Channels
II: Sending channels' status to the CN
III: Assign an appropriate channel
IV: Secondary transmission

Figure 2b. Secondary network workflow

3. ANALYTICAL MODEL

3.1. Spectrum Sensing and Cooperation Mechanism

Each cognitive device performs spectrum sensing independently and the outcome decisions are sent to the CN. According to the collected outcome decisions received at the CN, an algorithm can be employed to utilize appropriate unoccupied licensed channels (spectrum hole) in a cooperative manner. Assuming licensed channels are identical and independent, the received signals at the sensor

node i during sensing period of time slot (n) can be given by;

$$Y_i(n) = \begin{cases} \omega_i(n) \\ h_i(n)S_i(n) + \omega_i(n) \end{cases} \quad (1)$$

where $\omega_i(n), h_i(n)$ and $S_i(n)$ are independent parameters and represent the noise, channel gain between Primary User (PU) and SU, and primary signal at the spectrum sensor node. The noise is assumed to be an independent identically distributed (i.i.d) random variable with zero mean and variance σ_ω^2. Also, the primary signal is assumed to be a random variable with zero mean and variance σ_s^2. It is assumed that each channel gain $|h_i(n)|$ is Rayleigh-distributed with same variance σ_h^2, while the hypothesis of $\mathcal{H}_0$ and $\mathcal{H}_1$ denotes the absence and presence of primary users. Let's assume each sensor node employed energy detector and measured their received power during the sensing period. Let δ_i be the threshold parameter of the detector on the sensor i. Thus, the probability of detection (P_d^i) and probability of false alarm $(P_f^i)$of the detector can be approximated as;

$$P_d^i = Q\left(\frac{\delta^i - \tau f_s(|h|^2\sigma_s^2 + \sigma_\omega^2)}{\sqrt{2\tau f_s}(|h|^2\sigma_s^2 + \sigma_\omega^2)}\right) \quad (2)$$

$$P_f^i = Q\left(\frac{\delta^i - \tau f_s\sigma_\omega^2}{\sqrt{2\tau f_s}\sigma_\omega^2}\right) \quad (3)$$

where $Q(.)$ denotes the right-tail probability of a normalized Gaussian distribution $Q(x) = \frac{1}{\sqrt{2\pi}}\int_x^\infty e^{\frac{-t^2}{2}}dt$, f_s is the sampling frequency and τ represents sensing time slot. The equations above show that the probability of detection and probability of false alarm related to the detection threshold, sensing time and sampling frequency. Considering that each sensor observes the licensed channels and send their local decisions to the CN to make final decisions and proper coordination of all nodes and licensed channels around it, assuming that all decisions are independent, then the probability of detection and probability of false alarm of the SUs network under OR-rule, $\mathbb{P}_d$ and $\mathbb{P}_f$ respectively can be mathematically written as [9];

$$\mathbb{P}_d = 1 - \prod_{i=1}^{N}\left(1 - P_d^i\right) \quad (4)$$

$$\mathbb{P}_f = 1 - \prod_{i=1}^{N}\left(1 - P_f^i\right) \quad (5)$$

In the case of a homogeneous network, $\mathbb{P}_d$ and $\mathbb{P}_f$ can be simplified by$\mathbb{P}_d = 1 - (1 - P_d)^N$ and $\mathbb{P}_f = 1 - \left(1 - P_f\right)^N$. Where N represents number of secondary users participating in the sensing task.

3.2. Energy Consumption Analysis

The aim is to estimate and obtain the energy consumed by secondary users at different states. The power levels of SU are adjusted according to the Quality of Service (QoS) requirement and user's states. We estimate the energy consumption of SU at transmission, collision and idle states, which are explained in figure 3 and table I. Network notations are explained in table II. The network statuses are described in table II, and are at the instance of the sensing pulse time and the decisions taken are reflected on this.

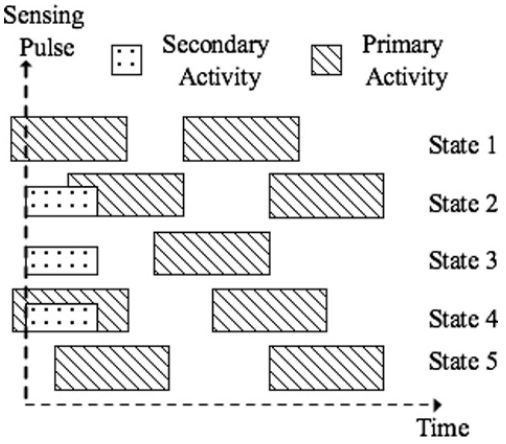


Figure 3. Primary user activity and various states at the SU side.

Table 1. Network Status

States	Primary User	Sensor Outcomes	Transmission State
1	Presence	Detection	No Transmission (Idle Mode)
2	Absence	Detection	Collision
3	Absence	Detection	Success
4	Presence	False Detection	Collision
5	Absence	False Detection	No Transmission (Idle Mode)

Table 2. Network Notation

Parameter	Description
P_{Tx}	Power consumption at transmit mode
P_{Rx}	Power consumption at receive mode
P_{Idl}	Power consumption at idle mode
W	Bandwidth
τ	Sensing time
M	Number of available channel
N	Number of SU
σ_w^2	AWGN variance
δ^i	Threshold value for sensor i
L	Frame duration of SU
$t_{ack.}$	Duration of acknowledge time
$\mathcal{T}^i$	Throughput of secondary user i
R	Channel rate of secondary user i

The probability of channel collision in the selected idle channel i is $P_{Coll}^i = 1 - \int_L^\infty \beta_i e^{-\beta_i t}\,dt = 1 - e^{-\beta_i L}$ [10]. Also, the probability of appropriate channels being

available (state 3) by employing a cooperative mechanism using (5) can be given as;

$$P_{av} = P(H_0)\left(1 - \mathbb{P}_f\right).\exp(-\beta.L) \quad (6)$$

Therefore, the average energy consumption of the secondary user at transmission, collision and idle states can be approximated as;

$$\mathcal{E}^{Trans.} = P(H_0)\left(1 - \mathbb{P}_f\right).\exp(-\beta.L).\mathrm{L}.P_{Tx} \quad (7)$$

$$\mathcal{E}^{Coll.} = \left[P(H_0)\left(1 - \mathbb{P}_f\right).(1 - exp(-\beta.L)).P(H_1).L + P(H_1).(1 - \mathbb{P}_d).[exp(-\alpha L).L + (1 - exp(-\alpha L))P(H_1).L]\right].P_{Tx} \quad (8)$$

$$\mathcal{E}^{Idle} = P(H_0)(1 - \mathbb{P}_d).\mathrm{L}.P_{idle} + P(H_1).\left(1 - \mathbb{P}_f\right).P_{idle}.\mathrm{L} \quad (9)$$

where $P(H_0)$ and $P(H_1)$ are the probabilities of the primary user being absent (P_{OFF}) and present (P_{ON}) in the channel respectively.

To obtain the energy consumption in a multi-channel licensed network, let v denote the number of available channels which are used by the secondary users, therefore, the probability distribution function of the random variable (v) can be obtained by $\binom{M}{v}(1 - P_{av})^{M-v}{P_{av}}^{v}$. The average probability of number of the channels can be written as;

$$CDF_{av} = \sum_{v=1}^{M}\binom{M}{v}(1 - P_{av})^{M-v}{P_{av}}^{v} \quad (10)$$

where M denotes the number of licensed channel. Hence, the average energy consumption at the successful transmission state of the network can be obtained by;

$$\mathcal{E}_{Trans.}^{Total} = \sum_{v=1}^{M} v\binom{M}{v}(1 - P_{av})^{M-v}{P_{av}}^{v}.\mathcal{E}^{Trans.} \quad (11)$$

Also, the average energy consumption of secondary users at idle and collision states are given by;

$$\mathcal{E}_{Coll+Idle}^{Total} = \left(N - \sum_{v=1}^{M} v\binom{M}{v}(1 - P_{av})^{M-v}{P_{av}}^{v}\right).P_{Tx}.L \quad (12)$$

Considering the secondary throughput as the successful secondary data delivery over unoccupied channels in the CRN, the average throughput of a secondary user i can be mathematically written as;

$$\mathcal{T}^{i} = \sum_{v=1}^{M}\binom{M}{v}(1 - P_{av})^{M-v}{P_{av}}^{v}.\left(\frac{L - t_{ack.}}{L}\right).R^{i} \quad (13)$$

where R denotes user's data rate.

4. NUMERICAL ANALYSIS

This section shows the evaluation of the total energy consumption of the proposed cognitive network within a homogeneous network. The proposed network is simulated in Matlab environment. In this simulation environment, the employed parameters and their values are given in table III (all values have been taken from 802.11 standard) [11]. The secondary delivery time is assumed to be constant for all secondary users and the threshold value δ is determined by equation (2).

Figure 4 shows the probability of detection of licensed signal versus received *SNR* of the PU signal at secondary node.

Table 3. Analytical Parameters

Symbol	Definition	Quantity
N_0	Noise power density	10^{-15} W/Hz
P_{Idl}	Power consumption at idle mode	0.8W
P_{Rx}	Power consumption at receive mode	1W
P_{Tx}	Power consumption at transmit mode	1.5W
$t_{ack.}$	Duration of acknowledge time	1 ms
L	Frame duration of SU	50 ms
M	Number of available channel	12
N	Number of SU	2, 5, 8, 10, 15, 20, 30, 40
R	Channel rate of SU	1Mbs
W	Channel bandwidth	1 MHz
α	Channel transit (ON to OFF)	0.4
β	Channel transit (OFF to ON)	0.3
P_f	Probability of false alarm	0.01
τ	Sensing time	10 ms

The probability of detection is assumed to be 0.8 given *SNR* equal to -7.4 dB. This shows that the sensors are able to detect signals with *SNR* greater than -7.4 dB with respect to the given bandwidth, sensing time and probability of false alarm detection as given in Table III.

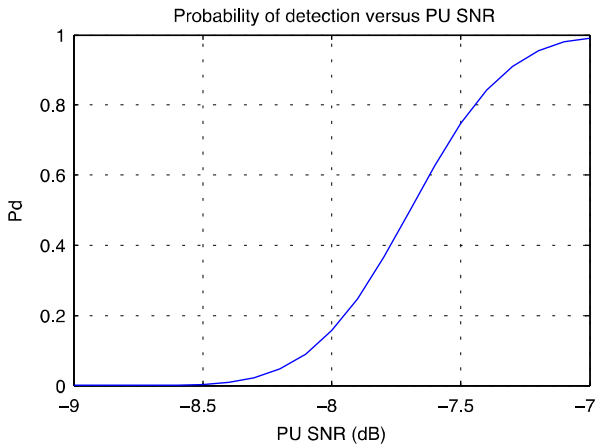


Figure 4. Probability of PU signal detection versus received SNR at secondary user.

Figure 5 shows the cumulative distribution function (CDF) of the number of spectrum holes (unoccupied channels) in

the proposed system supporting 12 licensed channels. It shows the various CDF functions versus number of SUs (from 2 to 40) and licensed channels. The figure shows that the increasing number of SU causes a decline in the available spectrum holes. Inevitably, the CN's decision is also affected by false alarm signal and number of SUs in cooperative mechanism.

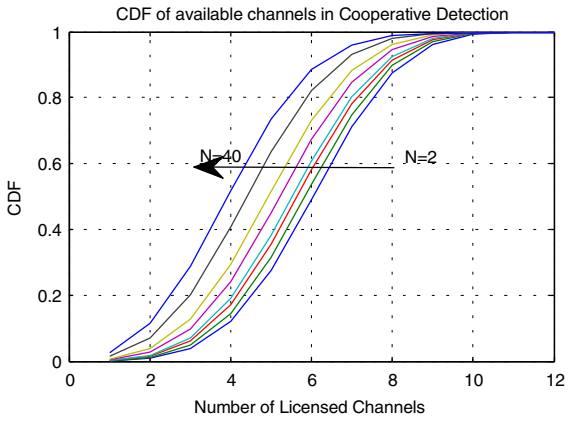


Figure 5. Cumulative distribution function of appropriate spectrum holes versus number of licensed channels and SUs.

The total energy consumption at the success transmission phase is shown in figure 6. The outcomes reveal that the total successful secondary data delivery gradually decreases due to the secondary network size. Obviously, the consumed energy at transmission state falls (e.g. 530mJ for two SU, and 250mJ for 40 SUs).

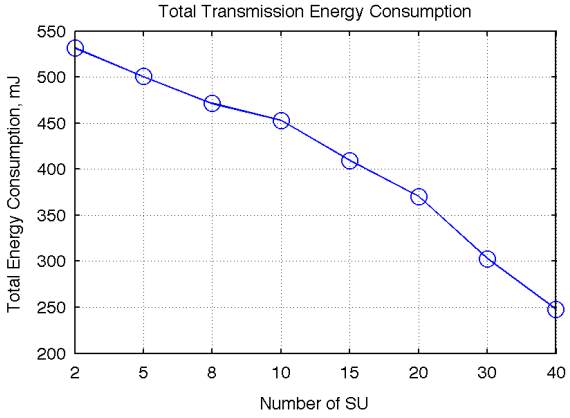


Figure 6. Total energy consumption at success transmission state versus number of SU

Total energy consumption at idle and collision states are shown in figure 7. The results reveal the increase in energy consumption as the number of secondary users in the proposed system increases.

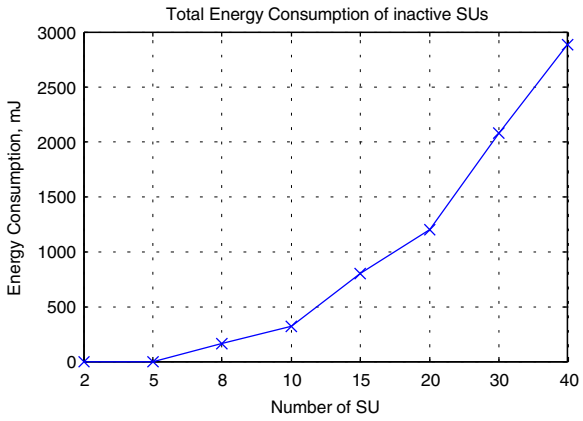


Figure 7. Total energy consumption (at collision and idle states) versus secondary network size.

Figure 8 depicts the SU's channel rate versus secondary network size. It can be seen that the SU's channel data rate is significantly dropped as the cognitive network size increases. This is influenced by the fact that the probability of false alarm is getting higher within the cooperative sensing mechanism as the network size increases, which is detrimental to the average channel rate.

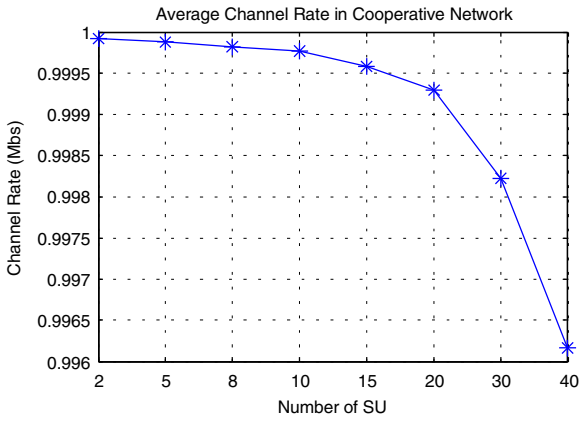


Figure 8. Average data rate versus cognitive network size

5. CONCLUSION

This paper has presented an analysis of the average energy consumption of cognitive radio network at transmission states (i.e. success, collision, idle modes). The numerical results proved that number of licensed channels, secondary users and licensed channel characteristics significantly affect energy consumption in CRN. In addition, the analysis of the average channel rate of the CRN versus network size revealed channel rate decreases as network size increases. Further studies to the work presented in this paper will aim to optimize energy consumption of the CRN with respect to the primary network's characteristics (such as bandwidth, sensing time and received SNR), required QoS of cognitive radio network, network size and power consumption constraint. Also the power consumption optimization in different states may be considered in architecture of CRN designing.

REFERENCES

[1] "Wireless Operations in the 3650-3700 MHz Band," Federal Communications Commission, Report 2007.

[2] D. Klaus, D. Sudhir, J. Nigel, "2020 Vision," *IEEE Vehicular magazine*, pp. 22-29, September 2010

[3] G.Gürkan, A. Fatih, "Green Wireless Communications via Cognitive Dimension: An Overview," *IEEE network*, vol. 25, no. 2, pp. 50-56, March-April 2011.

[4] N. Sergiu, P. Lucian, I. Gianluca, R. Sylvia, W. David, "Reducing Network Energy Consumption via Sleeping and Rate-Adaptation," in *NSDI'08: Proceedings of the 5th USENIX Symposium on Networked Systems Design and Implementation*, Berkeley, pp. 323-336, 2008.

[5] G. David, C. Jingxin, J. Tao, D.M. Paul, "Using Cognitive Radio to Deliver Green' Communications," in *4th International Conference on Crowncom*, 2009.

[6] C.O. Mert, B.A. Ozgur, "Energy-Efficient Packet Size Optimization for Cognitive Radio Sensor Networks," *IEEE Transaction on wireless communications*, vol. 11, no. 4, pp. 1544-1553, April 2012.

[7] L. Won-Yeol, A. Ian.F., "Optimal Spectrum Sensing Framework for Cognitive Radio Networks," *IEEE Transaction on wireless communications*, vol. 7, no. 10, pp. 3845-3857, October 2008.

[8] L. Yi, X. Shengli, Z. Yan, Y. Rong, C.M.L. Victor, "Energy-Effecient Spectrum Discovery for Cognitive Radio Green Networks," *Mobile Network Application*, vol. 17, pp. 64-74, March 2012.

[9] G. Amir, S.S. Elvino, "Opportunistic Spectrum Access in Fading Channels Through Collobrative Sensing," *Journal of Communication*, vol. 2, no. 2, pp. 71-82, May 2007.

[10] P. Mahdi, A. Olayinka, P. Christos, "Adaptive Power Control Scheme for Energy Efficient Cognitive Radio Networks," in *ICCC2012*, Otawa, 2012.

[11] "Power Consumption and Energy Efficiency Comparisons of WLAN Products," Atheros Communications, White paper 2003.

SOLAR-POWERED CELL PHONE ACCESS POINT FOR CELL PHONE USERS IN EMERGING REGIONS

Takuya Kato and Yoshihiro Kawahara,[†]

Graduate School of Information Science and Technology, The University of Tokyo, Tokyo, Japan
[†]School of Electrical and Computer Engineering, Georgia Institute of Technology, Georgia, USA
{kato, kawahara}@akg.t.u-tokyo.ac.jp

ABSTRACT

The availability of electrical power is a critical issue for building sustainable communication networks in emerging regions. Low-cost energy delivery encourages people to use communication services. This paper presents simulation results on the effect of the distribution of the surplus power generated for an access point (AP) to cell phone users. We show that 9.3% of the user population can use the excess power generated for the AP. Furthermore, we propose an energy-proportional server cluster to ensure computational resources for information services, such as for charging cell phones. The existing server hardware often wastes power in the idle state and is not energy proportional, and thus we designed the cluster to reduce this energy waste by matching the number of working servers to the number of incoming requests. Our prototype system with low-power and off-the-shelf devices cuts energy consumption in the frequently observed idle state by 50% compared with an existing server cluster with equivalent performance.

Keywords— Emerging Regions, Photovoltaic System, Energy Proportional

1. INTRODUCTION

For provision of sustainable information communication technology (ICT) services in emerging regions, economic obstacles are usually inevitable. A particularly serious problem is the lack of infrastructure, such as power plants, electrical grids, and landline connections. Because the cost of setting up solid infrastructure bears no relation to the expected returns for communication service providers, other inexpensive ways of establishing networks have to be found. Attention is currently focused on solar-powered wireless networks. For instance, efficient photovoltaic stations have been promoted to supply electrical power for telecom networks instead of diesel generators and fossil fuels, which require estimation of the maximum oil consumption and the impact on prices and the environment in terms of CO_2 emissions and noise reduction [1]. These limitations do not apply to renewable energy, which can be produced and consumed in the local area and can expand ICT services into emerging countries.

In contrast, from the viewpoint of information terminals, cell phones are mainly used for voice calls and SMSs in such regions [2]. The relatively low deployment cost and intuitive operations of cell phones are the reasons for their rapid penetration in areas that have low literacy rates and where landline connections are expensive. Not all cell phone users, however, can afford to have a personal photovoltaic system for charging their handset. Instead, they usually pay for this service at the nearest phone shop owned by a local entrepreneur.

Unfortunately, these charging expenses often have a negative impact on their cell phone usage. In Uganda, cellular subscribers spend an average of $2.25 per month on the charging process, which accounts for 10–50% of their monthly cell phone expenditure. Moreover, people living in remote areas spend up to $25 per month on transport to the nearest village with a charging spot. According to an interview, three out of four users answered that they would spend much more money on airtime if they could reduce their charging expenditure [3]. The loss of cellular airtime caused by high charging expenses also affects communication service providers in the form of revenue loss. Thus, service providers have an incentive to solve this problem.

In this paper, we present a feasibility study on one solution to the problem of cell phone charging: the distribution of electrical power generated at APs to cell phone users living around the AP. The additional cost of photovoltaic systems for cell phone charging might result in adverse balance, and the excessive provision of electrical power might cause a power shortage at the AP. Therefore, we focus on the expected surplus power—the power generated when the AP batteries are full and dumped wastefully—obtained when a PV system (solar modules and batteries) with the minimum capacity that can maintain the AP is installed. We use this base PV configuration in order to estimate the percentage of the user population that can be served without the additional expense of increasing the PV capacity.

Furthermore, to improve the convenience of charging and the quality of life of the local people by providing relevant local information and educational content, we suggest installing an energy-proportional low-power server cluster at the AP. The second and subsequent servers will handle user requests

Thanks to France Télécom S.A. for funding

in an overload situation in order not to disturb the signal-processing tasks at the AP. However, this wastes precious electrical power in low-load states because, when in the idle state, the existing server hardware is optimized to maximize the peak throughput and is thus not designed to save energy. Hence, we propose introducing the energy-proportional concept—the idea that the power consumption of a system should be proportional to its workload—to the server cluster by matching the number of active servers to the load. We also present the design and implementation of a prototype of the server cluster and evaluate it by examining power usage and response times to some kinds of traffic patterns. As a result, our scheme reduces power consumption effectively under low load, succeeding in saving over 50% energy compared to the conventional management policy where all servers are kept in service whatever the workload, without affecting the response time.

2. ANALYSIS OF AP SURPLUS POWER FOR CELL PHONE CHARGING

2.1. Wireless Networks in Rural Areas

A well-known example of a rural network is the Village Phone Program in Bangladesh, which is an initiative of the Grameen Bank [4]. The Grameen Bank elects a rural resident who satisfies the bank's criteria as a village phone operator (VPO). A VPO uses a loan to purchase a start-up kit, including a handset, an antenna, a solar charger, etc., for between $50 and $300 and earns revenue by providing call services to other people. This business style not only expands revenue opportunities but also extends communication services. Another approach to deploying rural wireless networks is the OpenBTS project [5]. OpenBTS is an open-source software application that works like a GSM base transceiver station (BTS). It can offer a GSM air interface to standard GSM compatible cell phones by using software radio and deliver calls with SIP soft switches or a PBX. Thus, it can deploy a cellular network at a considerably lower cost in rural areas with software-defined radio hardware, including Universal Software Radio Peripheral (USRP) costing several hundred dollars, antennas, and a host computer for signal processing and PBX software.

Based on these approaches, we assume a solar-powered GSM network system as shown in Figure 1. A rural GSM AP that adopts the design concept of OpenBTS is equipped with a photovoltaic system that will be supplied electrical power in an off-grid environment. Signal processing with GNU Radio and PBX software such as Asterisk are implemented on USRP hardware and a server. Because it is expected that the AP will be deployed sparsely in a thinly populated area, the AP has two antennas, one of which is used for communication with the phones in its cell and the other for a distant GSM BTS. According to the specification of a USRP device, the wave range is up to 35 km at 50 W power consumption [6]. Thus, it can provide communication service at a pinpoint location where residents reside within a 35 km radius of a GSM BTS, which can reduce the number of APs to be deployed.

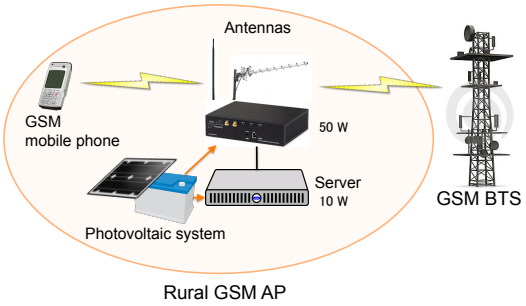


Figure 1. Solar-powered GSM network.

2.2. Surplus Power of AP

A GSM AP has a photovoltaic system that will be powered in an off-grid environment as shown in Figure 1. Most cell phone users cannot afford having their own photovoltaic system. They are forced to purchase electrical power from a phone shop, which is relatively expensive and thus reduces their usage of communication services, as described in Section 1. From the standpoint of the service providers, one feasible solution is to provide a part of the electrical power generated at the AP to their customers who reside near the AP. However, the additional cost of a photovoltaic system for the cell phone charging service might result in an adverse balance, and excessive provision of electrical power might cause a power shortage at the AP. Therefore, it is important to generate the expected surplus power with the minimum capacity required by the PV system (solar modules and batteries), so that the percentage of the population benefited by the charging service can be estimated without factoring in the additional expense of a larger PV capacity. Here, we define surplus power as the power generated when the AP batteries are full and are dumped wastefully.

To estimate the surplus power, first we determined the required capacity of the photovoltaic system of the AP. We calculated it by using the annual solar radiation data in Kampala in 2005 [7]. Because Kampala is located near the equator, we chose horizontal irradiation data. Moreover, we assume the power consumption of the AP is constant at 60 W, derived from the sum of 50 W of USRP hardware and 10 W of the server; the charge efficiency of the batteries is 85%; the temperature loss is 9.7%; and other losses are 5% [8]. Based on these data, we computed the annual AP system availability as shown in Figure 2. Allowing up to one day of system unavailability in a year, we can find the candidate minimum required capacity of solar modules and batteries to be (350 W, 157 Ah) or (400 W, 92 Ah). Here, because the price of the solar modules is $2.42/W and that of the batteries is $2.29/Ah, the latter is inexpensive and should be selected [9].

Next, we computed changes of surplus power in a day generated with the photovoltaic system (400 W, 92 Ah) as shown

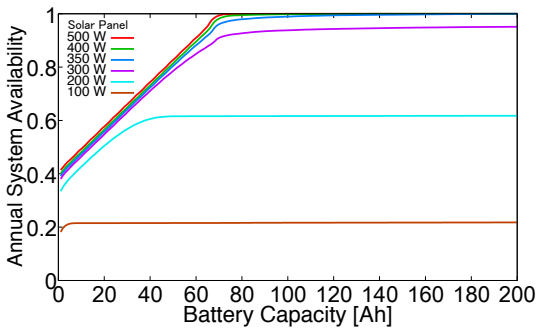


Figure 2. Relationship between photovoltaic system capacity and system availability.

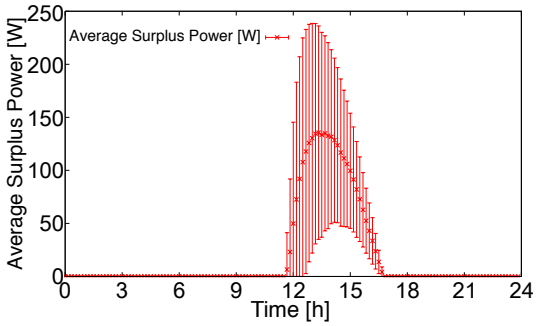


Figure 3. Fluctuation of surplus power in a day.

in Figure 3. These results suggest that the opportunity to charge in a day is limited, about 5.5 h on average. Additionally, large variance means daily fluctuation of charging opportunity. To provide further insight into the fluctuation, we plotted the annual data of surplus power in a day in Figure 4. The average surplus power slightly increases in February and March, when sunlight shines almost vertically down on the solar cells, but it is clear that the daily surplus power changes roughly randomly throughout the year. The yearly average surplus power in a day is 449 Wh, and the sample standard deviation is 278 Wh.

Following this line of thought, we came up with a speculative figure for the power consumed by cell phone charging for comparison with the AP surplus power, as in Equation (1):

$$E_{phone} = S_{AP} P_{own} P_{share} C_{phone} F \quad \text{[Wh/day/(people/km}^2\text{)]} \tag{1}$$

where E_{phone} [Wh/day/(people/km^2)] is the power consumption for cell phone charging per day per unit population density, S_{AP} [km^2] is the coverage area of the AP, P_{own} is the cell phone ownership, P_{share} is the market share of the communication service provider that installed the AP, C_{phone} [Wh] is the capacity of a phone battery, and F [/day] is the charging frequency per person per day. Each parameter is determined as follows. First, assuming the radio wave range of cell phones is 4 km, $S_{AP} = 50.3$. The ownership is derived from an interview in Uganda that says all respondents own at least one cell phone. This report also states that 35% of the interviewees own two or more phones, while 31% share their phones with other people. Thus, we set the average ownership $P_{own} = 1$. We set $P_{share} = 44\%$ from the largest market share in the GSMA report [10]. he battery capacity data is established based on Nokia 103 because Nokia is the dominant brand of cell phones used in Uganda [11]. According to the battery specification, the voltage is 3.3 V and capacity is 0.8 Ah, which implies $C_{phone} = 2.96$ [12]. The final parameter is calculated as $F = 0.375$, derived from the average monthly charging expenses of $2.25 and the average price per charge of $0.20 [3]. Consequently, $E_{phone} = 24.5$. Considering that the average population density in Uganda is about 167 people/km^2, the average power consumption is 4092 Wh/day. Assuming that the charge efficiency of lithium ion batteries of cell phones is 0.85, this means that about 9.3% ($= 449/4092 \times 0.85$) of customers on average can be supplied power by using the AP surplus power.

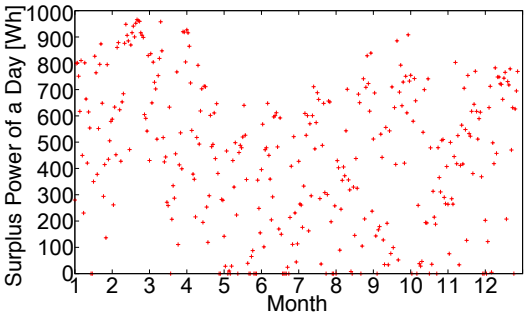


Figure 4. Distribution of surplus power in a day.

Finally, we should note that the available power varies with the cost of the photovoltaic system. It is possible that $1.0/W solar modules will become available due to a marked downward trend in their price. In fact, the minimum required photovoltaic system for APs is still (400 W, 92 Ah) even with the cheapest solar modules. However, service providers can add more photovoltaic systems at the current price. Because the total cost of the minimum required photovoltaic system becomes approximately half with $1.0/W solar modules, they can provide the same scale of energy supply to customers as the AP, which feeds 60 W constant energy to more 35.2% ($= 60 \times 24/4092$) customers. It is interesting to consider the return on investment of additional photovoltaic systems, but that is outside the scope of this paper.

3. ENERGY-PROPORTIONAL SERVER CLUSTER FOR WEB-BASED INFORMATION SERVICE

3.1. Significance of Energy-Proportional Server Clusters

We found that there is surely usable power generated at the AP. The amount of the power, however, is very limited with the minimum configuration and fluctuates greatly from day to day. One of the potential problems in offering the cell

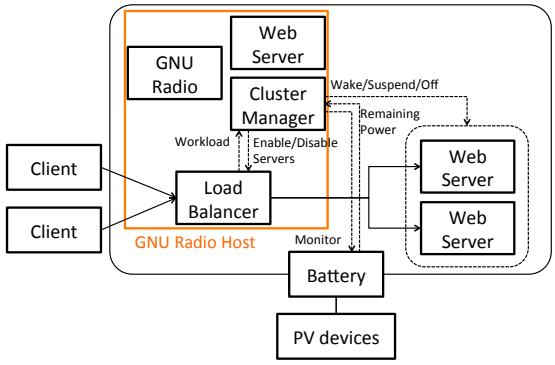


Figure 5. System architecture.

phone charging service is the inequality of charging opportunity. While people in the AP's neighborhood can check the battery state and charge their phones as soon as surplus power is generated, those who live far from the AP have to travel tens of minutes to charge their cell phone every time, without any guarantee of access to surplus power. It is easy to imagine the state of affairs: a lot of people who travel long distances are unable to charge their cell phones.

A solution to this inconvenience of the charging service is to provide a web-based information service. A web-based infrastructure that enables sharing of multimedia content through a graphical user interface is important where literacy rate is low. For example, battery-level notifications and weather forecasts should inform distant people about opportunities to charge their cell phones. It is better if a reservation service is offered because they can start out without the anxiety of a possible wasted trip. Furthermore, some educational content, such as training literature, might contribute to improving the quality of life of the residents. In this way, a web-based information service is valuable, and computational resources to provide it are needed.

In this case, utilization of cloud resources is not the best answer because it negates the advantage OpenBTS has, namely, the ability to construct local networks even in remote areas with poor Internet connectivity. That is why we focus on preparing a low-power server cluster at the local AP. For web services, additional servers are put into service for handling the overload, which is expected to occur around noon. However, unfortunately, it is known that existing servers do not save much power even in the idle state. For instance, even an energy-efficient server still consumes more than 50% of the peak power usage when idle, and commodity products often consume over 80% [13]. Thus, we introduce energy-proportional management to cut the wasted energy in the low-load state.

3.2. System Requirements

We described the significance of web-based content to the local people and computational resources needed to provide them in the previous section. In this section, we consider the features that must be built in when designing the system.

The first requirement is that low-power and inexpensive devices must be used. It is important to save limited electrical power and to implement the system at a low cost. The expected power consumption of a server is less than 10 W, as shown in Figure 1.

The second requirement is that the system must be autonomous and scalable, starting from a very small village. In order not to waste the harvested energy, the system should keep the power consumption very low if only a few families use the system at the same time. Furthermore, the number of clients varies depending on the time of day: it is high during the day and low at night. Hence, the system must autonomously match its capability and power consumption to the prevailing conditions.

3.3. Architecture

Our system design is based on the concept of standard web server architecture and the energy-proportional web cluster [14]. Figure 5 shows the overall view of this system. It comprises a combination of a GNU Radio host, two web servers, batteries, photovoltaic (PV) devices, and switches, where the GNU Radio host includes a cluster manager, a load balancer, and a web server.

Among them, the cluster manager plays the most important role for the energy-proportional operation. The cluster manager is responsible for receiving the current request rate periodically from the load balancer and controlling the number of active servers by Wake-On-LAN (WOL) according to the predicted workload. The cluster manager must collect important operational parameters such as the performance of each server, the maximum tolerable request rate, and the transition time to wake up and sleep in advance. Based on this information, the cluster manager determines the minimum number of active servers necessary for the predicted load while leaving room for coping with unexpected bursts of traffic. The cluster manager also should be able to modify the setting of the load distribution in the load balancer. Conventional load balancers automatically detect the status of each web server and reallocate the load distributions. However, in our scenario, the cluster manager intentionally shuts down some web servers to reduce the power consumption. Therefore, the load balancer should reallocate the load distribution proactively following the directives of the cluster manager.

We add the function of monitoring the remaining battery power to the cluster manager in order to support the power-aware control. When low battery power is detected, the cluster manager tries to turn off web servers to prevent them from abnormal shutdown due to power shortage. Moreover, it is useful to construct a resilient system that can provide a minimum level of quality service when not enough energy is provided. A simple energy management strategy that prioritizes system availability over performance can be formulated based on prediction of solar irradiation and load. First, it determines a time at which sufficient irradiation for charging batteries is expected to be available, from weather forecasts

Figure 6. Implementation of the prototype system.

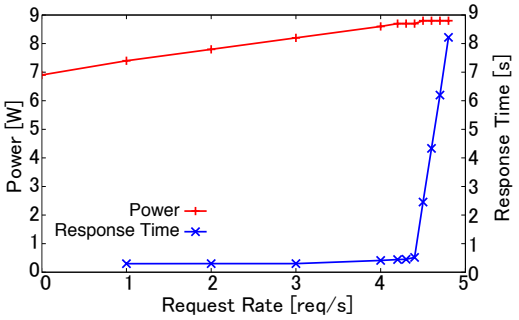


Figure 7. Performance of the web server.

Table 1. Server Specification [15].

CPU	Intel Atom Z530P/Z510P
Memory	1 GB (200 pin SO-DIMM)
Strage	8 GB (CompactFlash card)
Active State Power	8.6 W
Standby State Power	1.5 W

and statistics. Then, the expected power consumption before that time is derived from the predicted demand variation. The cluster manager restricts the maximum number of active servers if the power consumption exceeds the usable energy stored in the batteries.

3.4. Implementation

We have implemented a prototype system as described in Section 3.3 except for GNU Radio. Figure 6 shows all the devices that make up the system. We installed a 120 W solar panel on the balcony of our building that generates electrical power and charges a battery through a maximum power point tracking (MPPT) controller. The DC current generated by the solar panel is stored in the battery and then converted to AC using an inverter.

The three machines shown at the left side of Figure 6 are the web servers. The bottom one labeled EMBOX0 also plays the role of a load balancer and cluster manager because the load of these two modules is less than 1% of the CPU usage when handling 10 000 requests per second. The aggregation of the functions into one server reduces the number of machines and the total power consumption of the system. An embedded Linux platform with Intel Atom Z530P/Z510P was selected. Table 1 shows the specifications of this embedded device. Each server consumes only 8.6 W during operation according to the specifications. Moreover, it supports not only WOL but also a sleep state, which fulfills the second requirement defined in Section 3.2. WOL is a necessary feature for energy-proportional control in our system, and the sleep state is also helpful to minimize the time needed to change server states. It takes less than 4 s for this server to transition from the sleep state to the active state. All the system components are available off the shelf, and thus the first requirement is also fulfilled.

Linux OS is installed on these servers. Apache2, PHP5, and PukiWiki are chosen to provide web server functionality that supports the dynamic pages required for reservation of charging and other services. To understand the performance of this machine, we conducted a unit test with Apache Bench. Figure 7 shows the power usage and response times against the request rates of the web server. We took these measurements after applying a specific burden on the server for 30 s to clarify the limit of its capability. As shown in the result, the response time remains low as long as the load is equal to or less than 4.4 requests per second. However, it increases drastically after the load exceeds this threshold. We observed that CPU usage was close to 100% at over 4.4 requests per second, and power usage also reached the upper limit. In addition, we also measured power consumption in suspend and shutdown states, both of which were 1.5 W.

In addition to a web service, we configured a load balancer with Linux Virtual Server (LVS) in one of the servers. LVS is one of the load balancing solutions for Linux systems. LVS supports the Direct Routing request dispatching mode in which a load balancer forward packets to web servers after changing only the MAC address. Since this method allows the load balancer to omit IP-level analysis of packets, which reduces the workload of load distribution effectively, it is suited for low-power machines with relatively low computing capability. Furthermore, we can configure the LVS by using *Keeplived* and *ipvsadm* commands. *Keepalived* informs the LVS about the server states periodically for automatic load distribution configuration. For example, if *Keepalived* detects the accidental failure of servers, the LVS changes the weight of the load distribution to the servers to 0. Moreover, it also senses the participation of new servers, which is useful for energy-proportional control because the number of active servers varies frequently. On the other hand, the *ipvsadm* command enables/disables load distribution to a specific server forcibly by changing the weight and provides information on the number of current connections with the *--stat* option. In our prototype, we adopt weighted round robin as the scheduling method of the LVS. The weight for active servers is set to 1 and that for inactive servers is set to 0.

Next, we implemented a cluster manager in the same machine as the load balancer. It receives the request rates from the load balancer every second and calculates the anticipated

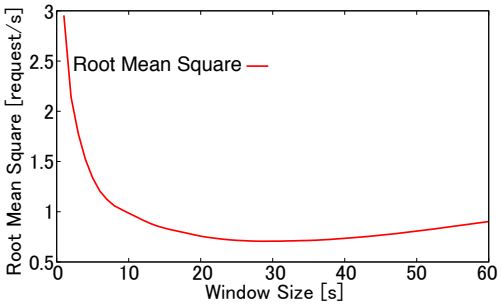


Figure 8. Differences between the predicted and actual workload for various window sizes.

workload. According to the obtained load, the cluster manager turns on/off the servers and updates the change of states at the load balancer. These functions are implemented in Python code that was written from scratch. Because we did not implement the mechanism of utilizing the remaining battery power as a system control parameter, we use incoming requests as workload parameters in this paper. There remains the problem that the load balancer and the cluster manager constitute the single point of failure of the system. Fortunately, *Keepalived* supports the Virtual Router Redundancy Protocol (VRRP), which provides a failover function to load balancers. Thus, the challenging task is for the cluster manager to identify the current master load balancer; this is a research problem for the future.

Regarding the workload prediction algorithm of the cluster manager, Krioukov et al. indicate that the moving window average (MWA) performs better than the other methods [14]. MWA predicts that the incoming request rate will be equal to the average request rates seen over the last n seconds. Unfortunately, the window size determination method is not defined in the paper. In order to find the best window size, we examined the accuracy of prediction for sample traffic while changing n from 0 to 60. In our experiment, we used the Poisson distribution, which fits the short-term traffic distribution well [16]. We can easily forecast that a larger window size predicts the arrival rate λ [req/s] well if λ is almost static. On the other hand, too large a window size may cause a slow reaction to the change of the base workload. Thus, we conducted the experiment while increasing the value of λ from 0 to 15 in 600 s to find the best window size. We concluded that $n = 30$ is the best value to absorb the influence of burst traffic and react to an increase in the base workload quickly, as shown in Figure 8.

Moreover, we determine the thresholds to wake up or suspend web servers by considering the Poisson process, given $P[N_t = k] = e^{-\lambda t}(\lambda t)^k/k!$. One server can handle up to 4.4 requests per second as shown in Figure 7. We select the maximum value of λ that gives over 99% probability of 44 or fewer requests arriving in 10 s ($\sum_{k=0}^{44} P[N_{10} = k] \geq 0.99$) as a threshold to start the second server, that is, $\lambda = 3.09$. Similarly, λ f for the arrival of 88 or fewer requests is the threshold to start the third server ($\lambda = 6.85$). On the other hand, the thresholds to suspend active servers are set for the arrival of 22 or fewer ($\lambda = 1.33$) and 66 or fewer ($\lambda = 4.94$) requests, which are lower than the rates to wake up in order to stabilize server states at the request rate near the thresholds. Additionally, we configure the cluster manager so that it will suspend an over-provisioned server after disabling load distribution to it to prevent connections between the server and clients from being terminated. This function can be implemented by using the*ipvsadm* command.

We connected these machines and clients with a 100 Mbps switch and examined the power consumption of the servers with an electric power meter. We measured AC power, that is, the output of the inverter, but not DC power, that of the battery, so that these data can be referred to in a different experimental environment, such as using commercial AC power for power supply.

4. EVALUATION

We conducted four kinds of experiments to evaluate the performance of our prototype system.

First, we carried out a load test in the same way as the unit test in Section 3.4 to examine the power usage and response times at specific request rates. We compared energy-proportional control to the conventional method in which all web servers are always active, in Figure 9. The results show that there are no differences between these two methods in terms of response times. We observed that response times were within 550 ms at request rates up to 13 requests per second, which is slightly longer than response times in the unit test shown in Figure 7 because of the additional overhead of the load balancing process. The response time starts to increase when the workload exceeds the capacity of all the servers, and the system becomes dysfunctional at 14 requests per second. On the other hand, power consumption differs between the two methods: energy-proportional control can reduce power usage effectively. The sharp slopes of power consumption of energy-proportional control signify the start-up of additional web servers. It occurs when the workload surpasses the wake-up thresholds as defined in Section 3.4. The effect of energy conservation is especially prominent at low request rates. The difference is almost double in the case of the idle state.

Second, we observed the transition behavior of web servers by generating a stepped traffic that increases or decreases request rates every 15 s. Figure 10 shows the results of this experiment. As expected, the cluster manager successfully issued commands to wake up additional servers when the request rate increased and put them to sleep when the request rate dropped. The transition of the server states is smooth; no responses exceed 550 ms, unlike the results of the previous experiment shown in Figure 9.

Third, we conducted a similar experiment to the second one, changing only the request rate interval from 15 s to 10 s. In this experiment, the response time was worse when the request rate was 5 or 6 requests per second, as shown in Fig-

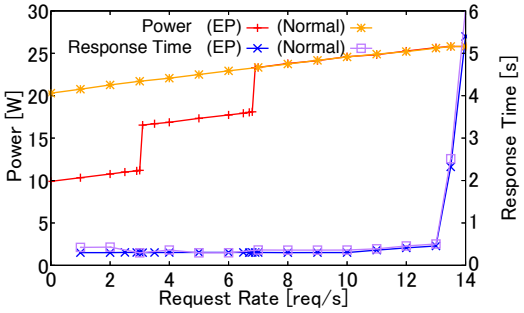


Figure 9. Comparison between energy-proportional control and conventional operation.

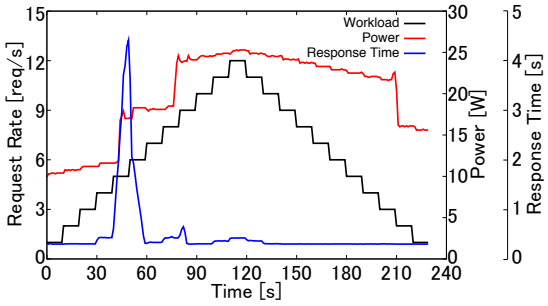


Figure 11. Energy-proportional control for stepped traffic (intervals of 10 s).

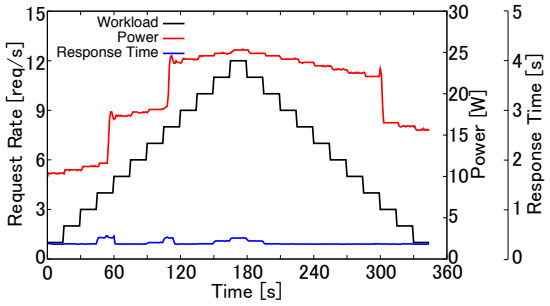


Figure 10. Energy-proportional control for stepped traffic (intervals of 15 s).

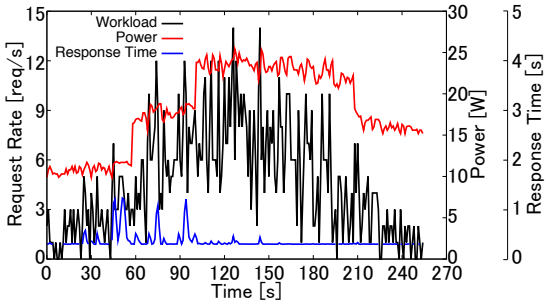


Figure 12. Energy-proportional control for Poisson traffic (intervals of 15 s).

ure 11. This degradation of response time occurred as a result of an accumulation of delay caused by a sudden increase in requests. In accordance with the configuration described in Section 3.4, the cluster manager tries to wake up the first additional server after 41 s from the start of this experiment. Because it takes 4 s before the additional server becomes ready, the server running from the beginning must handle the load of 5 requests per second for 5 seconds, which causes the response delay shown in Figure 7. For the same reason, the delay recurs after 82 s from the start. Thus, the operational parameter that determines whether to boot and suspend each server is very important to prevent overloading the server.

Finally, we observed the system behavior for the Poisson traffic. In the same manner as the previous experiments, the average request arrival rate λ is increased or decreased every 15 s. As shown in Figure 12, the state of each server is stable and tracks the value of λ in spite of the sharply oscillating request rates. Moreover, we can see the resemblance between the tracks of power consumption shown in Figures 10 and 12. This is because averaging request rates with the right length of window size absorbs the influence of the oscillation.

The delay in booting and sleeping a server affects the overall throughput. The decision making at the cluster manager—whether or not an additional server should be turned on—is quite important to cope with the sudden increase in the load. For instance, if we employ a pessimistic policy that activates one more web server at all times, the system will not experience any increase in response time even when the increase in the load doubles. However, this solution raises the problem of power consumption. Because the available power from solar modules is governed by the weather, this course of action is not always possible. There is a clear trade-off between saving energy and system performance and their respective priorities vary according to time and circumstances. To solve this problem, it will be necessary to add another function to change the management policy in each condition.

5. RELATED WORK

Existing CPUs support an energy-saving function called Dynamic Voltage and Frequency Scaling (DVFS) that adjusts CPU clocks and voltage dynamically in response to the load. This technique effectively lowers CPU power, but the ratio of CPU power to the total system power is almost always less than 40% [17]. Unfortunately, other components such as memories and disk drives do not have a wide dynamic power range like CPUs. Therefore, servers consume more than half their full power even when they are idling [13].

Several papers utilize Virtual Machines (VMs) to conserve energy. VMs can be consolidated into a hypervisor [18]. The consolidation of VMs keeps the resource utilization efficiency of hypervisors high and reduces the number of working machines. In our case, however, a hypervisor cannot host many VMs effectively due to its limited resources. More-

over, a boot or migration of VMs might cause additional overhead. For example, it takes 20 s for the low-power server we used to boot a new VM, which is 5 times longer than waking up a new physical server. Therefore, we managed the system in units of physical servers, not VMs. VM-based energy conservation is probably effective for larger-scale systems if low-power and high-performance servers become available in the future.

Other research focuses on energy-proportional computing proposed by Barroso et al. [13]. This phrase indicates an ideal feature of energy-efficient servers, namely, power consumption is proportional to the load or CPU usage. Since this feature is not available in existing servers as aforementioned, it is often emulated roughly in a server cluster by turning off excess servers depending on the load of the cluster.

Krioukov et al. proposed an energy-proportional web cluster consisting of a set of heterogeneous machines for use in data centers [14]. We employ their idea as the basis of our research and plan to improve the management approach by adding new variables related to photovoltaic generation and battery power. Moreover, we show how to determine a parameter used for load prediction, which is not clearly mentioned in the previous research.

6. CONCLUSION

We considered the feasibility of distributing excess power generated at the AP to cell phone customers in emerging regions. Calculations based on the premise of 60 W of constant power consumption at the AP with the minimum required photovoltaic system show that about 9.3% of users can charge their handsets on average from dumped energy. Furthermore, this calculation method can be applied to estimate additional solar modules and batteries required to charge a larger number of phones. It is an interesting economic exercise to compare extra equipment investment with the expected revenue from the charging and communication fees.

Then, we described the importance of a web-based information service for the cell phone charging service and presented a design for a solar-powered energy-proportional web server cluster. To utilize this system in emerging regions, we defined two requirements for our system: using low-power and inexpensive devices running on just solar power and being autonomous and scalable, starting from a very small village. According to these requirements, we designed a system framework so that our system can adjust the number of active servers in proportion to the workload, and implemented a prototype with inexpensive low-power machines.

We have evaluated the performance of our system in terms of power consumption and response times by comparing energy-proportional control with a conventional method that always keeps all web servers active. As a result, our prototype can reduce power usage up to over 50% when in the idle state. Furthermore, it does not show any response delay for stepped traffic load that increases and decreases the request rate at 15 s intervals. We expect this system can also be applied to information services in developed countries during a disaster recovery period after earthquakes where a cell phone AP is powered by solar modules.

REFERENCES

[1] D. Marquet, M. Aubrée, S. Le Masson, A. Ringnet, P. Mesguich, and M. Kirtz, "The first thousand optimized solar BTS stations of Orange group," in *Telecommunications Energy Conference (INTELEC), 2011 IEEE 33rd International*, Oct. 2011, pp. 1–9.

[2] J. C. Aker and I. M. Mbiti, "Mobile Phones and Economic Development in Africa," *Journal of Economic Perspectives*, vol. 24, pp. 207–232, June 2010.

[3] "Green Power for Mobile Charging Choices 2011: Mobile Phone Charging Solutions in the Developing World," GSM Association, July 2011.

[4] Grameen Foundation, *Village Phone Direct Manual*, 2 edition, Feb. 2008.

[5] "OpenBTS Public Release," `https://wush.net/trac/rangepublic/`.

[6] "5150 Series," `http://www.rangenetworks.com/store/5150-series`.

[7] "Solar Radiation Data," `http://www.soda-is.com/eng/index.html`.

[8] "Toshiba 240W Solar Module," `http://www.toshiba.co.jp/sis/h-solar/news/240w/index_j.htm`.

[9] K.H. Lee, K. Malmedal, and P.K. Sen, "Conceptual design and cost estimate for a stand-alone residential photovoltaic system," in *Green Technologies Conference, 2012 IEEE*, Apr. 2012, pp. 1–6.

[10] P. Leishman, "Is there Really any Money in Mobile Money?," GSM Association, Oct. 2010.

[11] A.C. Nchise, R. Boateng, I. Shu, and V. Mbarika, "Mobile Phones in Health Care in Uganda: The AppLab Study," *The Electronic Journal of Information Systems in Developing Countries*, vol. 52, 2012.

[12] "Nokia 103 Data Sheet," Nokia, 2012.

[13] L.A. Barroso and U. Holzle, "The case for energy-proportional computing," *Computer*, vol. 40, no. 12, pp. 33–37, Dec. 2007.

[14] A. Krioukov, P. Mohan, S. Alspaugh, L. Keys, D. Culler, and R. Katz, "NapSAC: Design and Implementation of a Power-Proportional Web Cluster," *ACM SIGCOMM Computer Communication Review*, vol. 41, no. 1, pp. 102–108, Jan. 2011.

[15] "EMBOX Type M20," `http://www.innotech.co.jp/products/product_list/embedded/cpubord/embox_m20.html`.

[16] M. Andersson, A. Bengtsson, M. Höst, and C. Nyberg, "Web Server Traffic in Crisis Conditions," in *Proceedings of the Swedish National Computer Networking Workshop*, 2005.

[17] D. Meisner, B. T. Gold, and T. F. Wenisch, "PowerNap: Eliminating Server Idle Power," *SIGPLAN Notices*, vol. 44, no. 3, pp. 205–216, Mar. 2009.

[18] L. Hu, H. Jin, X. Liao, X. Xiong, and H. Liu, "Magnet: A Novel Scheduling Policy for Power Reduction in Cluster with Virtual Machines," in *Cluster Computing, 2008 IEEE International Conference on*, Oct. 2008, pp. 13–22.

PROPOSAL OF A SUB-λ SWITCHING NETWORK AND ITS TIME-SLOT ASSIGNMENT ALGORITHM FOR NETWORK WITH ASYNCHRONOUS TIME-SLOT PHASE

Keisuke Okamoto[1], *Atsushi Hiramatsu*[2]

[1]Graduate School of Informatics, Kyoto University
[2]NTT Photonics Laboratories, NTT Corporation

ABSTRACT

We propose a sub-λ switching network which has fine granularity and low cost/power consumption. In this network, each wavelength is divided in time domain to achieve fine granularity, but buffers are eliminated from the core network to reduce switching cost and power consumption. Buffers are located only at the entrance of the network in order to groom ingress traffic, and all nodes in this network are operated synchronously under a certain time-control mechanism. The problem in this network is the rather long guard-time which is required to absorb the clock synchronization error and time-slot-phase difference, which is caused by the various fiber lengths between nodes (asynchronous phase network). To solve this problem, we propose a novel time-slot assignment algorithm using multi-time-slot bonding technique and delay-shift packing technique with global-time-based delay shift. By using the proposed method, we could improve the utilization of link capacity by 45% compared with the conventional method.

Keywords— sub-λ switching, time-slot assignment algorithm, asynchronous time-slot phase

1. INTRODUCTION

Reducing power consumption in broadband networks is an urgent problem since the Internet traffic keeps exponentially increasing. An all-optical network based on wavelength division multiplexing (WDM) seems one of promising solutions to this problem since many O/E and E/O converters in the network are eliminated, and it can realize very large capacity network by exploiting the advancing transmission technologies. A WDM network, however, exhibits inefficient bandwidth utilization because of its coarse bandwidth granularity. When the amount of traffic between two nodes is not large enough to fill one wavelength capacity, electrical switch/routers are often used with WDM nodes for traffic grooming, and thus the power consumption is still an issue in the case of carrying many small traffic paths [1]. Therefore, an efficient sub-λswitching with low power consumption is needed. One approach to realize a sub-λ switching network is all-optical time driven switching (TDS). In TDS network, optical bursts are all generated at the scheduled time-slots based on the precisely-synchronized global clock. And the intermediate optical switches transfer them to their destination links based on the time-slot position, considering the propagation delay. The TDS architecture provides fine bandwidth granularity by the time-slot division and reduce power consumption by at least 40% since OEO conversion is not needed at the optical node[2]. The problem of TDS network is the inefficiency due to the rather long guard-time. Usually the guard-time is required to absorb the switching time of optical switches. In TDS network, it is also used to absorb the clock synchronization error and time-slot-phase difference, which is caused by the various fiber lengths between nodes (asynchronous phase network).

Our goal in this paper is to provide a novel method to realize a very efficient TDS network by reducing the effect of the guard time. We first propose a multi-time-slot bonding technique. The second issue we address is time-slot phase synchronization. If time-slot phase of all the signals coming from different nodes are synchronized, switch operation and time-slot assignment at optical nodes become very easy. But, in order to make such a network, we need to adjust the propagation delay of each fiber between nodes to be the multiple of time-slot duration. And thus, for each transmission section, we have to add a fiber cable for one time-slot duration at worst case to adjust transmission delay (ex.: around 2 km fiber for a 10 μs time-slot). Therefore, we propose a delay shift packing technique with global-time-based delay shift. It can operate in an asynchronous time-slot phase TDS and enables high network utilization by efficiently allocating time-slots.

The remainder of this paper is organized as follows. We explain our proposed sub-λ switching network in Section 2. In Sections 3 and 4, we propose a multi-time-slot bonding technique and a delay shift packing technique. We describe related work in Section 5, and explain simulation results in Section 6. We conclude the paper in Section 7.

2. SUB-λ SWITCHING NETWORK

2.1. Network configuration/architecture

The network architecture of a sub-λ switching network is shown in Fig.1. A sub-λ switching network is constructed from one scheduler, optical time-slot switches of wavelength division multiplexing (written as optical switch hereafter) and optical time-slot adapters (written as adapter hereafter).

The scheduler creates a schedule table, which controls the timing of optical time-slot switching and that of burst generation from the electrical buffer of the adapter. The adapter stores packets at buffers dedicated for each destination adapter, and transmits those packets at the scheduled time-slots. Time-slots arrive at the optical switch with a certain propagation delay.

For example, as shown in Fig. 2, time-slot t1 transmitted from adapter A1 reaches optical switch C1 at time-slot t2, where the propagation delay between A1 and C1 is 1 time-slot length. Similarly, time-slot t2 from A2 reaches C1 at t5, where delay between A2 and C1 is 3 time-slot lengths. Therefore the scheduler must allocate time-slots to all those traffic hop-by-hop by taking the transmission delay into account. Moreover, in the above example, propagation delay between two nodes was assumed to be multiple of the time-slot duration, but this is not the case in the real environment. And thus the scheduling becomes a more complicated problem. We will discuss this later in Section 4.

In this network, the scheduler calculates the actions of the adapters and optical switches for a certain period called frame based on the traffic amount between adapters, and then stores the results in the schedule table for each adapter and optical switch. The adapters and optical switches perform the series of actions stored in the schedule table repeatedly and synchronously. This sub-λ switching network can be applied to backbone networks as a cross-connect network with fine bandwidth granularity as well as to leased line service networks. Therefore, here we consider a static scenario in which a static traffic matrix is constant and given.

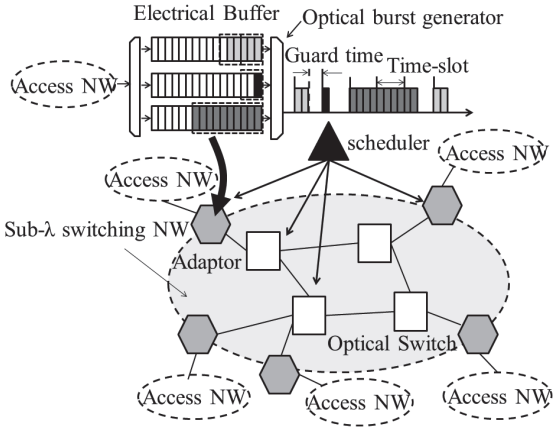


Figure 1. sub-λ switching network architecture

2.2. Time Synchronization and Guard-Time Length Estimation

Time synchronization at the sub-microsecond level is required in the proposed network, and Global Positioning System (GPS) is one of technology candidates to achieve this. The time synchronization mechanism of passive optical network (PON) can also be extended to wide area networks. The feasibilities of these two approaches have been demonstrated

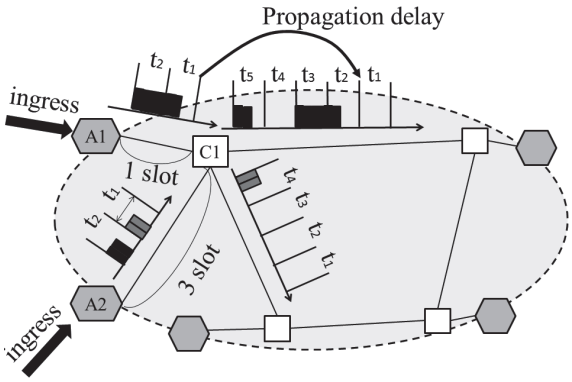


Figure 2. Example of time-slot flow in sub-λ switching network

[3], [4]. We may also apply the precision time protocol of IEEE 1588 [5].

The guard-time in this network is determined by (A) the timer synchronization error, (B) the propagation delay variation caused by temperature change and (C) switching time of the optical switch, as shown in Fig. 3. First we estimate the timer synchronization error (A) as ± 0.5 μs, assuming GPS. The typical upper-bound for the propagation delay variation (B) can be estimated as ± 1.2 μs assuming the temperature variation of ± 30 ℃ and the maximum network span of 1000 km (± 30 ℃ $\times$1000 km$\times$40 ps/km$\cdot$ ℃ [6]). When we assume the optical switching time C as 10 ns, the guard time should be larger than 4A + 2B + C = 4.41. In the latter part of this paper, we set guard time length to be 5 μs.

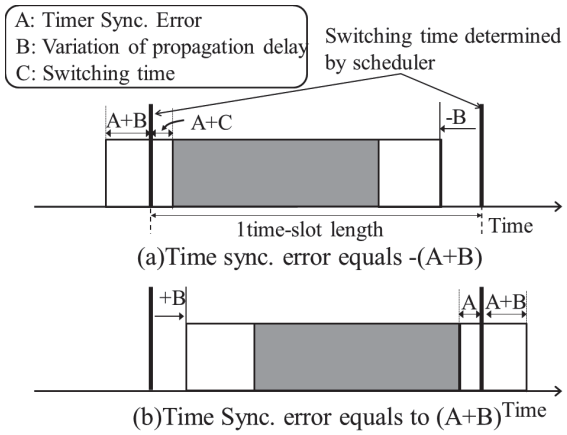


Figure 3. Guard-time design

3. MULTI-TIME-SLOT BONDING

In this section, we discuss the reduction of guard time. In many cases, time-slot length is designed much longer than guard-time length to increase the link utilization rate upper bound determined by the ratio of guard-time length to time-slot length. However, the network using a long time-slot is

inefficient when the traffic volume is small.

Accordingly, we propose multi-time-slot bonding technique where multiple optical bursts to the same destination are allocated in the successive time-slots without guard-time in-between as show in Fig. 4. This technique can reduce the number of guard-times and improve network utilization efficiency.

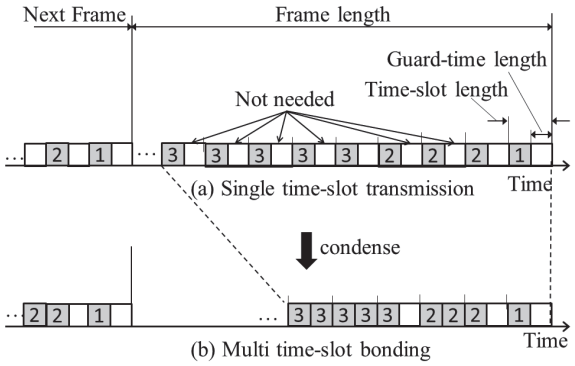


Figure 4. Concept of multi-time-slot bonding

The link utilization ratio upper bound can be calculated as follows. Assuming T_g, F, B, and V are the guard-time length, the frame length, the requested bandwidth, and the link capacity of the network respectively, T, which is the required time in a frame to carry the requested bandwidth, can be calculated by $F(B/V)$). The number n of time-slots required in single time-slot method and multi-time-slot-bonding are given by $\lceil T/(T_s - T_g) \rceil$ and $\lceil (T + T_g)/T_s \rceil$, respectively, and the link utilization ratio is represented by $T/(nT_s)$ The utilization ratios using single time-slot method and multi-time-slot bonding for T_g=5 μs and T_s=10 μs or 50 μs are shown in Fig. 5. The link utilization ratio using multi-time-slot bonding with T_s=10 μs is higher than that using single time-slot method with both of T_s=10 μs and 50 μs. Therefore, we adopt multiple successive time-slot bonding method to the sub-λ switching network, and set the time-slot length to 10 μs, which equals to $2T_g$.

4. DELAY SHIFT PACKING

An efficient time-slot assignment algorithm is important in proposed sub-λ switching network, and many studies have already been conducted for wavelength and time-slot assignment in TDM/WDM networks. However, the proposed network requires new features for the assignment algorithm; time-slot allocation considering multi-time-slot bonding and link propagation delay. Here we propose a "global time-based scheduling algorithm" in Section 4.1 to satisfy those requirements. Then we propose "hashed scheduling method" in Section 4.2, which can reduce the unused regions in the scheduling table for the asynchronous time-slot phase network shown in Fig. 6. The "time-slot phase" represents the optical burst arrival time according to the time-slot cycle

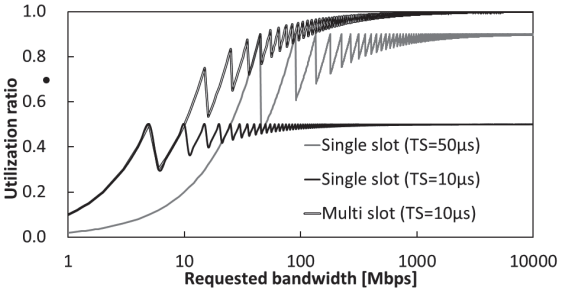


Figure 5. Efficiency of successive time-slot assignment

of the optical switch, and it varies to various values because the fiber lengths between optical switches are not aligned to the multiple of the time-slot duration in a general network. In this situation, the efficiency of the time-slot-based optical burst allocation is inefficient since many unused time regions remain in the time-slot scheduling table, as shown in Fig. 6.

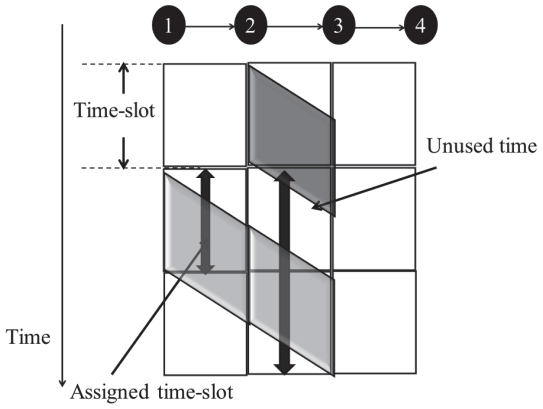


Figure 6. Time-slot-phase difference

4.1. Global Time-Based Scheduling

In the time-slot allocation, we usually start searching a vacant time-slot from the top of the frame and allocate the first found vacant time-slot. This method is called local time-based scheduling (Fig. 7(a)). In the global time-based scheduling, we shift the time-slot-search start point by the value determined by the propagation delay along the optical burst route from a certain reference node, which is pre-selected in the network. In Fig. 7(b), node 1 is the reference node, and the time-slot-search start point is shifted by 2 and 4 time-slots for the link between node 2 and 3 and the link between node 3 and 4, respectively.

When the network topology is complicated, it is not straightforwardly simple to determine shift volume. We explain the calculation of shift volume in an example in Fig. 8. In this

figure, D_{ij} means propagation delay between nodes i j. Here we consider the path from node 3 to node 4. When the reference node is 2, the shift value is determined uniquely. However, when node 1 is selected as the reference node, there are two ways to determining the shift value: shift value equals $(D_{12} - D_{32})$ when we calculate the link from node 4 to node 2, but it is $-(D_{21} + D_{32})$ when we calculate the link from node 2 to node 3. We call this the "shift lag", and we choose the shift value for the path with heavier traffic. Since the links where the shift lag occurred change with the reference node position, how to select the reference node is important in this method. The details of reference node selection are explained in Section6.

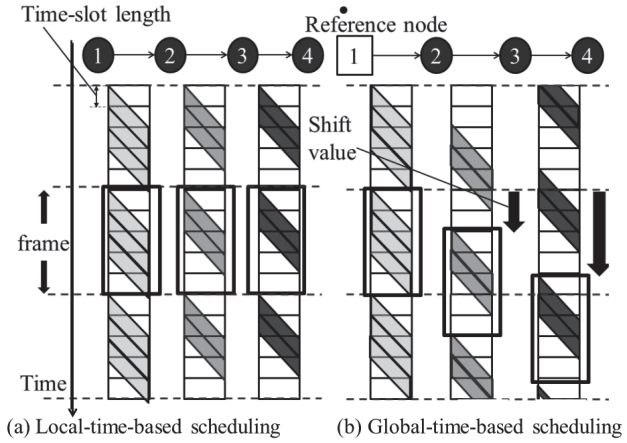
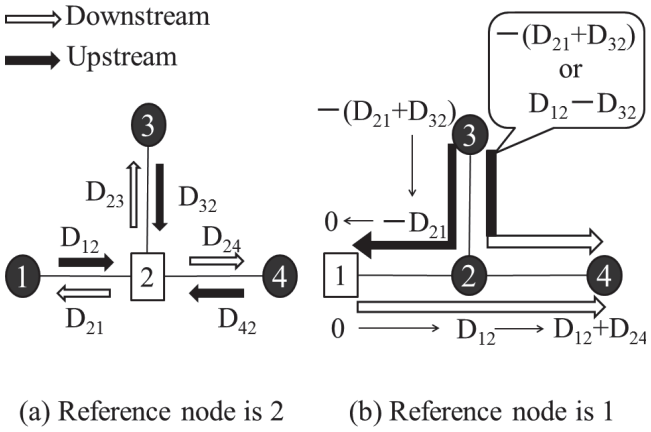


Figure 7. Global time-based scheduling vs. local time-based scheduling



(a) Reference node is 2 (b) Reference node is 1

Figure 8. Shift lag

4.2. Hash scheduling

The concept of hash scheduling is illustrated in Fig. 9. In this scheduling method, the positions of time-slots allocated are not aligned to the time-slot cycle but to the fine grid position dividing the time-slot duration into N sections. The grid interval (unit time) is denoted as $\phi = T_s/N$, and the size of the schedule table becomes N times larger than schedule table based on the time-slot duration. The lower bound for the unit time length will be the switching time of the optical switch, and we set it to 10 ns in this paper as the typical value for the current high speed optical switch [7]. In this approach, the time-slot phase of each optical burst is rounded up to the nearest smaller grid, and the difference shorter than ϕ is absorbed by extending the guard time by one unit time. Therefore, an additional ϕ-longer guard time is needed compared with synchronous phase network, but the difference becomes smaller when N is large. In this paper, we evaluate the time-slot assignment effectiveness due to the increase of N. Note that considering computational complexity is for future work.

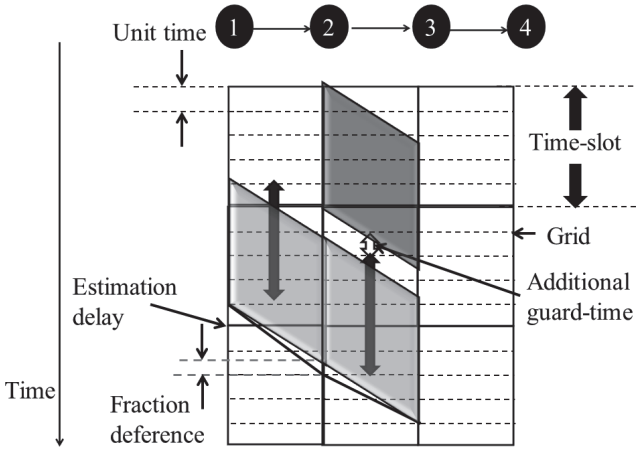


Figure 9. Concept of hash scheduling

5. RELATED WORK

In the proposed sub–λ switching network, as in the general TDM/WDM networks, we need to determine the route, wavelength, and time-slot assignment, which is referred to the routing, wavelength, and time-slot assignment (RWTA) problem [8]. This problem is a well-known hard problem since it can be regarded as an extended problem of the routing and wavelength assignment problem, which is NP complete [9]. Therefore, many studies solve the sub-problems of routing, wavelength assignment, and time-slot assignment separately to simplify the RWTA problem. In this paper, we focus on the heuristic approach for the time-slot assignment algorithm [10][11].Many types of heuristic time-slot assignment algorithms have been proposed, but many do not take into account propagation delay in WDM/TDM networks. For example, heuristic time-slot assignment algorithm involves wavelength conversion [10], but this algorithm ignores propagation delay on links. Liwe and Chao described a synchronous phase network [12], but the jitter dependence on

temperature is not written. On the other hand, Gadkar's work [11] is similar to our consider assumption, but the multi-time-slot bonding allocation is not considered. Therefore no studies on the time-slot allocation problem consider both multi-time-slot bonding and asynchronous time-slot phase without adjustment of fiber length.

6. SIMULATION ANALYSIS

we analyzed a performance algorithm to improve used time by computer simulation. In this section, we show simulation setup and results.

6.1. Simulation Setup

In this subsection, we explain the algorithm for the RWTA problem in our sub-λ switching network, simulation model, and algorithm performance metric.

First, we discuss the algorithm for the RWTA problem. Using the shortest path based on Dijkstra's algorithm with its metric being hop count, we performed routing of the traffic from node to node. To use the network capacity of a link for as long as possible, we did not determine the propagation delay of links but hop count. Wavelength assignment was considered by ignoring wavelength conversion. In other words, time-slot assignment was performed using the wavelength. The time-slot assignment algorithm is as follows. Considering the time-slot assignment problem, we need to first determine traffic-assignment order. We considered using an algorithm based on hop count as traffic assignment order. We first sorted the traffic flow according to their respective lengths and assigned time-slots to the longest traffic flow first. We assumed that longer traffic flow should be given priority since longer traffic flow requires empty an time-slot in many links. In addition, we need to determine time-slot position. Time-slot position was determined by random, local-based scheduling and global-based scheduling algorithms because we evaluated three algorithms by computer simulation, to compare their performance.

Second, we explain the simulation model. As shown in Fig. 10, we chose the topological structure of the National Science Foundation Network (NSFNET), and analyzed the performance of the proposed algorithmss. Using actual geographical distances, we calculated the propagation delay on the links as light velocity equal to 5 ns/m[13]. In Fig.10, the number of propagation delays on links (unit is ms) is between two nodes. Moreover,each adapter was connected to an access ring having five nodes.

The reference node, which is needed for global-based scheduling, was determined as follows. Using our delay shift packing algorithm, we select which adapters in the network as base node. This is important, especially, when proposed network is adapted to a network in which many types of phases exist, and has an asynchronous phase mesh topology. However, this is a difficult problem because many metrics of relevance which are propagation delay of link,

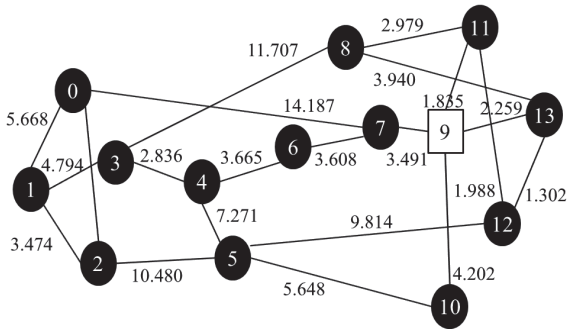


Figure 10. Topological structure

network topology, traffic volume between nodes, and routing should be considered. Therefore, we adopted two heuristic approaches. We used betweenness centrality[14] for the metric of relevance as the first approach because the node through which much traffic passes is suitable for the reference node. Let g_{jk} be the number of shortest path between node j to node k and $g_{jk}(n_i)$ be the number of shortest paths between nodes j k via node i. The betweenness centrality $C_b(n_i)$ is calculated using Eq. (1)

$$C_b(n_i) = \sum_{j<k} \frac{g_{jk}(n_i)}{g_{jk}} \tag{1}$$

However, only using betweenness centrality is insufficient for the metric of base tree which is shown by tree structure with base node as the parent node. Therefore, we used the number of nodes and links causing phase lag for the second approach. Figure 11 shows the types of node and link which need to be counted for the base tree metric. In Fig. 11, the divergence node means the node connected with more than two next-hop nodes, and a confluence node means the node connected with more than two former-hop nodes. In addition, equivalent hop link means the link between two nodes that have the same hop count in terms of the shortest path from the base node. Counting the number of nodes and links, we evaluated the suitability of the base tree to our sub-λ switching network. Therefore, we determined the base node by heuristic decision based on the previous two metrics. The reason the base node was selected as node 9 and the appropriateness of this approach is discussed in the next subsection.

Finally, we will give simulation parameters and a performance metric. The traffic volume between two nodes was determined at random following uniform distribution from 0 to 95 Mbps. The other parameters are listed in Table 1

we used the increase ratio of the minimum time-slot for the performance metric. The increase ratio of the minimum time-slot is calculated as follows. Let T_{max} be the required number of time-slots, which is the required number of time-slots to accommodate all data on the most heavy

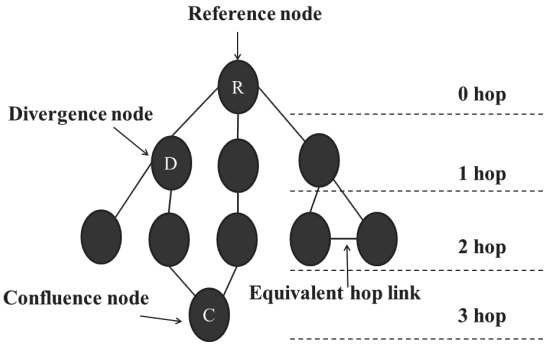


Figure 11. Base tree

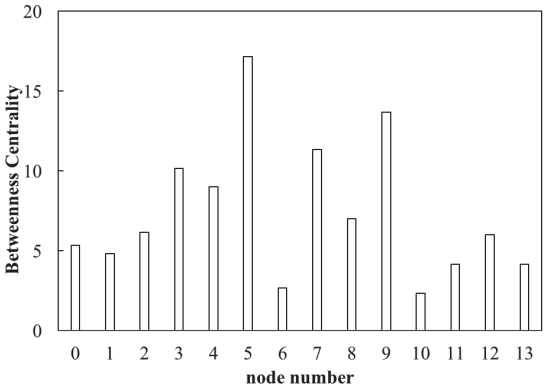


Figure 12. Betweenness centrality of NSFNET

time-slot length	10 μs
guard-time lengh	5 μs
frame length	10 ms
link capacity	10 Gbps

Table 1. Simulation parameters

link determined by routing. Let T be the required number of time-slots, which is the required number of time-slots determined using the time-slot assignment algorithm. In other words, T means the required frame length to accommodate all traffic. T is calculated by computer simulation, as frame length increases by 1 time-slot from the frame length with 1 time-slot and then scheduling is performed. The increase ratio of the minimum time-slot is defined as $(T-T_{max})/T_{max}$, that is to say this value is the increase ratio from the lower bound determined by routing. The simulation was repeated one hundred times, and then the obtained value was taken as the average. We used this value for the simulation results.

6.2. Simulation Results

We conducted computer simulation for evaluating delay shift packing.

First, we discuss reference node selection. Figure 12, shows betweenness centrality calculated using Eq. (1) in NSFNET. Note that we can choose centrality betweenness based on hop count or propagation delay, but we chose centrality betweenness based on hop count because we used the routing algorithm based on hop count. Figure 12 gives betweenness centrality of each nodes, and shows that node 5 has the highest betweenness centrality than other node in NSFNET and node 9 has the second highest. In addition, these nodes are considered appropriate as reference nodes since they are connected with four nodes. Therefore, we regarded these two nodes as candidates for base nodes.

Moreover, figure 13 gives base tree when the reference node are 5 and 9. when the reference node is 5, focusing on nodes 8 and 7, there are nodes with three confluences. The nodes are regarded as causing much shift lag. On the other hand, if the reference node is 9, there are no three confluence or three divergence node. Therefore, we regarded the base tree of node 9 as having better a structure than that of node 5. As a result, we selected node 9 as the base node.

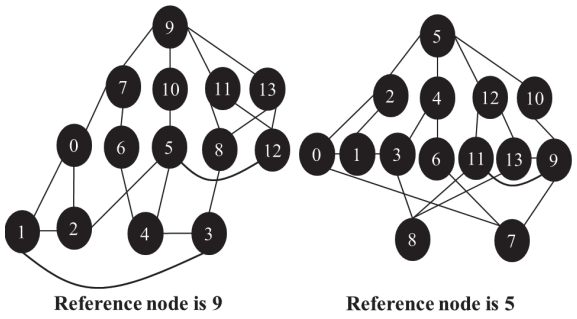


Figure 13. Example of base tree

In addition, to confirm the appropriateness of this approach, we conducted computer simulation with synchronous time-slot phase and no access ring. Note that the maximum value of traffic volume between two nodes was 995 Mbps. Figure 14 show the increase ratio of the minimum time-slot when reference node is selected respectively. When the reference node is 9, the increase ratio of the minimum time-slot is the lowest. The node that has the highest centrality betweenness of the nodes in NSFNET is node 5, but the increase ratio of the minimum time-slot is not the lowest. Therefore, we showed that the two metric was appropriate for base selection.

Second, we confirmed the effectiveness of our global time-based scheduling algorithm. Having assumed an synchronous time-slot phase network in order to eliminate the effect of asynchronous phase, we conducted a simulation analysis to evaluate this algorithm performance. We also conducted computer simulation using random and local

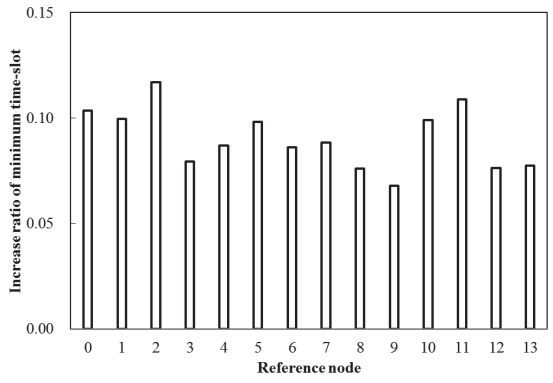


Figure 14. Appropriateness of base selection

time-based scheduling as references. Figure 15 shows the simulation results. These results show that local time-based scheduling, which ignores propagation delay on links, cannot assign time-slot in order because of much shift lag. On the other hand, having used global time-based scheduling, we reduced the increase ratio of the minimum time-slot by 0.04 compared to random assignment. Accordingly, we confirm that global time-based scheduling improved the effective time of network capacity.

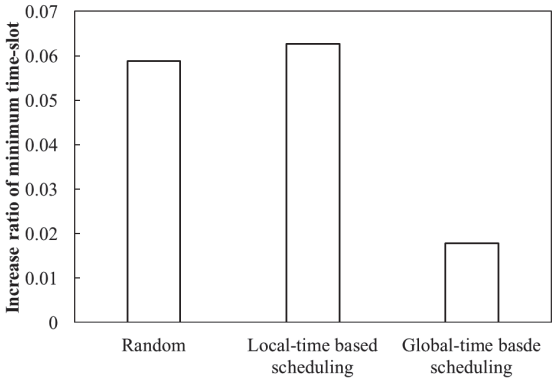


Figure 15. Evaluation of global time-based scheduling

Third, we discuss the effectiveness of hash scheduling. Figure 16 shows these results. The results in synchronous phase network are given as reference. when we used hash scheduling, time-slot was divided into five unit time, that is, unit time length equaled 2 μs , and scheduling was performed. Figure 16 shows that the increase ratio of the minimum time-slot reduced by 0.12 compared with the situation where no algorithm was used, when we used hash scheduling. we were able to bring the value in the phase synchronous network near the value in the phase asynchronous network . In this computer simulation, the time-slot-phase difference under unit time was not involved in guard-time, but we expect to obtain the same effectiveness by increasing the hash number

Therefore, we consider the relationship between hash number and increase ratio of the minimum time-slot. Figure 17 shows the increase ratio of the minimum time-slot when the hash number is changed. Note that we regard the minimum time-slot when hash number is equal to 1 as the base value. The increase ratio is then calculated using this value. In addition, to evaluate effect of hash number,the time-slot-phase difference under unit time was not involved in guard-time. Figure 17 shows that the increase ratio of the minimum time-slot deteriorated by increasing the hash number. However, the increase ratio of minimum time-slot does not rapidly increase with hash number. Therefore, by increasing the hash number, we can consider effective time of network capacity becoming long.

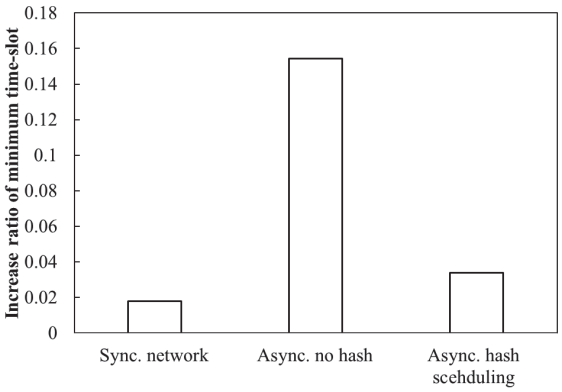


Figure 16. Increase ratio of minimum time-slot in asynchronous network

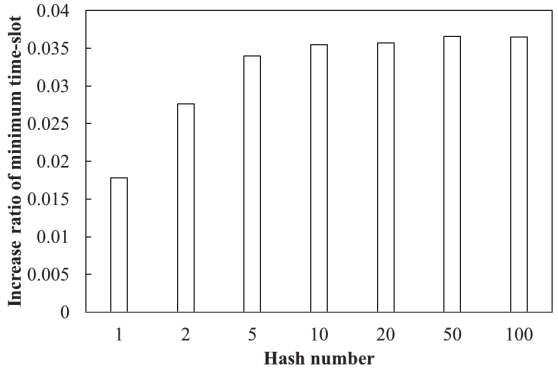


Figure 17. Increase ratio of minimum time-slot by changing hash number

Finally, compare each algorithm in the asynchronous phase network. Figure 18 shows that the increase ratio of the minimum time-slot equals 0.9 when a separate time-slot and random assignment algorithm are used. By using multi-time-slot bonding, we reduced the increase ratio of the minimum time-slot by about 0.7. Moreover, when we used delay shift packing and hash scheduling algorithms , it was possible to reduce the increase ratio of the minimum time-slot by about 0.15. By using proposed algorithm,the increase ratio

of the minimum time-slot reduced by 0.85 in total. This value equals to improving link capacity by 45% ($= \frac{0.85}{1+0.9}$) comparing with the case in which no bonding is used.

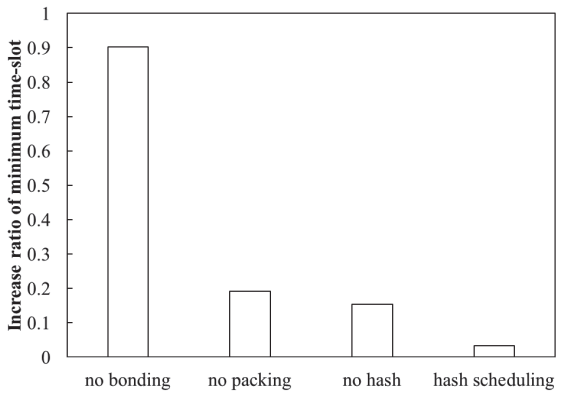


Figure 18. Comparison of each algorithm

7. CONCLUSION

We proposed a sub-λ switching network, which is all-optical, that achieves finer granularity than a λ switching network. In this network, it is possible to reduce power consumption by using the TDS architecture. We also proposed a multi-time-slot bonding technique and a delay shift packing algorithm to improve the effective time of the network. Guard-time is used with multi time-slot bonding and each traffic flow can be assigned in mass time-slots unit by global time-based scheduling. In addition, in an asynchronous time-slot phase network, using scheduling with the unit time into which a time-slot is divided , we showed that we were able to absorb the phase synchronization error. We also simulated time-slot scheduling in NSFNET using multi-time-slot bonding and delay shift packing, we showed that the increase ratio of the minimum time-slot can be reduced. By using multi-time-slot bonding and delay shift packing, we improved network utilization by 45% compared with the situation where no algorithm is used.

REFERENCES

[1] E. Oki, K. Shiomoto, D. Shimazaki, N. Yamanaka, W. Imajuku, and Y. Takigawa, "Dynamic multi-layer routing schemes in GMPLS-based IP+optical networks," *Communications Magazine, IEEE*, vol. 43, no. 1, pp. 108 –114, jan. 2005.

[2] F. Musumeci, F. Vismara, V. Grkovic, M. Tornatore, and A. Pattavina, "On the Energy Efficiency of Optical Transport with Time Driven Switching," in *Communications (ICC), 2011 IEEE International Conference on*, june 2011, pp. 1 –5.

[3] G. Fontana G. Marchetto Y. Ofek D. Severina M. Baldi, M. Corra and O. Zadedyurina, "Scalable fractional lambda switching: A testbed," ″ *IEEE/OSA Journal of Optical Communications and Networking*, vol. 3, pp. 447–457, May 2011.

[4] Yuji Shimada Kunitaka Ashizawa Kazumasa Tokuhashi Daisuke Ishii Satoru Okamoto Yutaro Hara, Takehiro Sato and Naoaki Yamanaka, "A Study of Next Generation Metro-Access Hybrid Scalable Network by Using PLZT Ultra High Speed Optical Wavelength Selective Switch," *1st International Symposium on Access Spaces (IEEE-ISAS 2011)*, pp. 1–6, June 2011.

[5] IEEE 1588, "Standard for a Precision Clock Synchronization Protocol for Networked Measurment and Control Systems," .

[6] Y. Mitsunaga M.Taneda N. Shibata, Y.Katsuyama and S.Seikai, "Thermal Characteristics of Optical Pulse Trainsit Time Celay and Fiber Strain in a Single-Mode Optical Fiber Cable," *Appleid Optics*, vol. 22, pp. 979–984, Apr 1983.

[7] K. Nashimoto, "PLZT waveguide Devices for High Speed Switching and Filtering," *IEEE International Conference on Communications*, pp. 406–410, Jan 1997.

[8] Ramakrishna Shenai Bo Wen and Krishna Sivalingam, "Routing, Wavelength and Time-Slot-Assignment Algorithms for Wavelength-Routed Optical WDM/TDM Networks," *J. Lightwave Technol.*, vol. 23, pp. 2598–2609, Sept 2005.

[9] I. Chlamtac, A. Ganz, and and G. karmi, "Lightpath Communications: An Approach to High Bandwidth Optical WAN ' s," *IEEE Transactions on communications*, pp. 1171–1182, July 1992.

[10] M. Sivakumar and S. Subramahiam, "Performance evaluation of time switching in TDM wavelength routing networks," in *Broadband Networks, 2004. BroadNets 2004. Proceedings. First International Conference on*, oct. 2004, pp. 212 – 221.

[11] A. Gadkar, "A comparison of optical time slotted networks," in *Advanced Networks and Telecommunication Systems (ANTS), 2009 IEEE 3rd International Symposium on*, dec. 2009, pp. 1–3.

[12] Soung.Y Liwe and H.Jonathan Chao, "On slotted WDM Switching in Bufferless All-Optical Networks," *High Performance Interconnects, 2003. Proceedings. 11th Symposium on*, pp. 96–101, Aug 2003.

[13] B. Moslehi, J.W. Goodman, M. Tur, and H.J. Shaw, "Fiber-optic lattice signal processing," *Proceedings of the IEEE*, vol. 72, no. 7, pp. 909 – 930, july 1984.

[14] Ulrik Brandes, "A faster algorithm for betweenness centrality," *Mathematical Sociology*, vol. 25, no. 2, pp. 163 –177, 2001.

POSTER SESSION

A PROPOSAL OF A NEW PACKET SCHEDULING ALGORITHM AND ITS EVALUATION

Tetsushi Matsuda

Information Technology R&D Center, Mitsubishi Electric Corporation
5-1-1, Ofuna, Kamakura-shi, Kanagawa-ken, Japan

ABSTRACT

ITU and other SDOs have launched oneM2M initiative recently and the standardization of M2M is now accelerating. The current access and core networks built for today's network services will be used as a common network infrastructure for M2M network with some modifications. When the current access and core networks are used for both the current network services and M2M services, communications equipments at the network edge need to handle a large number of communication flows which are a mix of large volume data communication such as web access and M2M data communication at the same time. To satisfy the QoS requirements of many applications including M2M applications, communications equipments at the network edge will need to support both minimum guaranteed rate service and low delay forwarding service for small sized packets. In this paper, we propose a packet scheduling algorithm which can provide minimum guaranteed rate service and which can reduce the scheduling delay of small packets. It can be used in access network communications equipment such as edge router and OLT. We also evaluate the proposed algorithm by simulation.

Keywords— Scheduling algorithm, Quality of service

1. INTRODUCTION

ITU and other Standards Developing Organizations such as ETSI have launched oneM2M initiative recently and the standardization of M2M is now accelerating. M2M technologies are widely deployed in real business world for remote asset management and other applications and the deployment of M2M technologies is expected to become more widespread. We think that the current access and core networks built for today's network services such as internet access will be used as a common network infrastructure for M2M network with some modifications because the present focus of M2M standardization is not on the access and core network.[1] But new requirements for the access and core network are expected to arise. When the current access and core networks are used for both the current network services and M2M services, communications equipments at the network edge need to handle a large number of communication flows which are a mix of large volume data communication such as web access and M2M data communication at the same time. To satisfy the QoS requirements of many applications including M2M applications, communications equipments at the edge of the access network will need to support minimum guaranteed rate of service which can secure the minimum data rate for each application. And sensor data communication, which is a kind of M2M communication, can be considered to send small sized packets in general because each sensor data is generally of small size. For applications such as remote fault management system which take some actions upon the information collected from the sensors, it will be necessary for packets carrying sensor data to be forwarded with less delay. To realize this requirement, access networks (especially communications equipments at the edge of the access networks) will need to support QoS functionality which allows small sized packets such as sensor data to be forwarded with low delay even when there are a lot of communication flows including large packet sized large volume data communication and small sized packet communication. These observations lead to a requirement to support both minimum guaranteed rate service and low delay forwarding service for small sized packets.

Packet scheduling algorithm is a candidate for a means to realize both minimum guaranteed rate service and low delay forwarding service for small sized packets. Many packet scheduling algorithms with O(1) computational complexity are proposed, such as Deficit Round Robin (DRR)[2], Surplus Round Robin (SRR)[3][4] and Eligibility Based Round Robin (EBRR)[5] and others[6][7]. EBRR is an extension of SRR computational complexity of which is O(1) even with a quantum less than the maximum packet size. EBRR does not try to schedule small packets prior to large packets among packets scheduled in one scheduling round. This can cause small packets to experience longer delay because they need to wait for transmission of large packets to finish. Considering the fact that delay sensitive sensor data packets are small, scheduling small packets prior to large packets is beneficial. In [8], we have proposed an extension of EBRR which can reduce the delay of a small sized packet and reported a preliminary evaluation result. In this paper, we propose a modified version of the packet scheduling algorithm in [8] which makes it possible to transmit packets smaller than a threshold before packets larger than the threshold among packets received in one scheduling round. We modified the algorithm in [8] to reduce the delay observed in the simulation results of the algorithm in [8]. The proposed algorithm can be used to reduce delay of sensor data in

access network communications equipments such as edge router and PON OLT which handles packets of various applications including sensor data and Web access at the same time. The rest of the paper is organized as follows. In section 2, we explain EBRR and the problem that the proposal aims to solve. In section 3, we describe proposed extension of EBRR algorithm and the evaluation result by simulation.

2. DESCRIPTION OF EBRR AND THE PROBLEM

To explain the problem to be solved by the proposal, we first explain EBRR algorithm and describe the problem next.

2.1. Description of EBRR Algorithm [5]

EBRR uses the following data structures. Packet queue Queue[i] to store packets for $flow_i$, Credit Counter Ci and quantum value Qi for each $flow_i$, array ActiveList[] an element of which is a list of Queue[i] which are to be scheduled in one scheduling round. The index of ActiveList[] corresponds to a scheduling round. We denote the maximum packet size as L_{max} and the size of array ActiveList[] as q. q is defined to be ceil(L_{max}/minimum of Qi), where ceil(x) is a function which returns the smallest integer which is not less than x. Queues to be scheduled in scheduling round RC (0,1,2,..) are in ActiveList[RC mod q] ("mod" means modulo). Processing in one scheduling round is as follows. Below, RC represents the current scheduling round. i) is invoked when one packet transmission ends.

i) If ActiveList[RC mod q] is empty, increment RC by 1 and start the next scheduling round and go to i). Otherwise, remove the first element (Queue[i]) from ActiveList[RC mod q]. Dequeue one packet at the head of Queue[i] and send it. Decrement Ci by the length of the packet and go to ii).

ii) If Ci > 0 and Queue[i] is not empty, append Queue[i] to ActiveList[RC mod q]. If Ci <= 0, find the scheduling round rc_i where Ci becomes more than 0 by adding Qi to Ci every scheduling round. (rc_i = RC+floor(-Ci/Qi)+1, where floor(x) is a function which returns the largest integer not larger than x.) If Queue[i] is not empty, append Queue[i] to ActiveList[rc_i mod q] and set Ci to the value at scheduling round rc_i. If Queue[i] is empty, record that packets in Queue[i] can be sent at scheduling round rc_i or later and set Ci to Qi. Goto i).

iii) When a packet for $flow_i$ is received in scheduling round RC, it is enqueued in Queue[i]. If Queue[i] was empty before enqueueing, let rc_i be the scheduling round where packets in Queue[i] can be sent (i.e. Ci at rc_i is more than 0) and append Queue[i] to ActiveList[max(rc_i, RC) mod q]. (max(x,y) is a function which returns the maximum value among x and y)

2.2. The Problem to be Solved by the Proposal

We consider a case where two flows (f1, f2) are defined and quantum for each of f1 (Q1) and f2 (Q2) is 750bytes (Fig.1). We assume that queues for f1 and f2 become ready to send packets at scheduling round 1. 1500 bytes packet (p1-1) is received for f1 and 300bytes packet (p2-1) is received for f2 in this order just before the beginning of scheduling round 1.

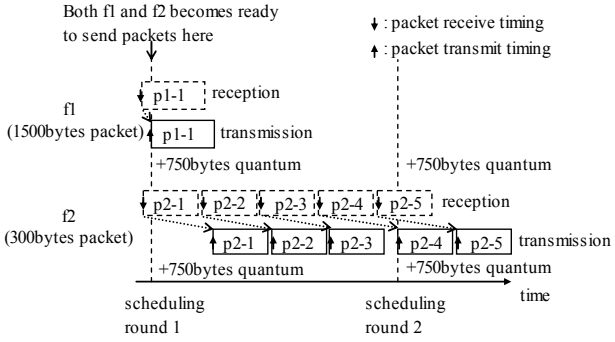


Figure 1. Example of packet scheduling by EBRR

EBRR scheduling algorithm decides to transmit p1-1 first and p2-1 next (Fig.1). This scheduling order forces p2-1 to wait for p1-1 to be transmitted on outgoing interface and to experience longer delay. If p1-1 is a packet for data application such as file transfer and p2-1 is a packet for sensor application, sending p2-1 first and p1-1 next would be preferable because p2-1 experiences shorter delay. We need some enhancement of EBRR to realize packet scheduling in which p2-1 is transmitted before p1-1 to lessen the delay experienced by a short packet.

3. PROPOSED ALGORITHM

In this section, we show the algorithm which solves the problem described in section 2. We also show the simulation results of the proposed algorithm.

3.1. Algorithm for Sending Small Packets prior to Large Packets

We modify EBRR in three ways. We proposed the first and the second modifications in [8] and we add the third one in this paper. Firstly, we split ActiveList[] into two arrays, ActiveListS[] for packets smaller than a threshold value THRESH and ActiveListL[] for packets larger than or equal to THRESH. Secondly, we define another threshold value TH which is less than or equal to 0. We decide that a packet can be transmitted if (current Ci - packet length) is larger than TH. If (current Ci - packet length) is less than or equal to TH, the packet transmission is scheduled at the scheduling round when (Ci at the scheduling round - packet length) is larger than TH. In order to guarantee O(1) computational complexity (i.e. at least one packet can be sent when Queue[i] is checked), we have to modify the way to calculate the scheduling round when a packet in Queue[i] can be sent in scheduling processing. By the first

modification, packets smaller than THRESH can be sent prior to packets which are not smaller than THRESH in one scheduling round. By the second modification, packets larger than the value determined by Credit Counter value and TH are scheduled at later scheduling round. Thirdly, we introduce maximum burst size max_burst_i configuration parameter for each $flow_i$. Credit counter of $flow_i$ (Ci) can be increased up to max_burst_i in step iii) of EBRR based on the difference between RC and rc_i. This is to allow credit counter to be incremented up to max_burst_i while waiting for a packet arrival for $flow_i$. We modify EBRR algorithm in 2.1 as follows. i') is invoked when one packet transmission ends.

i') If both ActiveListS[RC mod q] and ActiveListL[RC mod q] are empty, increment RC by 1 and start the next scheduling round and go to i'). If ActiveListS[RC mod q] is not empty, remove the first element (Queue[i]) from ActiveListS[RC mod q]. If ActiveListS[RC mod q] is empty and ActiveListL[RC mod q] is not empty, remove the first element (Queue[i]) from ActiveListL[RC mod q]. If Queue[i] is removed, dequeue one packet at the head of Queue[i] and send it. Decrement Ci by the length of the packet and go to ii').

ii') If Queue[i] is not empty and (Ci - packet length of the first packet in Queue[i]) is larger than TH, append Queue[i] to ActiveListS[RC mod q] or ActiveListL[RC mod q] depending on the size of the packet at the head of Queue[i] (i.e. if the size is smaller than THRESH, append Queue[i] to ActiveListS[RC mod q]. Otherwise, append it to ActiveListL[RC mod q]). If (Ci - packet length of the first packet in Queue[i]) is less than or equal to TH, find the scheduling round rc_i where (Ci - packet length of the first packet in Queue[i]) becomes more than TH by adding Qi to Ci every scheduling round. (i.e. rc_i = RC+ floor((packet length of the first packet in Queue[i] + TH - Ci)/Qi)+1) Then, append Queue[i] to ActiveListS[rc_i mod q] or ActiveListL[rc_i mod q] depending on the size of the packet at the head of Queue[i] and set Ci to the value at scheduling round rc_i. If Queue[i] is empty, record that packets in Queue[i] can be sent at scheduling round rc_i when Ci becomes positive or later and set Ci to Qi. (i.e. If Ci <0, rc_i = RC+ floor(-Ci/Qi)+1. If 0 <= Ci < Qi,, rc_i =RC + 1. Otherwise, rc_i = RC) Goto i')

iii') When a packet for $flow_i$ is received in scheduling round RC, it is enqueued in Queue[i]. If Queue[i] was empty before enqueueing, let rc be the scheduling round where packets in Queue[i] can be sent (i.e. If RC>rc_i, Ci=min(Ci+Qi*(RC-rc_i), max_burst_i). If (Ci - packet length of the received packet) > TH, rc = max(rc_i, RC). Otherwise, rc = max(rc_i, RC) + floor((packet length of the received packet + TH - Ci)/Qi) + 1) and append Queue[i] to ActiveListS[rc mod q] or ActiveListL[rc mod q] depending on the size of the packet. (min(x,y) is a function which returns the minimum value among x and y)

3.2. Example of Packet Scheduling by the Proposed Algorithm

We explain a packet transmission example by the proposed algorithm using the case in Fig.1. We set THRESH at 301 bytes and we also set TH at -300 bytes. For this example, we set both max_burst_1 and max_burst_2 at 0. We can get packet transmission sequence in Fig.2 by applying the algorithm in 3.1 to packet receive sequence in Fig.1 with these parameters. In Fig.2, p2-1, p2-2, p2-3, p2-4 and p2-5 are sent prior to p1-1.

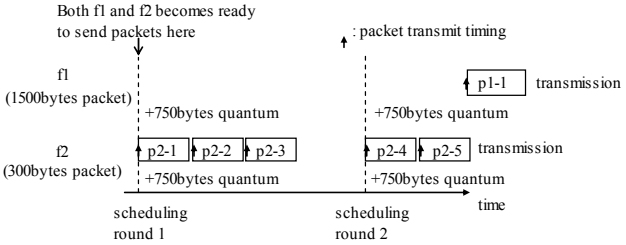


Figure 2. Example of packet scheduling

3.3. Simulation Result of the Proposed Algorithm

We have done simulations for 4 cases. In all the simulation cases, we assume that packets are received on four 1Gbps input interfaces (port1 – port4) and they are sent on one 100Mbps output port. The 4 simulation cases are explained in A), B),C) and D).

A) 333 flows of 1500bytes packets at 3.003Mbps are received on each of port1, 2, 3, which total up to 999 flows ($flow_1$–$flow_{999}$) and 1Gbps input on each of port1, 2, 3. Packets of each flow are received in a cyclic manner (i.e. on port1, packets are received from $flow_1$, $flow_2$, $flow_3$,.., $flow_{333}$, $flow_1$,..). Minimum guaranteed rate for each of those flows is 100kbps. One flow of 200bytes packets, one packet every 20ms (80kbps) is received on port4 ($flow_{1000}$). Minimum guaranteed rate for it is 80kbps. Simulated time interval is 20seconds (1001 packets for $flow_{1000}$). There is always at least one packet for each of large sized packet flow ($flow_1$-$flow_{999}$).

B) 33 flows of 1500bytes packets at 30.3Mbps are received on each of port1, 2, 3, which total up to 99 flows ($flow_1$–$flow_{99}$) and 1Gbps input on each of port1, 2, 3. Packets of each flow are received in a cyclic manner (i.e. on port1, packets are received from $flow_1$, $flow_2$, $flow_3$,.., $flow_{33}$, $flow_1$,..). Minimum guaranteed rate for each of those flows is 1Mbps. One flow of 200bytes packets, one packet every 20ms (80kbps) is received on port4 ($flow_{100}$). Minimum guaranteed rate for it is 80kbps. Simulated time interval is 20seconds (1001 packets for $flow_{100}$). There is always at least one packet for each of large sized packet flow ($flow_1$-$flow_{99}$).

C) We divide simulated time interval into 12ms intervals. 33 flows of 1500bytes packets at 1Mbps are received on each of port1, 2, 3, which total up to 99 flows ($flow_1$–$flow_{99}$) and 33Mbps input on each of port1, 2, 3. Packets

of each flow are received once at a random timing of every 12ms interval. Minimum guaranteed rate for each of those flows is 1Mbps. One flow of 200bytes packets, one packet every 20ms (80kbps) is received on port4 ($flow_{100}$). Minimum guaranteed rate for it is 80kbps. Simulated time interval is 20seconds (1001 packets for $flow_{100}$ and the number of packets in each of $flow_1$-$flow_{99}$ is 1667. There exist 1667 of 12ms intervals.). We have done the simulation 11 times with different random number seed.

D) 33 flows of 1500bytes packets at 1Mbps are received on each of port1, 2, 3, which total up to 99 flows ($flow_1$–$flow_{99}$) and 33Mbps input on each of port1, 2, 3. Packets of each flow are received at a random timing in simulated interval. Minimum guaranteed rate for each of those flows is 1Mbps. One flow of 200bytes packets, one packet every 20ms (80kbps) is received on port4 ($flow_{100}$). Minimum guaranteed rate for it is 80kbps. Simulated time interval is 20seconds (1001 packets for $flow_{100}$ and the number of packets in each of $flow_1$-$flow_{99}$ is 1667.). We have done the simulation 11 times with different random number seed.

Table 1 shows the parameters used for the proposed algorithm in each simulation case. The same values of Qi are used for the simulation of EBRR.

Table 1. Simulation parameters

	$Q_{1-99 \text{ or } 999}$	$Q_{100 \text{ or } 1000}$	THR-ESH	TH	max_burst 1-99 or 999	max_burst 100 or 1000
A)	50	40	201	-200	3000	1500
B)	50	4	201	-200	3000	1500
C)	50	4	201	-200	3000	1500
D)	50	4	201	-200	3000	1500

We set Q_1-Q_{1000} according to the ratio of minimum guaranteed rate of flows. We set THRESH at 201 and TH at 200 because the size of small sized packets is 200bytes. We determined the value of max_burst_i so that credit counter Ci can be incremented enough when appropriate.

Table 2 shows the maximum and average delay of small sized packet flow ($flow_{1000}$ for A) and $flow_{100}$ for B), C), D)) for EBRR and the proposed algorithm in each simulation case

Table 2. Maximum and average delay

	EBRR		Proposed Algorithm	
	maximum delay	average delay	maximum delay	average delay
A)	119.86ms	67.86ms	0.12ms	0.059ms
B)	11.87ms	5.95ms	0.12ms	0.059ms
C)	9.54ms	2.18ms	2.26ms	0.36ms
D)	9.15ms	1.76ms	0.12ms	0.059ms

The result in Table2 shows that the proposed algorithm is quite effective in reducing the delay of small sized packets.

Fig.3, 4, 5 and 6 show the delay distribution of each simulation case. These distributions show that most of the delay of the proposed algorithm are less than 1ms (actually less than 0.12ms)

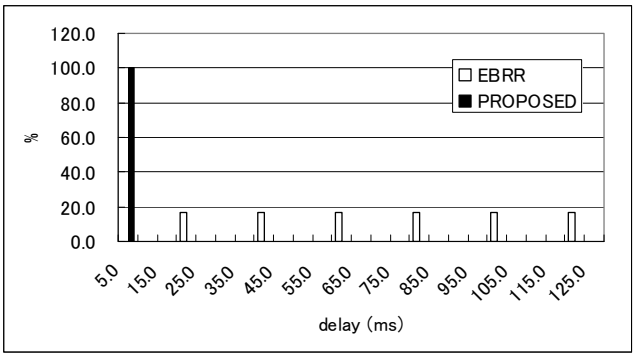


Figure 3. Delay distribution of A) (5ms interval histogram)

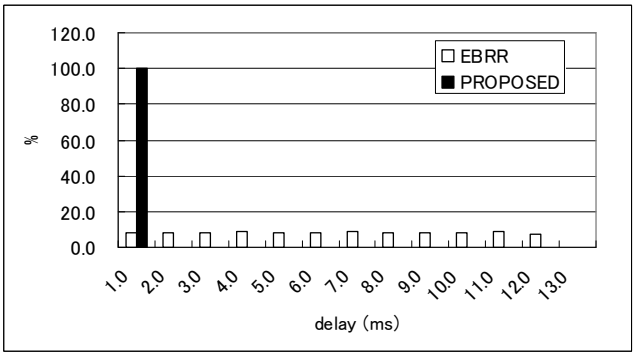


Figure 4. Delay distribution of B) (1ms interval histogram)

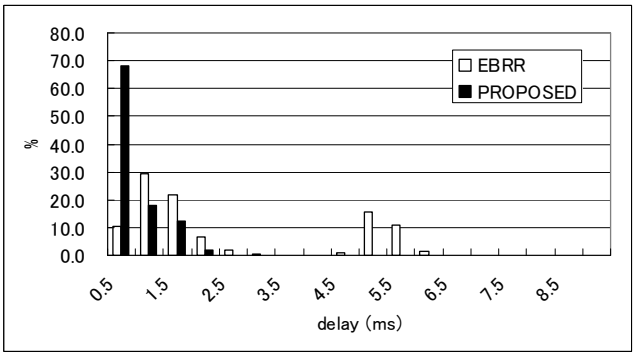


Figure 5. Delay distribution of C) (0.5ms interval histogram)

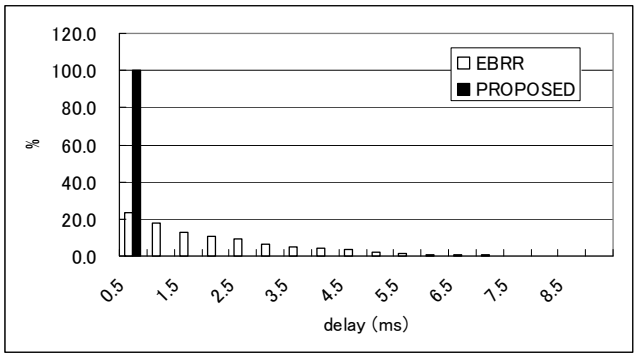


Figure 6. Delay distribution of D) (0.5ms interval histogram)

The delay of the proposed algorithm in C) is a little larger than those of other simulation cases. We think that this is because max_burst_i parameter allows large sized packets to be scheduled at an earlier scheduling round than a small sized packet and the small packet has to wait for the transmission of the large sized packets to finish and the probability of this event is higher in case C) than in case D).

4. RELATED RESEARCHES

Priority Queueing (PQ) packet scheduling algorithm is used to reduce the scheduling delay of a specified flow in real networks. PQ does not support minimum guaranteed rate service. And PQ does not allow a packet of one flow to be prioritized based on the size of the packet. The proposed algorithm supports both of them.

EBRR, which the proposed algorithm extends, is based on SRR scheduling algorithm. Aliquem[7] is a packet scheduling algorithm which extends DRR scheduling algorithm in the same way as EBRR extends SRR. The proposed algorithm extends EBRR so that a small sized packet can be scheduled at an earlier scheduling round than large sized packets. With Aliquem and DRR, a small sized packet can be scheduled at an earlier scheduling round than a large sized packet arriving at the same time if the value of (packet size/quantum value) is smaller for the small sized packet than for the large sized packet. So Aliquem can reduce the delay of small sized packets if the quantum value of each flow, which is determined by the ratio of minimum guaranteed rate of each flow, happens to satisfy the above condition. This condition does not always hold. The proposed algorithm allows parameters TH, THRESH and max_burst_i to be adjusted independent of quantum value and the delay of small sized packets can be reduced without regard to minimum guaranteed rate.

5. CONCLUSION

In this paper, we proposed an extension of EBRR algorithm which can transmit packets smaller than a threshold before packets not smaller than the threshold among packets received in one scheduling round. We also showed an evaluation result of the proposed algorithm by simulation and verified the effectiveness of the proposed algorithm in reducing the delay of small sized packets.

REFERENCES

[1] ETSI TS 102 690, “M2M functional architecture”

[2] M. Shreedhar and G. Verghese, “Efficient Fair Queuing Using Deficit Round-Robin,” IEEE/ACM Transactions on Networking, Vol.4, No.3, June 1996.

[3] D. Nikolova and C. Blondia, “Evaluation of Surplus Round Robin Scheduling Algorithm,” Proceedings of the 2006 International Symposium on Performance Evaluation of Computer and Telecommunication Systems, 2006.

[4] H. Adisshu, G. Parulkar, and G. Varhese, “A Relaible and Scalable Striping Protocol,” in Proc. ACM SIGCOMM, 1996.

[5] L. Lenzini, E. Mingozzi and G. Stea, “Eligibility-Based Rount Robin for Fair and Efficient Packet Scheduling in Wormhole Switching Networks,” Trans. On Parallel and Distributed Systems, Vol.15, No.3, 2004.

[6] S.Kanhere and H.Sethu, “Fair, Efficient and Low Latency Packet Scheduling Using Nested Deficit Round Robin”, in Proceedings of the IEEE Workshop on High Performance Switching and Routing, pp.6-10, 2001.

[7] L.Lenzini, E Mingozzi, and G. Stea, “Aliquem: a Novel DRR Implementation to Achieve Better Latency and Fairness at O(1) Complexity”, in Proc. 10th Int. Workshop on Quality of Service (IWQoS), pp77-86, May, 2002.

[8] T. Matsuda, E. Horiuchi and T. Yokotani, “A Proposal of a New Packet Scheduling Algorithm Which Can Reduce the Delay of Small Packets” in Proc. of IEEE GCCE, 2012.

DIGITAL SPACE TRANSMISSION OF AN INTERFERENCE FRINGE-TYPE COMPUTER-GENERATED HOLOGRAM USING IRSIMPLE

Masataka Tozuka†, Kunihiko Takano††, Koki Sato†, Makoto Ohki†

†Shonan Institute of Technology, ††Tokyo Metropolitan College of Industrial Technology

ABSTRACT

In this paper, we present a method to perform a digital-space transmission of an interference fringe-type computer-generated hologram using IrSimple. IrSimple was defined by IrDA technical standard, and it was developed for the purpose of sending image data at high speed using an infrared-rays. We performed infrared digital transmission using IrSimple, and we minimized the size of the transmitted file by using a suitable compression method. The size of the compressed file was very small compared with that of the bitmap file. The transmission preserved the quality of the representation while requiring a short transmission time.

Keywords— IrSimple, JPEG2000, space transmittance, interference fringe-type computer-generated hologram

1. INTRODUCTION

In recent years, camera, television broadcasting systems and communication have been digitized [1]. We consider a system for realizing 3D television [2]. We examine performing digital-space transmission of a CGH (computer-generated hologram) in real time by using infrared LEDs (Light Emitting Diodes). Visible light-space transmission using the white LED of an SSTV (Slow Scan Television System) and an NTSC (National Television System Committee System) has already been previously presented [3-6]. We used the NTSC System which is using the same signal spectrum as the bandwidth of white LED. If HDTV(High Definition TeleVision) is used, the amount of information will become huge. Therefore, visible light communications have many technical subjects.

In this experiment, we use IrSimple for digital communication [7], as IrSimple can perform infrared digital-space transmission of a suitably compressed file that contains an interference fringe-type computer-generated hologram.

We transformed the interference fringes created at 512 x 480 pixels to 256 x 480 pixels. The file size was set to 53.7 kB when compressing the file using the JPEG2000 (Joint Photographic Experts Group 2000) format [8-11]. The size of the compressed file was significantly smaller than that of the original bitmap file. Moreover, the infrared digital transmission time is short when using IrSimple, which can perform digital-space transmission with less degradation of the representation compared with analog transmission using an NTSC System.

2. THE PROPOSAL OF A SYSTEM

We consider that the same system as Figure 1 will realize in the near future. In this paper, we tried and experimented in the technology in the process of a system implementation.

A terminal transmits the data of interference-fringes type CGH with the LED light beam which was transmitted using the digital communication network and modulated by IrSimple, and we are conducting the confirm experiment for expressing a CGH reconstructed image as a terminal directly with the LED light beam used for the transmitting now. We have realized a part of the technologies in some past papers.

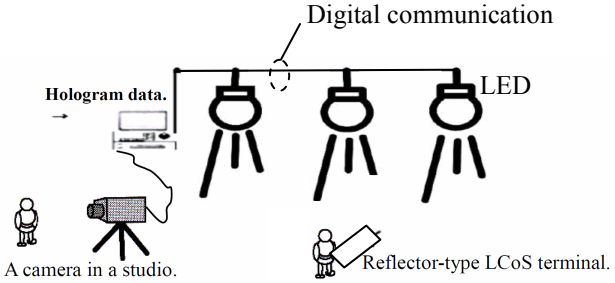


Figure 1. The overview of the system

3. THE TECHNOLOGY WHICH WE ARE USING IN THE EXPERIMENT

3.1. IrSimple.[1,7]

IrSimple is a communication method that uses infrared light and is based on the IrDA technical standard. IrSimple consists of a protocol-layer technical specification that specifies a transport layer from a data-link layer and a profile technical specification that specifies directions.

As shown in Table 1 and Figure 2, IrSimple has both uni- and bi-directional modes. Moreover, the time required to establish a connection is short and the transfer efficiency is good.

Uni-directional communication transmits data in only one direction and without the transmitter requiring a signal from the receiver. Because the receiver receives the data from the transmitter without replying, the packaging cost of the receiver is minimized.

Bi-directional communication is a conventional IrDA communications protocol and is shown in Figure 3. It uses the same communication function as IrSimple, and the

IrDA device and IrSimple device connections are compatible. The IrDA and IrSimple communication methods are distinguished by the response of the receiving device, and IrSimple can change a connection protocol into IrSimple or IrDA automatically.

IrSimple uses the following method to decrease the connection time. Conventional IrDA requires several seconds to connect each layer after searching for and discovering a partner device. However, IrSimple improved the connection sequence. It performs the search and connects each layer simultaneously, which results in a connection time that is approximately 0.1 to 0.2 seconds shorter.

Moreover, IrDA can perform a burst transmission of up to 64 kB; however, a transmission unit of size 2 kB is typically used. It communicates using a handshake as a unit, and for this reason, the communication efficiency is bad. Using the uni-directional mode of IrSimple, a burst transmission size of up to 2 GB is possible. For the bi-directional communication mode of IrSimple, the default the burst transmission size is 256 kB. Thus, using IrSimple significantly increase the communication efficiency.

Table 1. Comparison of IrDA and IrSimple.

	IrDA	IrSimple
Communication profile.	Bi-directional.	Bi-directional or uni-directional.
Connection time.	1 to 3 sec.	0.1 to 0.2 sec.
Transfer efficiency.	10 to 40%.	Not less than 90%
Communication configuration.	One-to-many.	One-to-one.

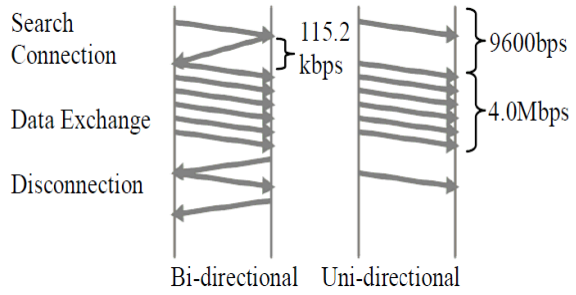


Figure 2. The IrSimple communication sequence.

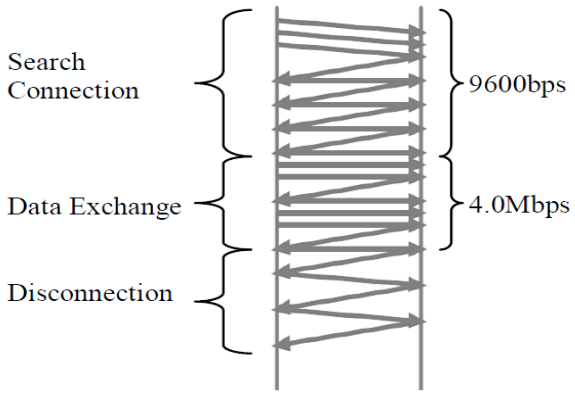


Figure 3. The IrDA communication sequence.

3.2. JPEG2000.[10-11]

The JPEG2000 encoding can be used to efficiently compress images such that the uncompressed image can be recovered from the compressed image at a high bit rate Moreover, when the JPEG2000 encoding is used, the user can specify the resolution, assign priority to specific regions of the image, and specify the error tolerance. This encoding system includes various functions that were not included previously. We compared JPEG2000 with JPEG for image quality by human eye in the same compression. It understood the following thing by this Confirm. The biggest difference between the JPEG2000 encoding and JPEG encoding is that when the compression is increased, block noise is not conspicuous. JPEG2000 has this advantage because it uses wavelet transforms in encoding. The primary steps of the encoding are shown in Figure 4.

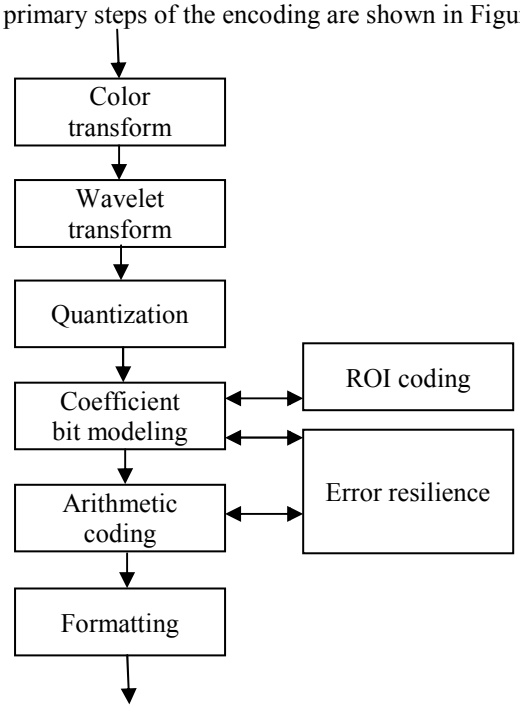
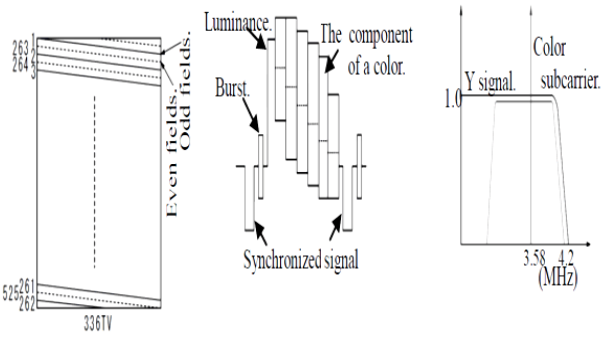


Figure 4. The primary steps of the JPEG2000 encoding method

3.3. Analog transmission using a white LED and the NTSC System.[5]



(a) NTSC scanning line (b) NTSC compound picture signal (c) NTSC Y/C signal spectrum

Figure 5. The NTSC standard

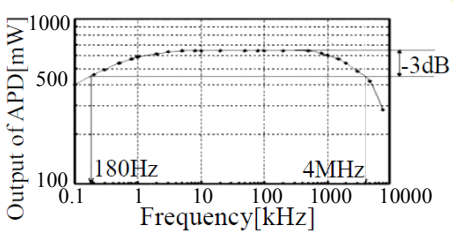


Figure 6. The frequency response of the white LED

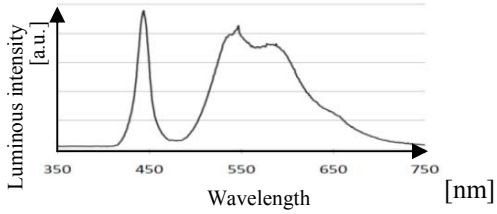


Figure 7. The wavelength distribution of the white LED

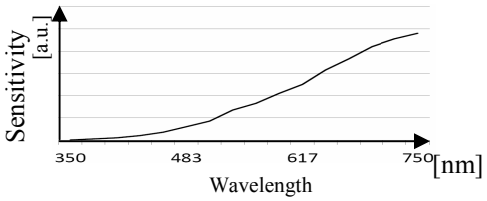
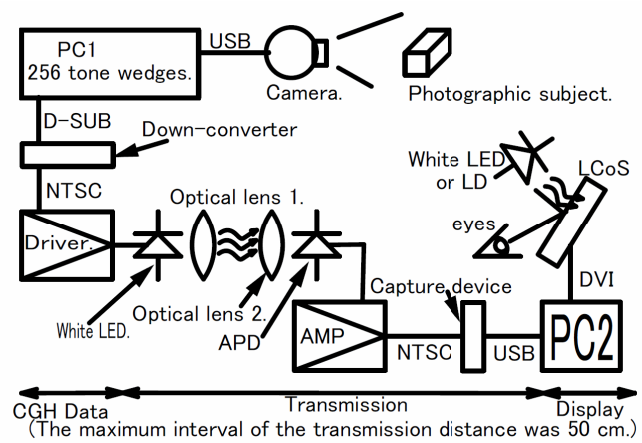
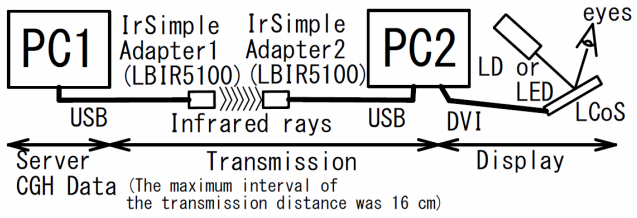


Figure 8. The sensitivity function of the APD



(a) NTSC System



(b) IrSimple System

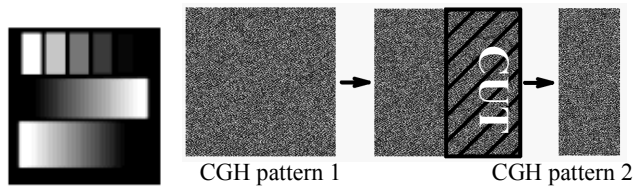
Figure 9. The spatial transmission of the experimental system

As shown in Figure 5, this experiment uses an NTSC System that has a 4.2-MHz bandwidth. As shown in Figure 6, it uses a white LED that has a 4-MHz bandwidth that starts at 180 Hz. The reason this experiment did not use HDTV was not settled in a white LED that has a 4-MHz bandwidth that starts at 180 Hz. A Phillips LXML-PWM1-100 LED is used for transmission. We modulate the light beam of the white LED, which is shown in Figure 7, to the NTSC signal and receive the modulated light by using the avalanche photodiode; the photodiode's sensitivity function is shown in Figure 8. As shown in Figure 9 (a), an NTSC signal modulates all the bandwidths of white LED, and receives in the bandwidth suitable for the photographic sensitivity of APD (avalanche photodiode). This system performs visible light-space transmittance using an analog form.

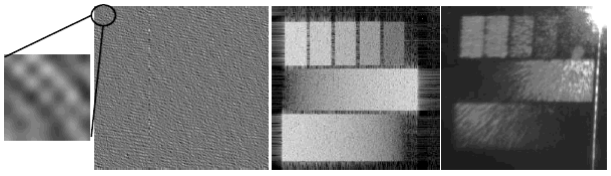
4. EXPERIMENTAL METHOD

4.1. Outline

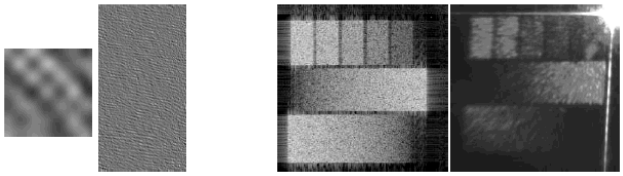
This experiment created an interference fringe-type CGH pattern using the PC (personal computer) system presented in Figure 9 (b). The reconstructed image with resolution 256 x 480 pixels is shown in Figure 10 (f). The file compressed using the JPEG2000 encoding was 53.7 kB in size, which implies that the compression ratio was 15%; the patterns are shown in Figure 11. We performed optical-space transmission from PC1(IrSimple adapter1:LB giken LLC LBIR5100) to PC2(IrSimple adapter2) using the uni-directional mode of IrSimple (see Table 2) and checked the PSNR (peak signal-to-noise ratio). We display the pattern after it was transmitted to the LCoS (Liquid crystal on silicon: see Table 3) and performed image reconstruction using the red-colored light of the laser diode(650-nm wavelength). The maximum interval of the transmission distance was 16 cm, when we experimented in infrared transmission.



(a) Input image. (b) Processing of a pattern.



(c) CGH pattern 1 (d) Simulation (e) Reconstructed image
(A corner of size 64 x 64 out of a total 512 x 480 pixels is shown.)



(f) CGH pattern 2 (g) Simulation (h) Reconstructed image
(A corner of size 64 x 64 out of a total 256 x 480 pixels is shown.)

Figure 10. The pattern before transmission

Figure 10 shows the reconstructed image in a simulation and for a laser diode based on the CGH pattern before transmission using IrSimple. The CGH pattern created by PC1 is the interference-fringe type computer-generated hologram of size 512 x 512 pixels. Figure 10 (c) of CGH pattern 1 shows the CGH pattern, which has size 512 x 480 pixels, after it is cut off the 512 x 512 to 512 x 480. Figure 10 (f) of CGH pattern 2 shows the pattern, which cut a CGH pattern 1 to one half become size 256 x 480 pixels.

We calculated the PSNR from the transmitted image, which is shown in Figure 11. The CGH pattern refers to some or all of the interference-fringe type computer-generated hologram of size 512 x 480 pixels. Figure 11 shows the CGH pattern, which has size 256 x 480 pixels, after it is transmitted and the reconstructed image in a simulation and for laser diode.

Table 2. General optical characteristics of the output

	Property (FIR method)	MIN.	TYP.	MAX.	Unit
LED	Peak emission wavelength	850	880	900	nm
	Spectral bandwidth	—	45	—	nm
	Radiant intensity (Std)	100	—	500	mW/sr
	Radiant intensity (LP)	9	—	—	mW/sr
PD	Peak receive wavelength	—	940	—	nm

Table 3. Reflective LCoS

Number	One piece
Phasing tone wedge	8 bit
Thickness of a liquid crystal	Approximately 4 [μm]
Pixel pitch	8 (H) x 8 (V) [μm]
The number of pixel	1920 (H)x1080 (V) [pixels]
Frame rate	60 [Hz]
The degree of phase modulation	Approximately 2.19π

4.2. PSNR

The PSNR of a monochrome representation can be calculated using formulas (1) through (3).

$$PSNR = 20\log_{10}\left(\frac{MAX}{\sqrt{MSE}}\right) \quad (1)$$

$$MSE = \frac{\sum_{i=0}^{x-1}\sum_{j=0}^{y-1}\{f(i,j) - F(i,j)\}^2}{x \times y} \quad (2)$$

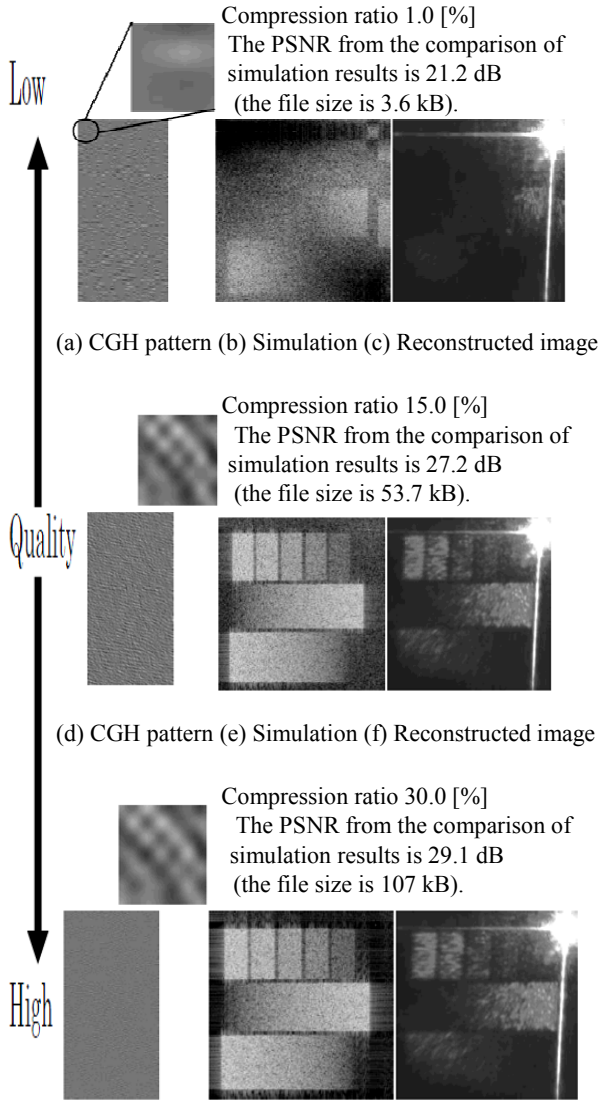
$$MAX = 2^B - 1 \quad (3)$$

The MSE is the mean squared error; f (i, j) is the pixel value of the transmitted representation at the point (i, j); F (i, j) is the pixel value of the received representation at the point (i, j); x and y are pixels; and B is the number of bits used to express a pixel's value.

Although a high PSNR value indicates high definition, a human may be unable to distinguish the original and received representation even when the PSNR is low.

4.3. Compression ratio and compression efficiency

A compression ratio is a ratio of the amount of input image datas to the amount of output image datas, and compression efficiency takes the reciprocal of a compression ratio. A compression ratio is because it consistents with the adjective which shows the merit of compression that it is high or large, and the fact of being a powerful coding machine, so that a value becomes large. If a discomfort is not felt for the word of "high compression ratio" with serving as a value small as a value, we do not take an inverse. The high compression efficiency is in order to point out a small ratio not a compression ratio but compression efficiency.



(A corner of size 64 x 64 out of a total 256 x 480 pixels is shown.)

Figure 11. The result after the transmission of CGH pattern 2

5. EXPERIMENTAL RESULTS

Degradation of the digital data transmitted using IrSimple occurs when part of a transmission is not received. A depletion of the digital data transmission of IrSimple is only success or failure of a reception of a transmittance. The digital data transmission of IrSimple is checked by calculating the PSNR relative to the uncompressed data, which are the data on the transmission side. And it could receive, it was the data as the transmission side of 0.0 with MSE. Moreover, MSE of the analog transmission of an NTSC System does not become a value of 0.0.

We created a CGH pattern with a resolution of 512 x 480 pixels using the JPEG2000 encoding. The file size of 216 kB was the largest in the experiment.

The file size was 53.7 kB when the CGH pattern was compressed at a resolution of 256 x 480 pixels, which corresponds to a compression ratio of 15%. The final file size was 1/4 of the maximum file size. The bitmap (BMP) file had size 720 kB; when the same 15% compression ratio was used, the final file size decreased to 1/14 of the maximum. The reconstructed images shown in Figure 10 (e) and Figure 11 (f) are of almost equivalent quality; thus, the compression maintained the image quality while achieving a 15% compression ratio.

When the uni-directional mode of IrSimple was used, the mean transmission time for 3 transmissions of a 768.1 kB was 2.5 sec, whereas for the compressed file with size 53.7 kB, it was 0.18 sec. We have put in practical use in the visible light communications which used the white LED of the NTSC System.

IrSimple is compared with analog transmission using the NTSC System in Table 4. IrSimple system became approximately 5.6 f/s and the IrSimple system approaching approximately 30 f/s of NTSC was created from the experiment.

Comparing Figure 12 (a) with (b) demonstrates that the analog transmission causes the representation to deteriorate.

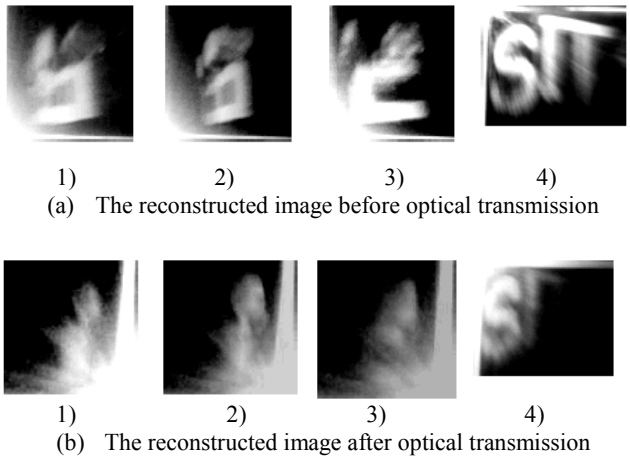


Figure 12. The CGH video transmission result when the NTSC System is used

Table 4. Comparison of frame rates

	IrSimple (f/s)	NTSC (f/s)
Frame rate	Approximately 5.6	Approximately 30

6. DISCUSSION

When analog transmission was performed using the NTSC System (the results are shown in Figure 12), approximately 1/4 of the bandwidth was required for transmitting the dot of a hologram pattern. It changed into the state where we applied the low pass filter. Moreover, the hologram pattern was expanded to correspond to an NTSC System. Because of this change, the range of observations of conditions in which an image can be reconstructed is restricted. The patterns on the transmission side and the receiving side are not similar because of degradation during the transmission.

If compression were not performed in the experiment, there was no depletion by digital transmission.

For example, when 15% of moderate compression is performed for CGH pattern, there was almost no degradation detectable with the human eye. Because the uni-directional mode of IrSimple requires only a small control signal, the transmission time was dominated by the transmission of the data file. It may be possible to perform a real-time transmission of animation using the uni-directional mode of IrSimple. Therefore, future work will also consider the transmission of an animation.

7. CONCLUSIONS

Digital-space transmittance of compressed interference fringes using IrSimple is optimal. This compression did not cause significant image degradation, when it is seen by human eyes. Moreover, IrSimple can be used to perform digital-space transmission at high speed.

In future work, we will examine real-time transmission and using IrSimple for transmission of visible light.

REFERENCES

[1] H.Naoe,F.Fukae,K.Yamaguchi,M.Matsumoto:"Ir Simple: High Speed International Standard Protocol at IrDA ",*IIEEJ Journal*, vol.35 no.5, pp.598-602,2006.

[2] M.Tozuka,K.Takano,M.Ohki,K.Sato:" An Improvement of the Characteristic of the Electronic Holography Using a Space Phase Modulation Device",*IIEEJ Journal,* vol.41 no.5, pp.554-559,2012.

[3] K.Takano,R.Wakabayashi,Y.Okamura,D.Nomura,H.Aoyama,Y.Akiyama,K.Muto,H.Suzuki,K.Shimada:"Transmission of Holographic 3D Images Using SSTV", *ITE Journal,*vol.57 no.12,pp.1770-1773 ,2003.

K.Takano,K.Muto,K.Sato,T.Lan,H.Zcho:"Transmission of Holographic 3D images using Infrared transmitter", *IEICE General Conference*,D-11-77, pp.77,2007.

[4] M.Tozuka,K.Sato,M.Ohki,K.Takano,Y.Takimoto,M.Matsumoto:"Improvement of Transmission of Hologram Data and 3D Image Reconstruction Using White LED Light",*Wireless Technology Park 2012 Journal*, pp.29-30,2012.

[5] K.Takano,M.Noguchi,Y.Kabutoya,S.Hochido,T.Lan, K.Sato,K.Muto:"An Elementary Research on Wireless Transmission of Holographic 3D Moving Pictures*",IIEEJ Journal,* vol.37 no.5, pp.645-650,2008.

[6] H.Naoe,F.Fukae,K.Sakai,S.Osawa,T.Kaminokado,T.Nakajima,K.Mameda:"Standardization of IrSimple, a High-Speed infrared Communications Protocol*", Sharp technical Journal,* vol.95, pp.63-68,2007.

[7] K.Takano, K.Sato:"Data Compression for Transmission of Holographic 3D Images Using Digital-SSTV",*IIEEJ Journal,* vol.34 no.5, pp.614-617,2005.

[8] K.Sasaki,E.Tanji,H.Yoshikawa:"Data Compression for Holographic 3D Image*", ITE Journal,* vol. 48 no.10, pp.1238-1244,1994.

[9] S.Ono,J.Suzuki:"Technology of intelligible JPEG2000",Ohmsha Co. LTD,2003.

[10] I.Ueno,E.Atsumi,F.Ono:"An Overview of the New ISO/ITU-T Still Image Coding Standard JPEG 2000", *ITE Journal,* vol.54 no.2, pp.164-171,2000.

INTEGRATED TELECOMMUNICATION TECHNOLOGY FOR THE NEXT GENERATION NETWORKS

V.I. Tikhonov, P.P. Vorobiyenko

A.S.Popov Odesa National Academy of Telecommunications, Ukraine

ABSTRACT

The paper focuses researches on next generation network (NGN) convergence. A set of comprehensive data-transfer axioms premise holistic approach to benefit diverse packet-to-circuit switching techniques. A novel dynamic flow switching (DFS) method introduced to facilitate digital telecommunication channels along with appropriate multipurpose network meta-protocol (MNP). The tenets of integrated telecommunication technology (ITT) platform developed for the transport stratum in ITU-T model of NGN. Two-dimensional quality of service (QoS) palette and related cost-to-quality ratio (CQR) function proposed for multimedia traffic control. An elastic ITT-address system originated for ITT-platform to meet the challenge of Internet-scope expansion. The paper intends to contribute future network engineering.

Keywords— next generation network, dynamic flow switching, integrated telecommunication technology

1. INTRODUCTION

The new challenges in the telecommunication market stimulate searching for enhanced technologies. In this respect, a new trend of network convergence emerged within the last decade. The ITU-T formulated the concept of network convergence in terms of the NGN directive network framework with all functions grouped in a two-layered model: service and transport stratum [1]. The reduced number of network layers is an essential feature of the NGN concept; it aims to minimize the network overhead and complexity. To satisfy the NGN framework the appropriate core and last mile technologies are required. The NGN concept implies a long-term moving process from the existing network infrastructure to the integrated network-transporting platform of the NGN. Some telecommunication companies introduced a broader concept of the NGN platform with decoupled service and application layers [2].

Deploying enhanced broadband services over existing network infrastructure results in increase of the network complexity and therefore, extra expenses on network management and maintenance. One of the most commonly used motto for the NGN perspective is “All over IP and IP over all”. The network and service convergence on the IP-basis could rather deem as an acceptable compromise towards the future and current interests of telecom companies.

However, this apparently cannot meet the increasing demands of the market because IP not originally conceived for the real time applications. There are some concerns of IP-based convergence. First, the addressing space of the currently popular IPv4 at meanwhile quite exhausted. The second issue is how to guarantee the NGN-promised quality of real time service (voice, video). These two problems seem partially mitigated via expected deployment of the enhanced IPv6 and the modified transportation technology MPLS-TP [3]. The IPv6 has an extended 128-bit address format, the Jumbo grams mode of the packet encapsulation extending the maximum transmission unit (MTU) up to the 4GiB, the flow label field in addition to the traffic class field, and some other improvements vs. IPv4. Nevertheless, a long-term transition from IPv4 to IPv6 is predicted.

Meanwhile evidently the NGN optical transport network remains the prior research direction for the telecommunication industry in the near future [4]. An important event on that way is the collaborative research project MPLS-TP launched by ITU-T and IETF joint working team (JWT) in 2008 [5]. The crucial question:”How to get a sophisticated compromise to benefit both circuit and packet switching in respect to the NGN’s promised quality of service (QoS)?” has no explicit answer yet. Therefore, more research about this aspect toward NGN is required.

The objective of this paper is to substantiate a new insight of the “packet/circuit switching dualism” in respect to the QoS demands compliant with the ITU-T framework of an NGN.

We focus the following three questions below:

1) What is the “packet” as a basic notion of digital data exchange in telecommunications?

2) What is the “true packet-switching” and “true circuit-switching” as two boundary points on the topological scale of data exchange status?

3) What is the “permanent service palette” connecting the two boundary points: packet-switching and circuit-switching (from both customer and network technology points of view)?

When studied these questions, we draft the concept of an advanced sophisticated NGN-aware technological platform (denoted as integrated telecommunication technology, or ITT).

2. ITT AXIOMATIC BASIS

In this section we provide a set of axioms to determine the data packet in the ITT-platform. Generic property of packet-switching techniques in diverse telecommunication technologies (Ethernet, Frame Relay, ATM, IP etc.) is statistical time division multiplexing (STDM). The STDM implies the physical communication channel division into an arbitrary number of digital streams mixed in a random order. The STDM provides sufficient link utilization improvement and therefore, in the context of NGN, it is a concurrent alternative to fixed sharing of a channel like conventional time division multiplexing (TDM). Any packet-switching technology has its own unique PDU − protocol data unit (e.g. Ethernet frame, IP-packet, ATM cell, UDP segments etc.).

Here we formulate a pervasive definition of an abstract digital "packet" with minimal confines, satisfying a wide spectrum of PDU frameworks and being extendable to the TDM circuit-switching techniques. We define an abstract digital "packet" as a labeled capsule constituted by two segments: packet header (PH) and packet body (PB). The PH acts as a labeled shell of the digital capsule. The solely presumed requirement for PH structure is being distinct in digital stream of a channel, as well as being related to some specific packet handling mechanism. The PB segment conceivably involves a variety of realizations with minimal restrictions.

Axiom 1: *In ITT-platform, an extra distinct attribute A attached to an arbitrary non-drop digital data segment D forms a labeled capsule C, where A, D and C consider being PH, PB and PDU, if some A -aware mechanism of C -handling exists.*

From this point of view, a distinct data byte of a TDM-container (like an STM module of SDH) carries a minimal non-drop amount of information. Any distinct byte of an STM is identified due to its unique time-slot allocation aligned with the STM-synchronization benchmarks. Therefore, a TDM-distinct byte looks like a primitive PB of an abstract "packet" (see axiom 1) related to its time-slot allocation property, in combination with the STM de-multiplexing procedure.

Both common logic mind and practical experience testify the following phenomenon: the less PB and PDU are the faster data interaction of communicating parties is available. The PB increase may result in latency of multimedia data playing back. Eventually it suppresses the overall dynamics of party's interaction in real-time mode. The interaction dynamics is critical issue in voice communication as well as in distributed automatic control systems. On the other hand, in some real time application (e.g. streaming video data) the bulk transportation in large-sized packets is more expedient. Consider these arguments we postulate the axiom 2 that specifies PB-length boundaries. This modus tolerates both dynamic real time interaction and efficient bulk transportation of digital data.

Axiom 2: *In ITT-platform, the minimal PB length is one informational byte; the maximal PB length is not limited; the bit length of the information byte depends on adopted agreements; meanwhile one byte commonly refers to 8 bits (one octet), however, it may extend this value.*

The market exposed integrated circuits (IC) support transmission of octet-units through optical and cooper wire links. These octet-units circulate the channel being shaped in 10-bit array on the low physical sub-layer [6]. Based on these IC an array of network interfaces with one-octet PDU exchange is commercially available. This fact premises the axiom 2 capability.

The PH-to-PDU-length ratio is a critical moment in various packet switching techniques. Any PH of a PDU carries control data needed for PDU handling. However, the PH control data impairs the PDU utilization. Apparently it is no use of carrying small PB amount (capsule payload) within an enormously PH-extended capsule. On the other hand, scanty packet header constrains the packet maintenance ability. Hence, no one packet header of fixed size uniformly satisfies diverse application requirements; therefore, dynamically adaptable PH needed. We formulate this idea in the axiom 3.

Axiom 3: *In ITT-platform, the PDU packet header (PH) tolerates dynamically changed PH-structure and length; the minimal PH-length is 1 byte (of 8 bits or more); the maximal PH-length is not limited.*

For instance, the ATM cell of 53-octet is an example of a compact fixed length packet container. It has a fixed length 5-octet header and 48-octet body. The ATM cells are worthwhile for transporting PB-data amount of about 40 bytes length through the pre-established permanent virtual circuits (PVC). Let L_{PB} the length of PB, L_{PDU} the length of the PDU measured in octets. The maximal encapsulation factor of ATM cell (denoted as μ_{ATM}) results

$$\mu_{ATM} = \frac{L_{PB}}{L_{PDU}} = \frac{48}{53} = 0.906. \quad (1)$$

The $\mu=0.906$ is a quite appropriate option for an arbitrary transportation container. However, to achieve the best quality of voice communication in a packet-based network, it is mandatory that any detected voice sample (e.g. of one octet size) is transmitted immediately with no latency. Another QoS-critical real time application refers to high dynamic automatic control systems running over distributed sensor networks. For a PB of one-octet length (due to the axiom 1) the ATM encapsulation factor results $\mu_{ATM}=0.019$. Such value will merely satisfy any network design. The axioms 2 and 3 result the minimal PDU-length in the ITT-platform of 2 bytes (1 byte of PH plus 1 byte of PB), as well as minimal encapsulation factor $\mu=0.5$ for the most dynamic interaction (with 1-byte PB).

The dynamically modified PH-length adopted in the ITT-platform may cause the following issue: how to recognize the PDU boundaries within a serial PDU-flow? The utilization of a special packet delimiter symbol may solve this problem. Mentioned above commercially available IC allow generating special control symbols imbedded in 10 serial bit-array on the low physical sub-layer. Thus, one of those control symbols may act as PDU-delimiter in the ITT digital flow. The copper twisted pair channels mainly exploit the pulse code modulation (PCM). To avoid

electrical damages caused by the electric potential difference of linked parties, the network interfaces maintain input/output transformers (IOT). The coupling of PCM and IOT is susceptible to the mean value of the serial digital signal (DC-offset problem). Special coding methods could moderate this problem mapping an arbitrary data set into the pseudo-random consequence (with near-to-zero mean value within the certain time-interval). Therefore, the maximum transmission unit (MTU) of PDU in a PCM system is typically limited. Thus, the packet sequence transmission over the PCM-channel implies the segregation of any digital flow into variety of distinct frames; these frames are separated due to the inter-frame gaps (IFG) intended to pacify the DC-offset. In contrast to the copper links, the optic fiber systems with the amplitude shift keying (ASK) do not suffer the DC-offset problem. Consider the NGN optic fiber perspective we postulate the following axiom 4 to provide particular PDU delimiter.

Axiom 4: *In ITT-platform, no inter-frame gaps (IFG) facilitate the PDU recognition; the single PDU is distinguished due to the special delimiter-symbol (DS) of one-byte length; the DS is encoded by 10 serial bits on the low physical sub-layer in the ITT-platform.*

Let T the total length for both PDU and delimiter symbol (counted in bytes). We define the packet-switching efficiency s as the following ratio:

$$s = \frac{L_{PB}}{T}. \qquad (2)$$

If PB=1 byte, then the minimal packet-switching efficiency s in ITT is about 0.333.

It is clear, that small PH (of 1÷4 bytes) provide few capabilities for packet handling. Therefore, the ITT packet headers of 1÷4 bytes presumably support a particular array of high QoS services (e.g. CBR voice virtual connections, automatic control system application etc.).

To maintain the PDU with dynamically modified PH length (see axiom 3) a specific mechanism of PH recognition needed. This addresses the next axiom.

Axiom 5: *In the ITT-platform, the PH-structure is a self-extracting framework driven by the first PH-byte.*

The set of axioms 1÷5 substantiates an abstract data packet framework compliant with general idea "all services over packet-switching networks" that is common in NGN approach. At the same time, these axioms notably extend the scope of the packet-switching paradigm transforming it into the blanket concept of statistical time division multiplexing (STDM) that ignores its particular realizations.

Another crucial aspect of "circuit-switching vs. packet-switching" in the NGN-architecture is the known drawback of IP-networks referred as unpredictable packet time delay, which may result packets jitter. That is to say: how to meet the NGN challenge of high QoS performance in a wide range of customer demands? Is it possible to guarantee the circuit-switching quality in a packet-switching NGN? Next section focuses this problem.

We assume the customer's behavior to search in the best "cost-to-quality ratio" (CQR). The CQR is eventually individual and varies dynamically within a short time. Imagine the one-dimensional CQR as the function $p(s)$, $s=0,1,2,\ldots, N$ where s is type of service (ToS) for given class of services with average bit rate $r=8Kbyte/s$ that is guaranteed on the time-interval $\Delta t(r,s)$ dependent on r, s; $p(s)$ is the cost of s.

Let b the guaranteed byte amount within the time-interval $\Delta t(r,s)$. We define the best quality ToS for $s=N=15$ implying the constant bit rate (CBR) data transfer of $8Kbyte/s$. That means, for any time-interval $\Delta t=1/8Khz=0.125ms$ the one-byte data transfer is guaranteed. Next, we will double Δt each case of decreasing s: $\Delta t(r=8,s=14)=0.25$ ms, $\Delta t(r=8,s=13)=0.5ms$ and so on. Any case of s results the same average bit rate that is guaranteed on diverse intervals Δt: *1byte/0.125 ms=2byte/0.25 ms=4byte/0.5 ms=… = 8Kbyte/s.*

The ordered collection of variables (s, p, Δt, b) exposes the tab.1. The first two rows in tab.1 present the function $p(s)$. The last two rows in tab.1 refer guaranteed time intervals Δt and correspondent values of byte amount b. The column $s=0$ in tab.1 is reserved for implicitly guaranteed average bit rate of $8Kbyte/s$ with no specific time-interval Δt (referred as available bit rate, or ABR transfer mode). The ABR is character to packet-switching technique, therefore we consider the $s=0$ column in tab.1 be the true packet-switching phenomenon.

The $s=15$ case is inherent to circuit-switching technique with CBR transfer mode. Therefore, we consider the $s=15$ column in tab.1 be the true circuit-switching phenomenon. The $q(r=8, s=15)$ type of service in tab.1 guaranties the constant bit rate of $8Kbyte/s$ that is compliant with conventional digital telephony standard of $64Kbit/s$ in full duplex communication.

The columns of $s=1\div14$ we denote as intermediate flow-switching cases referred to variable bit rate, or VBR transfer mode. All the cases of $s=0\div15$ we unite into generalized class of dynamic flow switching for given average bit rate of $8Kbyte/s$. The two rows in tab.1 (s and Δt) reflect a discrete function $\Delta t(s)$ denote as one-dimensional "QoS-palette" $q(s)=\Delta t(s)$. The set of 16 values of s in tab.1 presents the s-axis of the QoS-palette. This palette we define as the circuit/packet switching dualism in respect to the network transport layer (NTL) of the ITT-platform.

Onwards we construct a two-dimensional QoS-palette as a $q(r,s)$ function for different average bit-rate values r. Let $\{r_m\}$, $m=1, 2, \ldots, 15$ the set of service-classes with predefined average bit-rate values r_m. For any given r_m the one-dimensional QoS-palette $q(r_m,s)$ is possible, fig.1. The gradient-vector in fig.1 is directed from the left-down corner (low-cost service location) to the right-upper corner (best-quality of service location).

All the upper case items ($s=15$) of the QoS-palette in fig.1 present the constant bit-rate of data transfer, or CBR-domain for diverse average bit rate values (ranged from $r_1=64Kbit/s$ to $r_{15}=1Gbit/s$). The CBR transfer mode is an essential property in conventional circuit-switching technique. Reciprocally, all the down-case items ($s=0$) of the QoS-palette in fig.1 present the available bit rate of data

Table 1. One dimensional CQR-table for 8Kbyte/s average bit rate

s	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*
p	p_0	p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8	p_9	p_{10}	p_{11}	p_{12}	p_{13}	p_{14}	p_{15}
Δt (ms)	-	2048	1024	512	256	128	64	32	16	8	4	2	1	0.5	0.25	0.125
b	-	16384	8192	4096	2048	1024	512	256	128	64	32	16	8	4	2	1

transfer, or ABR-domain which is natural to conventional packet-switching technique.

The *s=1÷14* cases in fig.1 refer to ITT-specific intermediate flow-switching mode of data transfer (VBR-domain). The entire two-dimensional QoS-palette in fig.1 reflects dynamic flow switching (DFS) concept developed in this paper.

The particular values of average bit-rates r_m and time-intervals Δt for any *s* are solely exposed for an example; actually, they may vitally adopt an experience. A fairly designed QoS-palette will evidently comprehend a wide spectrum of customer demands, whereby the minimal-capacity kit of QoS items needed. Following this approach the two-dimensional cost-to-quality ratio table, or CQR-function *p(r,s)* may be designed to help customers in benefit the QoS-palette at any current moment. The previously discussed concept of two-dimensional QoS-palette we formulate in axiom 6.

Axiom 6:

In ITT-platform, the two-dimensional circuit/packet switching QoS-palette of virtual connections is available on the network transport layer; the lowest QoS-type is associated with "true packet-switching" mode of dynamic flow switching (DFS); the highest QoS-type means 'true circuit-switching' mode of DFS; the QoS-types $q(r,s_m)$, $m=1\div(s_{max}-1)$ *refer to the ITT-specific "intermediate flow-switching" mode of DFS for any given average bit rate r.*

To experience the packet/circuit switching benefits in QoS-palette, the ITT-platform provides two-dimensional cost-to-quality ratio, or CQR-function $p(q(r,s_m))$ related to the QoS-palette.

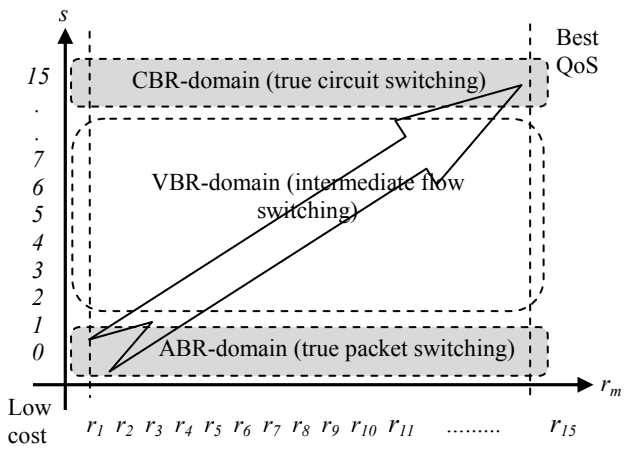


Figure 1. Two-dimensional packet/circuit switching QoS-palette

The axiom 6 justifies duality concept for the packet/circuit switching techniques: no one of them is predominant, yet both are potentially important. Another to say: not excluded "Or" (circuit-switching XOR packet-switching) but dynamically achieved compromise based on the CQR will presumably benefit the two spoken above opposite categories in the context of the more distant future generation networks. The axiom 6 compliant technology implies a quite diverse and granulated service palette enabling customer's dynamic migration within a wide range of virtual connection services. The visual image of the axiom-6 depicts fig.1.

3. THE ITT-PLATFORM FRAMEWORK

The spoken above axioms 1÷6 form theoretical basis of integrated telecommunication technology (ITT) as an alternative novel platform to facilitate more distant future generation networks. The ITT-platform intends to support transport function for end-point network devices in accordance with the ITU-T model of NGN. Nevertheless, more benefits of the ITT-platform expected in the access and aggregation network layers ranged from the end-user terminal devices (TD) to the edge of the core transportation network (CTN), fig.2.

Onwards we define in more details the two sub-layers of ITT-platform. The CTN implies the circuit-switching tunneling technology (like MPLS-TP) that is transparent to the tunneled traffic. It may be either IP (v4, v6) or any other packetized digital flow (e.g. ITT-traffic). The ITT-network considers being an open domain: any ITT-node may have a local environment with terminal devices (TD), as well as any ITT-edge node may extend the scope of ITT-domain. Therefore, an ITT-network device embraces the following two sub-layers:

1) Physical link layer (PLL) for data exchange between the two ITT-network adaptors of the adjacent nodes. The postulated PDU of the PLL sub-layer is *byte-quanta* (of 8 bit or more) as information product-primitive. We denote this primitive as an abstract "*letter*" of the transmission grammar-formalism. The one-byte PDU of the PLL sub-layer is compliant with axioms 2 and 3 spoken above.

2) Network transport layer (NTL) for data exchange between network devices within an ITT-domain. The postulated PDU of the NTL sub-layer is *packet-quanta* compliant with axioms 1÷5, namely: various PDU of the NTL sub-layer drive variable-sized PH and PB (of one byte and more with no upper border). We denote this packet-quanta as an abstract "*word*" of the transmission grammar-formalism.

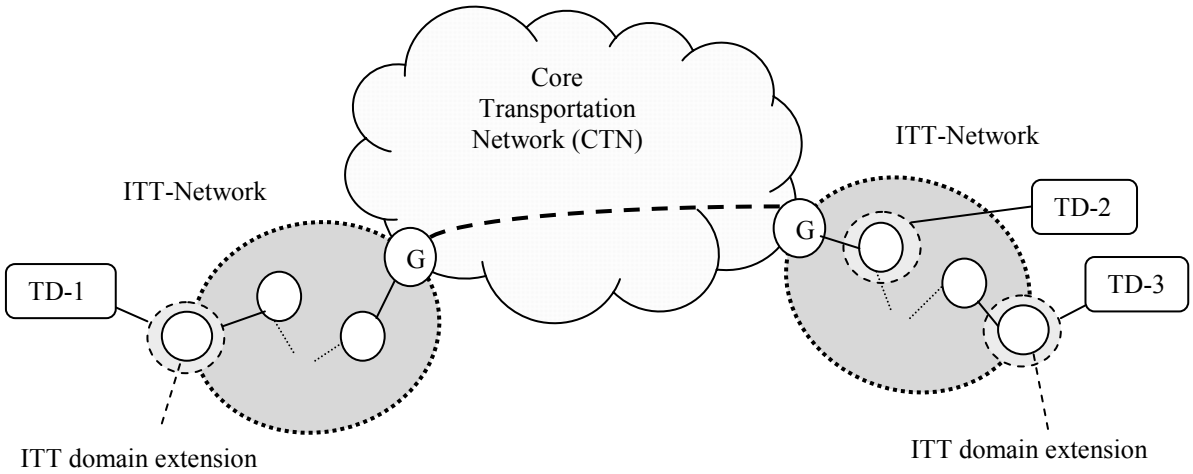


Figure 2. The ITT open network architecture

The ITT concept of a flexible data packet framework enables quite diverse palette of intermediate packet-switching processes, covering the range between the packet-switching and circuit-switching data transfer (compliant with axiom 6). In accordance with axiom 6, the ITT-platform intends to benefit two alternative switching techniques in telecommunications: packet-switching (PS) and circuit-switching (CS). The coupling of two diverse switching methods (packet /circuit switching dualism) is the background of the ITT-platform.

To maintain the complete QoS-palette of packet/circuit switching (spoken in the axiom 6) an appropriate algorithm of ITT-signaling and resource reservation is investigated along with the correspondent packet-queues procedures (out of this paper scope). The synthetic notion of the packet/circuit switching dualism of the ITT-platform we denote as *dynamic flow switching* (DFS). That means the DFS-capable switching technique enables the customer-demand vital migration within the QoS-palette framework (see fig.1).

To enable the DFS, an original multipurpose network meta-protocol (MNP) was developed for the ITT-platform based on the axioms 1÷6, [7]-[9]. The MNP solely specifies a set of generic rules for DFS, enabling a great deal of MNP-profile protocols sufficient in particular applications. In fact, the MNP implements the packet-switching technique in a wide range of packet formats. In contrast to the conventional packet-switching methods (like IP, FR, ATM etc.), the MNP does not require any packet header to deliver comprehensive control data for packet handling. Instead, the control data for currently handled packet may previously drop in several preceding packets. For that purpose, the special flow control table (FCT) provided in MNP. The FCT enables dynamic storage of an open set of the distinct packet header options. The two identical FCTs act on both emitting and receiving adjacent network nodes, fig.3. The multipurpose network meta-protocol (MNP) may alternate the conventional packet switching technique in an ITU-T compliant NGN.

The tab. 2 depicts FCT-structure. The amount of FCT-options is open and not size-restricted. The index-number of any particular MNP-profile determines the necessary collection of FCT-options for any type of service. The PH of any packet sent by Node-1 towards Node-2 can modify several options of FCT-2-In (see fig.3). Therefore, a packet-bulk of identical PH may stream the channel via expanded ITT-packet with no abundant repetition of the PH data. The tab.2 illustrates the two diverse profiles of MNP. Profile 1 embraces a typical set of control data inherent in conventional IP-packet networking; this profile is responsible for true packet-switching mode in the ITT-platform. Instead, the profile 2 includes solely one control option (flow label). This option provides a wide range of packet/circuit switching techniques in the ITT-platform in accordance with QoS-palette (fig.1).

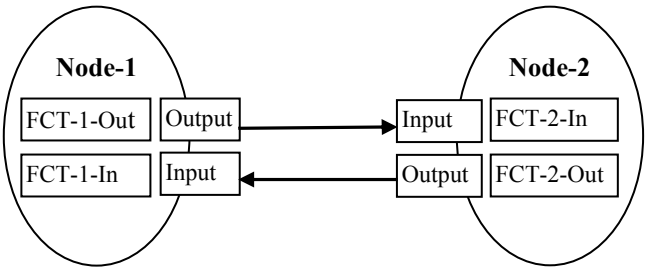


Figure 3. The ITT-platform flow control table linkage

The low redundant ITT data streaming may cause resilience issues if some input FCT corrupted. Therefore a special recovery mechanism provided in the ITT-platform: regularly the correspondent output FCT-checksums are transmitted by any network node towards their linked neighbor parties. For instance, if FCT-1-out checksum does not match FCT-2-in checksum, then Node-2 requests Node-1 for complete FCT-retransmission.

The ITT axiom 3 prompts how to harness generic IP-addressing concern. The restrained pool of conventional 32-bit IPv4-addresses eventually hinders the Internet-scope expansion. At the same time, the NGN-promised

implantation of 128-bit IPv6-addresses aggravates the today superfluity of IP-packet header in multimedia applications.

To meet the NGN challenge of IPv6-address surplus, an elastic ITT-address system developed in the ITT-platform [10]. This system implies open multilayer (global and local) identification of vast network devices due to the recurrent address-coding mechanism driven by the first address-byte. The ITT flow control adopts instantly modified packet-addresses ranged from 1 to 16 byte-length. Herewith, skimpy address options (of 1 to 4 bytes) are predominantly used in real-time applications of high QoS-demands (e.g. voice communication). Instead ample addresses (of 5 bytes and more) are actualized as needed in less critical applications (e.g. video or data streaming). The flexible network-objects addressing improves the overall channel utilization.

Table 2. The flow control table structure

MNP Profile	Source Address	Destina-tion Address	Type of Service	TTL	Check Sum	Flow Label	…	…	…
1	x	x	x	x	x				
2						x			
…									

4. CONCLUSION

The study addresses the NGN researches focusing the network transport stratum. A novel integrated telecommunication technology (ITT-platform) is developed to contribute in advanced telecommunication networks engineering. The background of ITT-platform is an original approach to digital flow segmentation with dynamically changed packet header/body length and structure. The theoretical basis of the ITT-platform is constituted by six axioms.

The axioms 1÷5 substantiate an abstract packet notion tolerant to particular technological solution. Hereby, the two ITT sub-layers grounded: physical link layer (PLL) drives distinct-byte PDU exchange for any two adjacent ITT-network nodes; network transport layer (NTL) provides diverse multi-byte PDUs data exchange for ITT-domain nodes. The NTL communication is controlled due to the flow control tables (FCT) relevant for any couple of adjacent network nodes.

The axiom 6 formulates the ITT-concept of dynamic flow switching (DFS); it premises the two-dimensional service palette to benefit the packet/circuit switching compromise.

To progress the ITT-platform a framework of the multipurpose network meta-protocol (MNP) is designed. It determines generic rules for various profiles of particular transmission protocols. The MNP-protocol declares minimal requirements for its profiles. The openness of the MNP-profiles family forms a solid background for sustainable network evolution towards ITU-T adopted NGN architecture.

To meet the NGN challenge of Internet-scope expansion, an elastic addressing system developed in the ITT-platform based on the recurrent address-coding mechanism.

The ITT-platform originates clear and simple packet-based networking model. It reduces the header abundance inherent in conventional IP-networks. The MNP protocol of the ITT-platform aims to alternate the IPv6 networking technology in the context of the more distant future generation networks.

5. LIST OF ACRONYMS

ABR – Available Bit Rate
ASK – Amplitude Shift Keying
ATM – Asynchronous Transfer Mode
CBR – Constant Bit Rate
CQR – Cost-to-Quality Ratio
CTN – Core Transportation Network
DC – Direct Current
DFS – Dynamic Flow Switching
DS – Delimiter Symbol
FCT – Flow Control Table
IC – Integrated Circuit
IFG – Inter Frame Gap
IOT – Input/output Transformer
IP – Internet Protocol
ITT – Integrated Telecommunication Technology
JWT – Joint Working Team
MNP – Multipurpose Network meta-Protocol
MPLS-TP – the Transport Profile of MPLS
MTU – Maximum Transmission Unit
NGN – Next Generation Network
NTL – Network Transport Layer
PB – Packet Body
PDU – Protocol Data Unit
PH – Packet Header
PLL – Physical Link Layer
PCM – Pulse Code Modulation
PVC – Permanent Virtual Circuit
QoS – Quality of Service
SDH – Synchronous Digital Hierarchy
STDM – Statistical Time Division Multiplexing
STM – Synchronous Transport Module
TD – Terminal Device
TDM – Time Division Multiplexing
TOS – Type of Service
UDP – User Datagram Protocol
VBR – Variable Bit Rate

REFERENCES

[1] "ITU-T Recommendation Y.2001: General overview of NGN", http://www.itu.int/rec/T-REC-Y.2001-200412-I/en (visited on 2012-11-21).

[2] A.Yakupov. “Cisco IP NGN: servisnyi uroven i upravlenie IP-adresatsiei v setiah operatorov mobilnoi i fiksirobannoi sviazi”, http://www.ciscoexpo.ru/expo2011/downloads/materials/sp/D2_NGN-BB-CNR_aryakupo.pdf (visited on 2012-11-21).

[3] “Understanding MPLS-TP and Its Benefits”, http://www.cisco.com/en/US/technologies/tk436/tk428/white_paper_c11-562013.pdf (visited on 2012-11-21).

[4] Y. Nakano, M. Takihiro, Y. Fukashiro, M. Mizutani, “Optical Network Systems for Next-generation Networks”, *Hitachi Review*, Hitachi, February 2009, http://www.hitachi.com/rev/archive/2009/__icsFiles/afieldfile/2009/02/24/r2009_feb_002.pdf (visited on 2012-11-21).

[5] “JWT Report on MPLS Architectural Considerations for a Transport Profile”, http://tools.ietf.org/html/draft-bryant-mpls-tp-jwt-report-00 (visited on 2012-11-21).

[6] Cypress Semiconductor, “EZ-USBR Technical Reference Manual”, http://www.cypress.com/?docID=27095 (visited on 2012-11-21).

[7] P.P. Vorobiyenko, V.I.Tikhonov, Patent 56395: “Sposib dinamichnoi komutatsii potokiv v telekomubikatsinih merezhah”, Ukraine, bulletin 1, 2011-01-11.

[8] P.P. Vorobiyenko, V.I.Tikhonov, Patent 46762: “Sposib pobudovi telekomubikatsinih paketnih merezh z dinamichnou adresatsieu vuzliv”, Ukraine, bulletin 1, 2010-01-11.

[9] P.P. Vorobiyenko, V.I.Tikhonov, Svidotstvo pro reestratsiu avtorskogo prava na tvir 29486: “Bagatofunktsionalni merezhni meta-protokol”, Ukraine, 2009-07-20.

[10] P.P. Vorobiyenko, V.I.Tikhonov, I.V. Smirnov, U.I. Sopina, “Algoritm dinamicheskoi adresatsyi obyectov telecommunicatsionnoi seti”, Sb. “Tsifrovyie tehnologyi”, № 8, 2010, p.11÷18.

RESEARCH ON ICT SERVICE ENERGY IMPACT ASSESSMENT METHOD: HOW MUCH ENERGY TO MANUFACTURE A CHIP

S.Schinella[*†], *D.Marquet,S.Le Masson, T.Tanaka*

France Télécom - Orange Labs
Dpt Research for Energy and Environment
38 rue du Gal Leclerc
92130 Issy-les-Moulineaux, FRANCE

X.Chavanne, J.P.Frangi

Université Paris Diderot - IPGP
Dpt Dynamique des Fluides Géologiques
35 rue Hélène Brion
75013 Paris, FRANCE

ABSTRACT

Telecommunications are expected to reduce the energetic impact of human society through dematerialization. But in addition to their utilization consumption, the ICT equipments are responsible of energy consumption for their fabrication. To assess as precisely as possible their consumption and the gain compared to physical services, we use a new modular and open method intended to reduce errors and highlight possible improvements.

The lithography phase of chips manufactured from extra-pure silicon wafers is responsible for about 70% of the consumption of chips fabrication. The main elements of this phase are the tools, which make the operations, and the air circuit, which cleans the air and control its levels of temperature and humidity 24h/24. This element is deeply analyzed in this paper, as a key module of the method which can be reused for other studies.

Many parameters take part in the air circuit consumption, particularly the climate of the place where the factory is built, the class of the clean room and the amount of particles emitted. We try to put them ahead to understand this consumption and to know how to improve the energy efficiency.

Keywords – *ICT materials, chips fabrication, energy efficiency, life cycle*

1. INTRODUCTION

Telecommunications are expected to be able to reduce human energy impact with substitution to physical services, but their own consumption is a preoccupation of international institutions like IEA and ITU [1, 2]. These agencies care about both the electric consumption in use of ICTs, and the large amounts of energy consumed during the manufacturing phase, which includes the study of this paper. To assess the real energetic impact of telecommunications, we need to consider their whole life cycle. LCAs already exist, but they are difficult to adapt to a service because of many hypotheses made and parameters hidden. It is a method which does not allow comparison to evaluate the substitution potential. This is why we use an other method explained part 2.

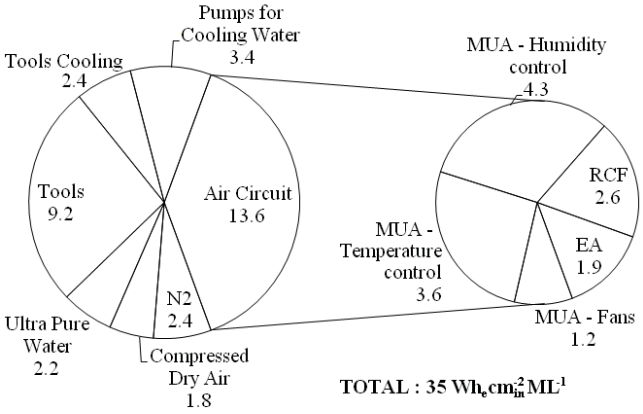


Figure 1. Principal electrical consumptions of the Hu 2008 factory [3]. The consumption due to purification of chemicals is not taken account. Consumptions are given in $\mathrm{Wh_e cm_{in}^{-2} ML^{-1}}$ (The unit "ML" is the number of mask layers necessary to manufacture a chip, characteristic of the number of operations in the clean room and the complexity of the chips. For the most powerful technologies, the number of mask layers is generally greater than for light technologies.)

For computers involved in telecommunications, the manufacturing phase seems to be the most energy-consuming phase [4]. In this phase, we separate the manufacturing consumptions of each element of the computer, and it appears that the energy necessary to manufacture the chips seems to be the highest [4]. This consumption results mainly from the lithography phase, a surface treatment where CMOS are lithographed on pure silicon wafers in a clean room [5]. This is the system analyzed in this paper schematized figure 2 and divided in all the subsystems identified by S.C.Hu [6, 3].

A benchmark of all the consumptions of these subsystems is given figure 1[1], but there are differences between all the fac-

*This paper is a part of a PhD work between France Télécom - Orange Labs and the Université Paris Diderot. It is the CIFRE convention n ° 1535/2011 established by the ANRT.

†To contact the author: sebastien.schinella@orange.com or +33.1.45.29.62.49

[1]We give the unit of energy in Wh, more common for our study (1 Wh = 3600 J). We differentiate all the kinds of energy: a Wh of electricity is noted $\mathrm{Wh_e}$, a thermal Wh is noted $\mathrm{Wh_{th}}$

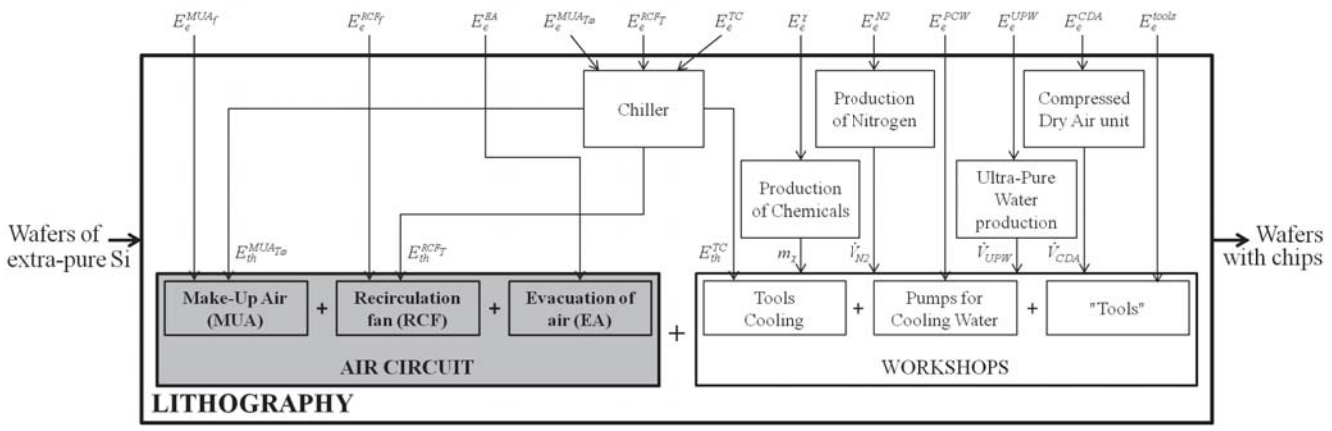


Figure 2. Schematization of the processes of the lithography factory [6, 3]

tories studied by Williams and Hu [5, 6, 3], and we also noted differences with another factory based in France. Our goal is to point out the parameters which cause these differences.

We separate the system in three subsystems:

- The air circuit (in grey in figure 2), which is potentially the most energy consuming subsystem, has three sub-subsystems:
  - The Make-Up Air unit (MUA), detailed part 3
  - The Recirculation Fan unit (RCF) detailed part 4
  - The Evacuation of Air unit (EA)
- The workshop, studied part 6
- The other consumptions, studied in part 7

We do not take account of the energy used to build the factory with the tools in the operating consumption.

2. METHOD USED TO ASSESS THE ENERGY CONSUMPTION

We have modelled energy consumption with the Chavanne & Frangi method, which is recognized for assessments of energetic fields yields [7]. An adaptation for ICT services has been started [8].

The goal is to determine the **global rate** of consumption of a service R^{syst}, which is the consumption per unit of service. It can be defined at different levels: to manufacture a computer (then noted PC) it is R^{PC} expressed in $\mathrm{Wh_e PC^{-1}}$, the manufacture a chip it is R^{chip} expressed in $\mathrm{Wh_e cm_{chip}^{-2}}$.

The formulation of a global rate is complex, so the method consists in analyzing a complex system by splitting it in subsystems which can be studied independently. The consumption of the whole system R^{syst} is the sum of the global rates of the subsystems, which all have the same unit.

$$R^{syst} = \sum_j R_j^{syst} \tag{1}$$

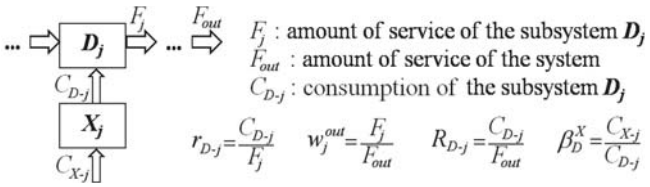


Figure 3. Schematization of a subsystem in the Chavanne & Frangi method [7]

A subsystem is schematized figure 3.

Among the consumption of the computer manufacturing, we can isolate the energy consumed for the chips in a computer, noted R_{chip}^{PC} and expressed in $\mathrm{Wh_e PC^{-1}}$. It is difficult to obtain this value directly, which is not representative of the amount of energy spent to manufacture chips. So we introduce the **local rate** of consumption r_j (here r_{chip} expressed in $\mathrm{Wh_e cm_{in}^{-2} ML^{-1}}$ because the energy is characterized by the surface of chips manufactured and their complexity), which should be as invariable as possible and is a data provided by specialists (in the example, the study on chip factories).

To relate it to the consumption of the whole system, we use its **weight** w_j^{syst} (in our example, it is the surface of chips in a computer, in $\mathrm{cm_{chip}^2 PC^{-1}}$). Then we can calculate the global rate:

$$R_j^{syst} = r_j \cdot w_j^{syst} \tag{2}$$

We differentiate the different kinds of consumption. The direct consumption, noted C_{D-j} (for example the cooling energy provided by the chiller), can be issued from an auxiliary system (the chiller in our example) and has its own consumption C_{X-j} (in practice it is the electric energy consumption). We note β_D^X the yield of conversion of the auxiliary system (for example the COP, coefficient of performance, of the chiller)

$$C_{X-j} = \beta_D^X \cdot C_{D-j} \iff r_{X-j} = \beta_D^X \cdot r_{D-j} \tag{3}$$

3. MAKE-UP AIR UNIT (MUA): CONTROL OF TEMPERATURE AND HUMIDITY

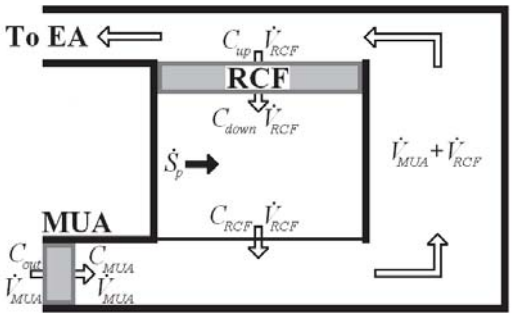


Figure 4. Schema of the air circulation in a clean room (grey elements of figure 2). The flow rates $\dot{V}_j$ and concentrations of particles C_j are also given.

The air circuit is schematized figure 4. Its first part is the MUA, which has several functions with dedicated tools (shown figure 5):

- The filters, which purify air, because the chip fabrication is very sensitive to impurities. They create a pressure drop which must be compensated with a pump or a fan. They also bring outside air in the air circuit with a low excessive pressure (negligible regarding the pressure drop) [6, 3]
- The air conditioning, which controls the humidity and the temperature of the clean room, because the processes are very sensitive to the variation of these parameters. It is divided in two parts
  - The temperature control, which brings the outside air to a determined level of temperature at point G (see figure6). For the Hu 2008 factory, $T_{CR} = 22\,^{\circ}C$ and $T_G = 15.3\,^{\circ}C$ [3]
  - The humidity control, which brings the outside air to a determined level of humidity ω_G. For $\omega_{CR} = 8.32\ g_o kg_{air}^{-1}$, $\omega_G = 7.4\ g_o kg_{air}^{-1}$ [3][2]

For all the clean rooms, the values of the temperature and humidity are almost the same, so we suppose it is the same for T_G and ω_G.

Obviously, the local consumption rate of the MUA, noted r_{e-MUA}, is the consumption per amount of air treated. To report it to the consumption of the air circuit, we multiply it by (All the numerical values given are calculated from the example of the Hu 2008 factory [3])

$$w_{MUA}^{Air} = \dot{V}_{MUA}/S_{CR} = 88.4\ \mathrm{m^3_{MUA} h^{-1} m^{-2}_{CR}} \quad (4)$$

S_{CR} : Surface of the clean room

[2]Every subscript "o" concerns water

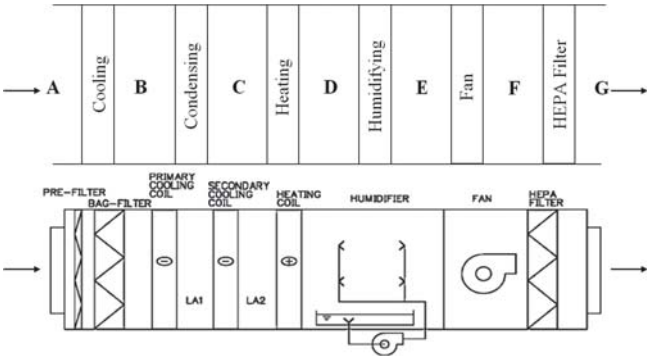


Figure 5. Detail and schematization of the MUA elements [9]

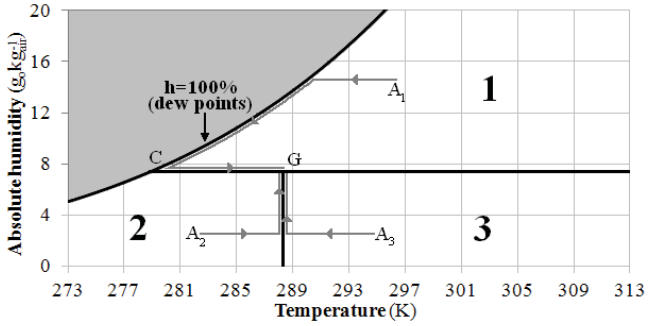


Figure 6. Diagram of $\omega_i = f(T_i)$ with the 3 cases. Zone 1: too humid. Zone 2: not humid enough, too cold. Zone 3: not humid enough, too hot. Grey zone: forbidden zone ($h_i > 100\%$). Point "G": point to reach.

3.1. Control of particles

For the purification, a HEPA filter (The different types of filters are given by the norm EN 1822:2009 [10]) is needed. It is responsible of pressure drop in the air flow. It is the local consumption rate

$$r_{m-MUA_f} = \Delta P_{MUA} = 1234\ \mathrm{Pa} = 0.343\ \mathrm{Wh_m m^{-3}_{MUA}} \quad (5)$$

To calculate the electric consumption associated to this mechanic effort, we need to multiply it by the efficiency of the MUA fan:

$$\beta^e_{m-MUA} = 1.67\ \mathrm{Wh_e Wh_m^{-1}} \quad (6)$$

3.2. Air conditioning

For the air conditionning, there are 3 cases (see figure 6):

- When the outside air is too humid (point A_1): the air is cooled to the dew point to reach a humidity rate of 100% at a constant humidity. Then we continue to cool the air to dehumidify it to the humidity of the clean room. Finally the air is heated to reach T_G.
- When the outside air is not humid enough and too cold (point A_2): the air is heated to reach T_G, then the air is humidified by boiling water to create steam.

- When the outside air is not humid enough and too hot (point A_3): the air is cooled to reach T_G, then the air is humidified by boiling water to create steam.

The physics laws give us the equations which control humidity and temperature. We put ahead the parameters almost constant in equations 7 and 8, and the variables in the equations 9 and 10.

$$r_{th-\Delta T} = c_{p(air)}\rho_{air} = 0.335\ \mathrm{Wh_{th}m^{-3}_{MUA}K^{-1}} \quad (7)$$

$$\begin{aligned} r_{th-\Delta\omega} &= L_{v(o)}(T_{CR})\rho_{air} \\ &= 0.815\ \mathrm{Wh_{th}m^{-3}_{MUA}}\left(\mathrm{g_o kg^{-1}_{MUA}}\right)^{-1} \end{aligned} \quad (8)$$

$$\begin{aligned} w^{T\omega}_{\Delta T} = \left(1 - Hs(-\Delta T)Hs(-\Delta\omega)\left(1+\frac{COP}{COP'}\right)\right)\Delta T \\ +Hs(\Delta\omega)\left(1+\frac{COP}{COP'}\right)(T_G - T_C) \end{aligned} \quad (9)$$

$$w^{T\omega}_{\Delta\omega} = \left(1 - Hs(-\Delta\omega)\left(1+\frac{COP}{COP''}\right)\right)\Delta\omega \quad (10)$$

To take account of the different cases, we introduced

- $\Delta T = T_{out} - T_G$
- $\Delta\omega = \omega_{out} - \omega_G$
- The Heaviside function $Hs(x)$
- COP' and COP'', respectively the coefficient of performance of the heater and the boiler. They are about 10 times higher than COP [3].
- T_C the dew point at the humidity of $7.4\, g_o kg^{-1}_{air}$ (about $5.7\,^\circ$C)

The consumption rate of the air conditioning is

$$R^{T\omega}_{th}\left(\mathrm{Wh_{th}m^{-3}_{MUA}}\right) = r_{th-\Delta T}\cdot w^{T\omega}_{\Delta T} + r_{th-\Delta\omega}\cdot w^{T\omega}_{\Delta\omega}$$

To determine the electric consumption associated to this thermal energy, we multiply R^{MUA}_{th} by

$$\beta^e_{th}\left(\mathrm{Wh_e Wh^{-1}_{th}}\right) = COP^{-1} \quad (11)$$

The COP of the chiller depends on the thermal energy produced and the outside temperature. But in our cases, the variations are low and

$$COP \approx 4 \implies \beta^e_{th} \approx 0.25\ \mathrm{Wh_e Wh^{-1}_{th}} \quad (12)$$

3.3. Synthesis

The total electric consumption of the MUA is

$$\begin{aligned} R^{MUA}_e\left(\mathrm{Wh_e m^{-3}_{MUA}}\right) &= r_{m-MUA}\cdot\beta^e_{m-MUA} \\ &+ R^{T\omega}_{th-MUA}\cdot\beta^e_{th} \end{aligned} \quad (13)$$

We see in the equations 9 and 10 that the difference of temperature and humidity are the main parameters of this consumption. The inside temperature and humidity are almost the same whatever the clean room, so the difference between the clean rooms is the difference of climate of the place where the factory is built.

We have studied 4 different climates to make a one-year average of the air circuit global rates. The pressure drops and the EA consumption are considered as unchanged. We can see in the figure 7 that in a hot and humid climate (Taipei), the consumption is higher than in a temperate (Grenoble) or in a dry (Canberra) or a cold (Stockholm) climate.

The part of consumption in the air circuit in the total lithography consumption is lower in a French factory studied (near Grenoble) than in the Hu fabs [6, 3] (near Taipei), so this confirms this result.

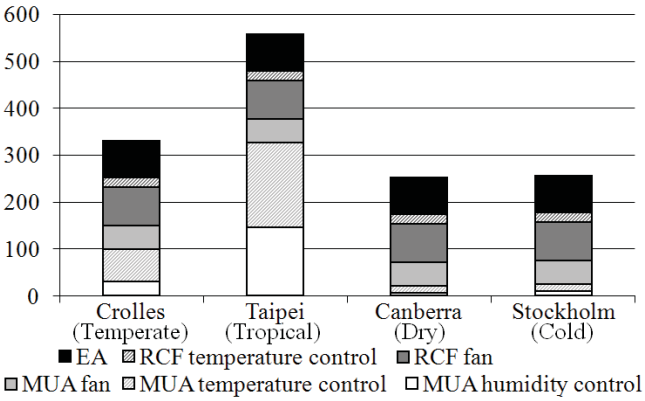


Figure 7. Influence of the climate on the air circuit rate. The energy consumed for the filtration is supposed to be constant.

4. RECIRCULATION FAN UNIT (RCF): CONTROL OF THE PARTICLES CONCENTRATION

This unit brings the pure air in the clean room. It is composed with a fan to lead the air in the clean room, a filter to eliminate particles, and a heat exchanger to evacuate the heat produced by the fan and the pressure drop [3].

As the MUA unit, the rate of the fan is the pressure drop of the filter:

$$r_{m-RCF_f} = \Delta P_{RCF}\left(\mathrm{Wh_m m^{-3}_{RCF}}\right) \quad (14)$$

It is very different between all the clean rooms [6, 3], and depends on the amount of particles diffused in the clean room air and the class of the environment required, which indicates

the concentration of particles in the air, defined by norm ISO 14644-1. For a clean room of class N and a size of the particles d_p, the concentration of particles is:

$$\Sigma C_N\left(d_p \geq D_p(\mathrm{m_p})\right) = 10^N \left(\frac{10^{-7}}{D_p}\right)^{2.08} \text{part m}_{\text{air}}^{-3}$$

$$C_N\left(d_p\right) = 2.08 \cdot 10^{N+7} \left(\frac{10^{-7}}{d_p}\right)^{3.08} \text{part m}_{\text{p}}^{-1}\text{m}_{\text{air}}^{-3} \quad (15)$$

The schema of the concentrations of particles in the air circuit is given in figure 4. We can calculate the concentrations in relation with the diffusion of particles in the clean room $\dot{S}_p(\text{part m}_{\text{p}}^{-1}\text{m}_{\text{CR}}^{-2}\text{h}^{-1})$, the filter efficiencies CE_{MUA} and CE_{RCF}, and the flow rates.

The filters are less efficient for $d_p = 0.1\ \mu$m [10], so we use this size of particle in our study. The concentrations C_{RCF} and C_{out} are defined with the equation 15, with respectively the class of the clean room (4.5 for Hu 2008 [3]) and $N = 9$ (class of the outside air according to the norm ISO 14644-1).

By definition, we have

$$CE_{MUA} = C_{out}/C_{MUA} \quad (16)$$

$$CE_{RCF} = C_{up}/C_{down} \quad (17)$$

The filter efficiency can be expressed by [11]:

$$CE_f = \exp\left(\frac{4}{\pi} \cdot \frac{\alpha_f}{d_f(1-\alpha_f)} \cdot e_f \cdot \eta_f\right) \quad (18)$$

$$\eta_f = 2.6 \cdot \left(\frac{1-\alpha_f}{Ku}\right)^{1/3} \cdot \left(\frac{v_{air} \cdot d_f}{D}\right)^{-2/3} + \left(\frac{1-\alpha_f}{Ku}\right) \cdot \left(\frac{d_p^2}{d_f \cdot (d_p + d_f)}\right) \quad (19)$$

α_f : Filter density

Ku : Kubawara number $\left(-\frac{\ln \alpha_f}{2} - \frac{3}{4} + \alpha_f + \frac{\alpha_f^2}{4}\right)$

d_f : Diameter of the filter fiber

v_{air} : Air speed in the filter

D : Coefficient of diffusion of particles in the air ($7.20\ \text{m}^2\text{s}^{-1}$ in our conditions)

e_f : Thickness of the filter

In the Hu 2008 study, the MUA filter is a HEPA filter [3]. The document [10] gives us a typical HEPA filter characteristics which allows to calculate $CE_{MUA} = 1.74 \cdot 10^3$. We deduce with equation 16: $C_{MUA} = 1.19 \cdot 10^{13}\ \text{part m}_{\text{p}}^{-1}\text{m}_{\text{MUA}}^{-3}$.

C_{up} is an average between the air concentration of RCF and MUA:

$$C_{up} = \frac{C_{MUA}\dot{V}_{MUA} + C_{RCF}\dot{V}_{RCF}}{\dot{V}_{MUA} + \dot{V}_{RCF}} = 2.13 \cdot 10^{12}\ \text{part m}_{\text{p}}^{-1}\text{m}_{\text{RCF}}^{-3} \quad (20)$$

We note that the filter is fold up, so e_f and v_{air} are not the apparent filter thickness and the apparent air speed.

After approximations, with typical values ($\alpha_f \approx 0.1$; $d_f \approx 1\ \mu$m ; $d_p \approx 0.1\ \mu$m ; $v_{air} \approx 5\ \text{cm s}^{-1}$), the second term of η_f can be ignored. So $\ln CE_f$ is proportional with $v_{air}^{-2/3}$, and the coefficient of proportionality, called λ_{CE}, depends only on the characteristics of the filter.

We also have the value of the pressure drop in a filter [10]:

$$\Delta P_f = 16\mu_{air} \cdot \frac{\alpha_f \cdot e_f}{d_f^2 \cdot Ku} \cdot v_{air} \quad (21)$$

μ_{air} : Dynamic viscosity of air ($1.8 \cdot 10^{-5}$ Pa s)

So the pressure drop is proportional to the air speed. The coefficient of proportionality, called λ_P, depends only on the characteristics of the filter. By combining these equations, we deduce

$$\ln CE_f = \lambda_{CE} \cdot \lambda_P^{2/3} \cdot \Delta P_f^{-2/3} \quad (22)$$

After having contacted S.C.Hu, we learnt that the RCF filter is an ULPA. A designer of ULPA filters gave us characteristics of such filters [12]. We deduce for the Hu 2008 case ($\Delta P_{RCF} = 153$ Pa) $CE_{RCF} = 2.6 \cdot 10^7$. With this value of CE_{RCF} and the equation 17, we deduce $C_{down} = 8.2 \cdot 10^5\ \text{part m}_{\text{p}}^{-1}\text{m}_{\text{RCF}}^{-3}$.

We also have a relation between the production of particles and the concentrations in the clean room:

$$C_{RCF} = C_{down} + \frac{\dot{S}_p \cdot S_{CR}}{\dot{V}_{RCF}} \quad (23)$$

$$\dot{S}_p = (C_{RCF} - C_{down})\frac{\dot{V}_{RCF}}{S_{CR}} \approx \frac{C_{RCF}\dot{V}_{RCF}}{S_{CR}} \approx 3.9 \cdot 10^{14}\ \text{part m}_{\text{p}}^{-1}\text{m}_{\text{CR}}^{-2}\text{h}^{-1} \quad (24)$$

The class of a clean room is not established at its construction: it is possible to change it by modifying the flow rate. Indeed, with the equation 24, we can say that

$$\dot{V}_{RCF} = \frac{\dot{S}_p \cdot S_{CR}}{C_{RCF}} = \frac{\dot{S}_p \cdot S_{CR}}{2.08 \cdot 10^{N+7}} \quad (25)$$

The production of particles and the surface of the clean room are supposed constant.

v_{RCF} is linked to $\dot{V}_{RCF}$ with this equation:

$$v_{RCF} = \frac{\dot{V}_{RCF}}{S_{CR}} \cdot \frac{1}{\tau_{fold} \cdot \tau_{cov}} \quad (26)$$

τ_{fold} : Fold rate of the filters (a typical value is 30 [12])

τ_{cov} : Coverage rate of the clean room by the filters (Typical values of this data are between 25%-50%. After having contacted Hu, we obtained the value of $\tau_{cov} = 30\%$ for the Hu 2008 factory.)

We can deduce from equations 21 and 25

$$\Delta P_{RCF} = \frac{\lambda_P}{\tau_{fold} \cdot \tau_{cov}} \cdot \frac{\dot{S}_p}{2.08 \cdot 10^{N+7}} \quad (27)$$

We can consider the left term, constant, as the local rate r_{m-RCF_f}. For the Hu 2008 factory,

$$r_{m-RCF_f} = 7 \cdot 10^{-4}\ \mathrm{W_m m^2_{CR}} (\mathrm{m^3_{RCF} h^{-1}})^{-2} \quad (28)$$

To calculate the electricity consumption associated, we multiply $R_m^{RCF_f}$ by the efficiency of the RCF fan

$$\beta^e_{m-RCF} = 3.33\ \mathrm{Wh_e Wh_m^{-1}} \quad (29)$$

According to Hu, the pressure drop induces a release of heat equivalent to the amount of electricity consumed [3]. We multiply this value by β^e_{th} (see equation 8) to have the electric energy consumed by the chiller to remove this heat.

In conclusion, we can say that the rate of the RCF has the following expression:

$$\begin{aligned} R_e^{RCF}\left(\mathrm{Wh_e m^{-3}_{RCF}}\right) &= r_{m-RCF_f} \cdot w_{RCF} \quad (30) \\ w_{RCF} &= \frac{\dot{S}_p \cdot \beta^e_{m-RCF} \cdot (1+\beta^e_{th})}{2.08 \cdot 10^{N+7}} \end{aligned}$$

To get the consumption rate of the RCF system, we multiply R_e^{RCF} by

$$w^{Air}_{RCF} = \dot{V}_{RCF} / S_{CR} \quad (31)$$

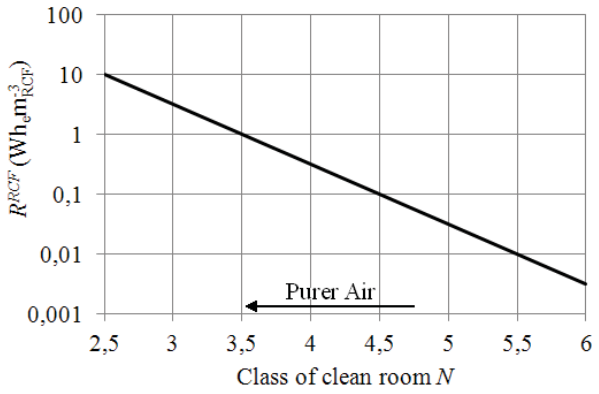


Figure 8. Influence of N on the consumption rate of the RCF with the analysis of the Hu 2008 factory [3].

We see that the consumption of the RCF is higher for high $\dot{V}_{RCF}$, which is linked to the class of the clean room. We see on figure 8, for a constant value of $\dot{S}_p$, we have a great growth of the consumption rate with the increase of the air purity. For this reason, it is important to adjust the purity to the real processes needs.

5. CONCLUSION FOR THE AIR CIRCUIT

The consumption rate of the air circuit is noted R_e^{Air}, expressed in $\mathrm{Wh_e m^{-2}_{CR}}$, because the more the clean room is wide, the more we have to purify air. The consumption rate of the air circuit is

$$R_e^{Air} = r_{e-MUA} \cdot w^{Air}_{MUA} + r_{e-RCF} \cdot w^{Air}_{RCF} + r_{e-EA} \cdot w^{Air}_{EA} \quad (32)$$

For the EA, we could not get the parameters influencing its consumption. However, the consumption of this element takes a little part of the air circuit consumption compared to the other elements (figure 1). So we consider $r_{e-EA}(\mathrm{Wh_e m^{-3}_{EA}})$ as the rate of the EA. To relate this local rate to the global rate of the air circuit, we multiply it by

$$w^{Air}_{EA} = \dot{V}_{EA} / S_{CR} \approx \dot{V}_{MUA} / S_{CR} \quad (33)$$

The surface of the clean room is linked to the surface of input wafer which can be treated $N_{max}(\mathrm{cm^2_{max} h^{-1}})$ and number of mask layers $N_{ML}(\mathrm{ML})$:

$$w^{max}_{CR}\left(\mathrm{cm^2_{max} h^{-1} m^{-2}_{CR} ML}\right) = N_{max} \cdot N_{ML} \,/\, S_{CR} \quad (34)$$

$$w^{in}_{max}\left(\mathrm{cm^2_{in} cm^{-2}_{max}}\right) = N_{prod} \,/\, N_{max} \quad (35)$$

w^{in}_{max} is the rate of utilization of the clean room and N_{prod} is the real production of wafers.

We represented on figure 9 the product of w^{max}_{CR} and w^{in}_{max} for the factories studied by Hu in 2003 [6]. Knowing that the factories I and G did not work near their full capacities, we can say that w^{max}_{CR} is quite the same in the factories, and is equal to about 40 $\mathrm{cm^2_{max} h^{-1} m^{-2}_{CR} ML}$. With this value, we can link R^{Air} with the capacity of the clean room.

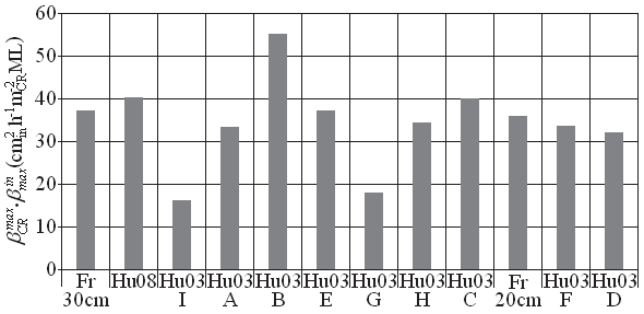


Figure 9. Calculation of $w^{max}_{CR} \cdot w^{in}_{max}$ for the Hu [6, 3] and the Crolles factories. The values are expressed in $\mathrm{cm^2_{in} h^{-1} m^{-2}_{CR} ML}$

6. STUDY OF WORKSHOPS

It is the other most energy consuming post (figure 1), divided in three parts:

- The lithography tools
- The tools cooling (TC), which corresponds to the evacuation of heat
- The pumps which allow the process cooling water (PCW) circulation

Some consumptions are directly proportional to the production and the number of mask layers used (see caption of the figure 3), for example etching tools), and other consumptions are independent from the production, for example vacuum pumps). We separate these two parts to calculate the local rates.

Table 1. Other consumptions (Compressed Dry Air (CDA), Ultra Pure Water (UPW) and pur nitrogen (N2)) in the Hu 2008 factory[3]. The production was 504 000 wafer per year.

i	**Consumption**	**Utilization**	R_e^i	w_i^{other}	r_{e-i}^{other}
UPW	$22.6\,\mathrm{MWh_e/day}$	$24\,\mathrm{h/day}$	$\mathrm{Wh_e m_{UPW}^{-3}}$	$\mathrm{m_{UPW}^3 cm_{in}^{-2} ML^{-1}}$	$2.17\,\mathrm{Wh_e cm_{in}^{-2} ML^{-1}}$
CDA	$20.0\,\mathrm{MWh_e/day}$	$24\,\mathrm{h/day}$	$\mathrm{Wh_e m_{CDA}^{-3}}$	$\mathrm{m_{CDA}^3 cm_{in}^{-2} ML^{-1}}$	$1.84\,\mathrm{Wh_e cm_{in}^{-2} ML^{-1}}$
N2	$221\,\mathrm{Wh_e m_{N_2}^{-3}}$	$4909\,\mathrm{m_{N_2}^3 h^{-1}}$	$221\,\mathrm{Wh_e m_{N_2}^{-3}}$	$\mathrm{m_{N_2}^3 cm_{in}^{-2} ML^{-1}}$	$2.40\,\mathrm{Wh_e cm_{in}^{-2} ML^{-1}}$

6.1. Part depending on the production

Its consumption rate is noted R_e^{WS-var}, and it is expressed in $\mathrm{Wh_e cm_{in}^{-2} ML^{-1}}$. It is composed by the variable consumption of the workshops:

$$R_e^{WS-var} = R_e^{tools-var} + R_e^{TC-var} + R_e^{PCW-var} \quad (36)$$

Heat comes out of the tools, and its amount is equivalent to the electric energy consumed by the tools. To obtain the electric consumption necessary to deliver this thermal energy and to compare it to the other elements of the workshops, we need to multiply it by β_{th}^e.

The cooling water is brought by pumps. The characteristic service of the pumps is the flow rate of cooling water $\dot{V}_{PCW}$, so the rate is the consumption per unit of volume of water $R_e^{PCW-var}$. To compare to the other variable elements of the workshop, we multiply it by

$$w_{PCW-var}^{WS-var} = \frac{\dot{V}_{PCW}}{N_{prod} \cdot N_{ML}} \quad (37)$$

In conclusion, we replace the expression 36 by

$$\begin{aligned} R_e^{WS-var} &= R_e^{tools-var} \cdot (1 + \beta_{th}^e) \\ &+ R_e^{PCW-var} \cdot w_{PCW-var}^{WS-var} \end{aligned} \quad (38)$$

6.2. Part independent from the production

This part depends on the number of masks and number of wafer maximum treated by the clean room. This consumption rate is noted R_e^{WS-fix}, expressed in $\mathrm{Wh_e cm_{max}^{-2} ML^{-1}}$ Like the variable part, it is the sum of the fixed consumption of the workshops.

The remarks made at the previous subsection are available, but instead of being based on wafers really treated, the calculations are made regarding the capacity of the clean room. Like the equation 38, we have

$$\begin{aligned} R_e^{WS-fix} &= R_e^{tools-fix} \cdot (1 + \beta_{th}^e) \\ &+ R_e^{PCW-fix} \cdot w_{PCW-fix}^{WS-fix} \end{aligned} \quad (39)$$

7. OTHER ELEMENTS OF THE FACTORY

7.1. Chemicals

A lot of chemicals are used during the lithography. The energy necessary to manufacture them may be consequent: $45.2\,\mathrm{g}$ of chemicals are necessary to treat $1.6\,\mathrm{cm^2}$ of wafers [5]. We calculate the local rate of the chemical production as a variable rate. So for 25 ML, the local consumption rate: $R_\chi^\chi = 1.13\,\mathrm{g_\chi cm_{in}^{-2} ML^{-1}}$ [3].

$2.3\,\mathrm{MJ}$ are necessary to manufacture the $45.2\,\mathrm{g}$ of chemicals [5], so $\beta_\chi^e = 14.1\,\mathrm{Wh\,g_\chi^{-1}}$. We can report R_χ^χ to the global rate thanks to the followed operation:

$$R_e^\chi = R_\chi^\chi \cdot \beta_\chi^e = 16\,\mathrm{Wh\,cm_{in}^{-2} ML^{-1}} \quad (40)$$

This rate may be different from the one given by E. Williams, and it is comparable to the other higher rates of the lithography phase given figure 1. For example, for the French factory, the rate is $R_e^\chi = 35\,\mathrm{Wh\,cm_{in}^{-2} ML^{-1}}$. So we can't generalize the value of equation 40 and a further analysis is necessary, like for the workshops.

7.2. Other consumptions

For the other elements (UPW, CDA, N2), we have data presented table 1. These consumptions certainly depend on the surface of wafer treated and the number of masks.

These consumptions are not very high regarding the lithography phase consumption (see figure 1). We regroup all these consumptions in the following consumption rate:

$$\begin{aligned} R_e^{other}(\mathrm{Wh_e cm_{in}^{-2} ML^{-1}}) &= R_e^{UPW} \cdot w_{UPW}^{other} \\ + R_e^{CDA} \cdot w_{CDA}^{other} &+ R_e^{N2} \cdot w_{N2}^{other} \end{aligned} \quad (41)$$

8. CONCLUSION

Thanks to this study, we put ahead the main parameters and the local rates to model the consumption of the clean room air circuit.

To consume as few energy as possible, the outside air humidity and temperature must be at levels close to the point G levels. The building place of the factory is very important: we prefer colder and dryer climates than hotter and more humid ones (see figure 7).

For the particles management, it is important to adapt the class of the clean room to the air purity really needed to avoid overconsumption of the RCF unit as said at the end of part 4.

A deeper study of the workshops may put ahead the parameters influencing the consumption of the tools and the chemical products which are the other parts which consume a lot in the clean room (see figure 1).

We can calculate the rates of consumption of the part depending on the production ($R_e^{var}(\mathrm{Wh_e cm_{in}^{-2} ML^{-1}})$) and the independent part ($R_e^{fix}(\mathrm{Wh_e cm_{max}^{-2} ML^{-1}})$):

$$R_e^{var} = R_e^{WS-var} + R_e^{other} + R_e^{\chi} \quad (42)$$

$$R_e^{fix} = R_e^{WS-fix} + R_e^{Air} \cdot w_{CR}^{max} \quad (43)$$

The system consumption rate of the lithography phase is finally given by this expression:

$$R_e^{Litho}\left(\mathrm{Wh_e cm_{in}^{-2} ML^{-1}}\right) = R_e^{var} + R_e^{fix} \cdot w_{max}^{in} \quad (44)$$

Then, the consumption of the rest of the manufacturing phase and the utilisation phase has to be assessed to have the life cycle impact of ICT services, and then we will be able to compare it to material services and find ways to improve the environmental footprint of human society thanks to ICTs.

9. ACKNOWLEDGMENTS

The authors thank S.C.Hu for his availability to give precision on the data of his documents, and STMicroelectronics at Crolles for their environmental report data.

10. REFERENCES

[1] International Telecommunication Union, "Using icts to tackle climate change," November 2010.

[2] M. Ellis, N. Jollands, International Energy Agency, Organisation for Economic Co-operation, and Development, *Gadgets and Gigawatts: Policies for Energy Efficient Electronics*, OECD/IEA, 2009.

[3] S.C. Hu, J.S. Wu, D.Y.L. Chan, R.T.C. Hsu, and J.C.C. Lee, "Power consumption benchmark for a semiconductor cleanroom facility system," *Energy and Buildings*, vol. 40, no. 9, pp. 1765–1770, 2008.

[4] R. Kuehr and E. Williams, *Computers and the environment: Understanding and managing their impacts*, vol. 14, Kluwer Academic Pub, 2003.

[5] E.D. Williams, R.U. Ayres, and M. Heller, "The 1.7 kilogram microchip: Energy and material use in the production of semiconductor devices," *Environmental science & technology*, vol. 36, no. 24, pp. 5504–5510, 2002.

[6] S.C. Hu and YK Chuah, "Power consumption of semiconductor fabs in taiwan," *Energy*, vol. 28, no. 8, pp. 895–907, 2003.

[7] X. Chavanne and J.P. Frangi, "Comparison of the energy efficiency to produce agroethanol between various industries and processes: Synthesis," *Biomass and Bioenergy*, 2011.

[8] D. Marquet, M. Aubrée, J.P. Frangi, and X. Chavanne, "Scientific methodology for telecom services energy consumption and co2 emission assessment including negative and positive impacts," in *Telecommunication-Energy Special Conference (TELESCON), 2009 4th International Conference on*. VDE, pp. 1–14.

[9] J.M. Tsao, S.C. Hu, T. Xu, and D.Y.L. Chan, "Capturing energy-saving opportunities in make-up air systems for cleanrooms of high-technology fabrication plant in subtropical climate," *Energy and Buildings*, vol. 42, no. 11, pp. 2005–2013, 2010.

[10] INERIS, "Filtres tres haute efficacite - http://www.ineris.fr/ badoris/ pdf/ substances_toxiques/ tox_filtre_the_v0.pdf," Juin 2006.

[11] RA da Roza, "Particle size for greatest penetration of hepa filters-and their true efficiency," Tech. Rep., Lawrence Livermore National Lab., CA (USA), 1982.

[12] "Website of filters hepa et ulpa designers – http://www.fcr.it/ air-filtration/ absolute-filters-epa-hepa-ulpa," .

ROBUST AUDIO WATERMARKING BASED ON DYNAMIC DWT WITH ERROR CORRECTION

Hemam A. Alshammas
Computer Engineering Department, The University of Jordan, Amman, Jordan, hemam_ayed@yahoo.com

ABSTRACT

Audio watermarking was introduced as a solution for the arising challenges facing audio ownership verification. These challenges are a result of easiness and high speed of copying and distribution digital audio. This paper presents enhancements in the performance of an audio ownership verification system that has been reported previously. The proposed system is based on the Discrete Wavelet Transform (DWT). A new approach for audio signal framing, dynamic DWT leveling, error correction code and new embedding methods are suggested to improve the watermark bit rate, minimum audio-cover period, quality of the watermarked audio and watermark robustness against audio attacks. The evaluation of the suggested system showed the following improvements: the watermark bit rate increased 23.4 times, 92% reduction in the minimum required audio-cover period, 54% increase in the Signal-to-Noise Ratio (SNR) of watermarked audio, and it demonstrated better robustness against watermarking benchmark attacks.

Keywords— audio watermarking, ownership verification, audio copyright protection, discrete wavelet transform, error correction

1. INTRODUCTION

Digital watermarking has been a powerful technique of hiding information into digital media. It can be applied to the many types of media, including image, audio and video in order to protect copyright or verify ownership. In this paper, the audio ownership verification is in concern. The great revolution in digital audio production resulted in new challenges facing ownership verification, as unauthorized distribution of copyrighted material became very common. A number of techniques for digital audio watermarking have been introduced to stop or limit the impact of these challenges [1] – [11].

In digital watermarking, data can be hidden in another object (host data) then the watermarked object can be distributed. Another method used for audio ownership verification is audio fingerprinting [1]. Audio characteristics of a given audio signal are extracted. A unique ID for this audio is formed using the extracted characteristics and saved in a database as a fingerprint for the original audio. Unlike watermarking, additional information are not embedded into the signal. Watermarking process can be performed in the spatial domain where the watermark is added directly to an audio signal. An alternative way is to transfer the audio to a frequency domain where the watermark is embedded [2], [3]. Ramakrishnan et. al. [4] introduced a method for image watermarking. Their method aims to the improvement of the robustness of still images using Discrete Wavelet Transform (DWT) and Singular Value Decomposition (SVD) techniques. Lalitha et. al. [5] evaluated the performance of the Discrete Cosine Transform (DCT) and SVD based audio watermarking and compared it with the performance of DWT-SVD based audio watermarking. The comparison led to the conclusion that the DWT-SVD had shown better outcomes than the DCT-SVD techniques.

The human auditory system is more sensitive than the human visual system. Therefore, it is more challenging to attain the required imperceptibility [6], [7]. Furthermore, it is difficult to achieve the same levels of robustness and imperceptibility with the same amount of information embedded because audio signals usually have less samples rate. Accordingly, research about audio watermarking is constrained due to the sensitivity of the human auditory system and audio signals. Al-Haj et. al. [8] proposed a DWT based technique for audio signals watermarking which has the least effect on the original signal. Al-Haj and Mohammad [9] suggested a digital audio watermarking based on DWT-SVD techniques. This suggestion was designed to improve watermark robustness after the audio is subjected to various audio attacks.

Nevertheless, in the systems related to audio copyright-ownership verification, robustness is not the only requirement a watermarking technique should satisfy. Another requirement is imperceptibility of the embedded watermark to obtain low levels of distortion in the audio, and persistence and uniqueness to identify the owner from all parties claiming copyright. Al-Yaman et. al. [10] used code assignment to minimize the effect of the ones in the embedded digital watermark bits to improve the quality of audio. Using a hash function encoding, they improved the minimum audio signal cover period. Al-Yaman et. al. [11] introduced new framing approach of the audio signal. A new way to form the matrix after DWT transformation was presented. They proposed embedding methods to improve the minimum audio-cover period, quality of the watermarked audio and its robustness against various attacks.

In this paper, an improved system for audio ownership verification is proposed. It builds on the results reported in

[11]. The improvements presented include modifications on the audio signal framing to improve minimum cover period and maximum bit rate, dynamic DWT leveling, using error correction codes and changes on the watermark embedding procedure which lead to dropping the use of SVD. These enrichments aim to increase the quality of the watermarked audio and the robustness of the watermark. The rest of the paper is organized as the following: Section 2 describes the procedure in which the watermark data is embedded and extracted into and from the original audio signal. The suggested improvements are discussed in detail in Section 3. In Section 4, the experimental results are discussed and compared with the results of the old systems. Finally, Section 5 draws conclusions of the proposed paper.

2. SYSTEM OVERVIEW

Fig.1 shows the conceptual representation for the suggested audio ownership verification system. As shown, the watermark image that belongs to the owner is embedded into the audio. This is performed in a way that it results in the minimal effect on the quality of the original audio. To achieve this, dynamic DWT leveling and enhanced embedding is considered. Then, by comparing the given watermark image with the one embedded in the audio, the audio ownership can be verified. The basic operation of the system being investigated is outlined briefly as follows.

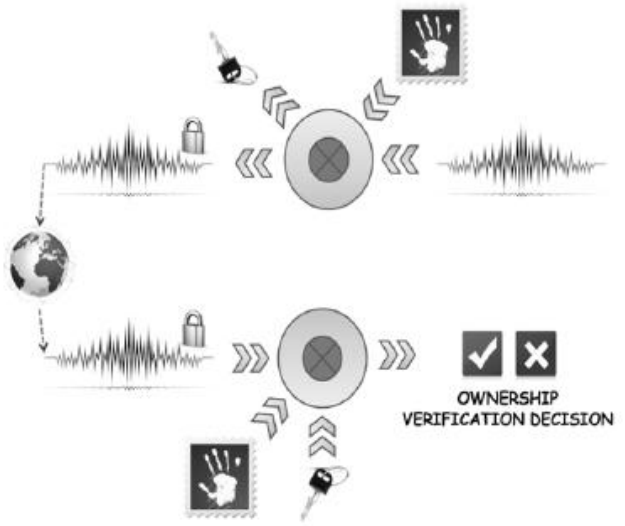


Figure 1. System overview.

A. Watermark Embedding Procedure

An audio signal runs into the following procedure when the watermark data is to be embedded into it: sampling the original audio signal, dividing the samples into frames (framing) and then a dynamic n-level DWT is applied to each frame.

The digital image to be used as watermark is then encrypted using SHA-1 hash algorithm, which returns the digest of the watermark image as a 160 bits string [10]. The hash bits are then encoded using extended Golay code, and then embedded into the n-level DWT as will be described in the proposed enhancements section.

To produce the final watermarked audio, inverse n-level DWT is performed. The stages of the described watermark embedding procedure are shown in Fig. 2. In the proposed system, the framing, dynamic DWT leveling, the use of error detection-correction code and the embedding process are the core enhancement suggested.

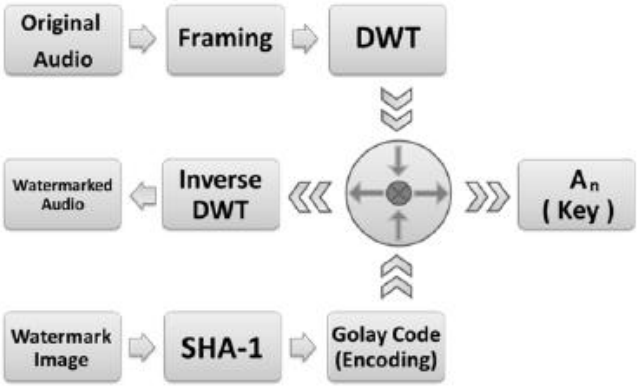


Figure 2. Watermark embedding procedure.

B. Watermark Extraction Procedure

To extract the encrypted watermark from the watermarked audio, a procedure similar to the watermark embedding is followed. The stages of the watermark extraction procedure are shown in Fig. 3.

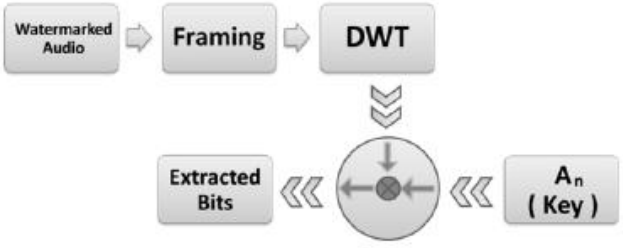


Figure 3. Watermark extraction procedure.

C. Ownership Verification Procedure

After watermark bits are extracted, they are decoded using Golay codes to recover the hash bits of the watermark image. Then, the decoded bits are compared with the bits obtained from performing the Hash (SHA-1) on the watermark image. Verification is confirmed when the compared bits match completely. This procedure is shown in Fig. 4.

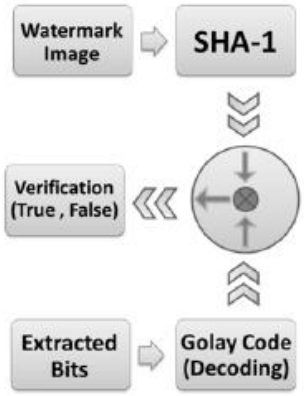


Figure 4. Ownership verification process.

3. THE PROPOSED ENHANCEMENTS

This section explains the suggested improvements on the proposed audio ownership verification system. These improvements include new audio signal framing method, dynamic DWT leveling, the use of error correcting code and new embedding methods to improve the performance of the system in the magic triangle of watermarking requirements: the bit rate, quality of the watermarked audio and watermark robustness against noise, attacks or even normal audio processing. The proposed improvements and their effects on the system are detailed below.

A. Audio framing

Applying the watermark to all frames will increase the bit rate of the watermark and therefore will improve the audio minimum cover period. However, applying it only to selected frames would decrease SNR as mentioned in the results of [11]. The issue of SNR is solved in the suggested system using different approaches. Higher bit rate is required due to the new overhead bits added to the hash bits. These bits are used for error correction proposed stage (Golay code). An example to demonstrate the increase of bit rate and the minimum audio-cover period, when applying the watermark to an audio with sampling frequency of 44100 Hz and the audio is divided into frames of length 1024 samples, according to the Eqn. 1, the new framing approach allows adding 43.07 watermark bits in each second which is a significant improvement compared with [11] which reported a bit rate of 1.84 bps.

$$\boldsymbol{Bit\,rate = Fs/FL} \qquad (1)$$

where Fs corresponds to the sampling frequency and FL is the used frame length.
The proposed framing method reduced the audio minimum cover period by 92%, to become 7.8 seconds instead of 90 seconds as reported in [11] regardless of the watermark image size.

B. Dynamic DWT Leveling

The discrete wavelets transform (DWT) produces a time-frequency representation of a signal. Given the original audio signal S, the 1st level DWT finds two sets of coefficients; the approximated coefficients A and the details coefficients D. The approximated coefficients A represent low frequencies and are found by passing the signal through a low pass filter. The details coefficients D represents high frequencies and are found by passing the signal through a high pass filter. The low frequencies part (A) could be decomposed again into two parts of low and high frequencies. Fig. 5 shows n-level DWT for S. The original signal S can be recomposed using the inverse DWT process [9].

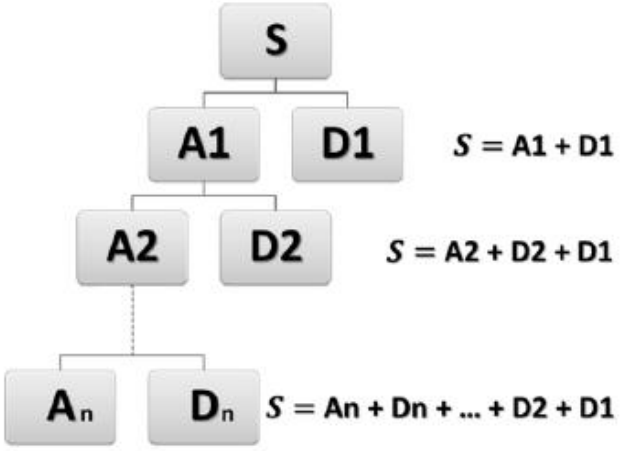


Figure 5. n-Level DWT decomposition.

The maximum number levels of DWT could be applied depends on the frame length, as described in Eqn. 2.

$$\boldsymbol{n = log_2 FL} \qquad (2)$$

where FL is the frame length described previously. The new watermark embedding procedure requires applying DWT at the maximum number of levels in order to obtain the last low frequency coefficient (A_n) in the audio signal. This coefficient is a [1×1] matrix that can be used in the embedding process. Therefore, the system must perform n-level DWT and (n) will be selected dynamically depending on the previously selected frame length. To apply the method correctly, (n) must be a real integer. Thus, the selected frame length must be a power of 2.

C. Error Correction Code

In order to increase the robustness of the watermark against different audio attacks, the proposed system uses the (24, 12) extended Golay code for error correction. The extended Golay code encodes 12 bits of data and returns a 24-bit word in such a way that is capable of performing 7-bit error detection and 3-bit error correction per word [12].
In the (24, 12) extended Golay code, the polynomial code used is:

$$\bar{c}(x) = 1 + x^2 + x^4 + x^5 + x^6 + x^{10} + x^{11} + x^{23} \quad (3)$$

It has a weight of 8 and is expressed as a sum of the generator polynomial g(x) and the parity-check polynomial x23, where g(x) is a polynomial code in the (23, 12) Golay code and is represented by:

$$g(x) = 1 + x^2 + x^4 + x^5 + x^6 + x^{10} + x^{11} \quad (4)$$

Each 12 bits of the 160 hash bits are grouped together and encoded with (24, 12) extended Golay code to create a 24-bit sequence. There will be 14 sequences of encoded bits, that is 336 bits in total, as illustrated in Figure 6. The extended Golay code is used to correct at most 3 errors, which means the system can correct up to 42 errors.

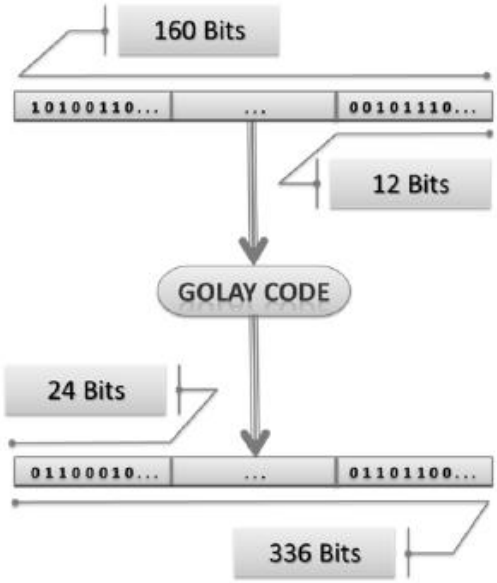


Figure 6. Applying Golay code to hash bits.

D. Embedding Process

In the suggested system, the embedding process is performed according to the following formula:

$$A_{nw} = A_n [\, 1 + \alpha \times W(i) \,] \quad (5)$$

where W(i) is a bit from the watermark bits after Golay code is applied, α is the watermark intensity, A_n is the last approximation coefficient from the dynamic n-level DWT, and A_{nw} is the watermarked A_n. For example, if the watermark intensity (α) is set to 0.5, then A_{nw} will equal (1.5 A_n) when W(i) is 1, and to (A_n) when W(i) is 0. This formula replaced the one used in [9], [10] and [11] which included a part of the S matrix obtained from SVD.

Increasing the value of watermark intensity (α) used in Eqn. 5 will lead to increasing the robustness of a watermark. Higher values of α will add more noise immunity to the watermark bits and they are easier to extract. Conversely, the probability of the watermark bits being lost using smaller values of α will increase. In the other hand, increasing α will decrease SNR [11]. In Fig. 7, the relationship between the value of watermark intensity (α) used in the proposed system and the resulting SNR value is shown. The curve for each sound starts ate the point of the minimum value of α at which the watermark can be retrieved successfully.

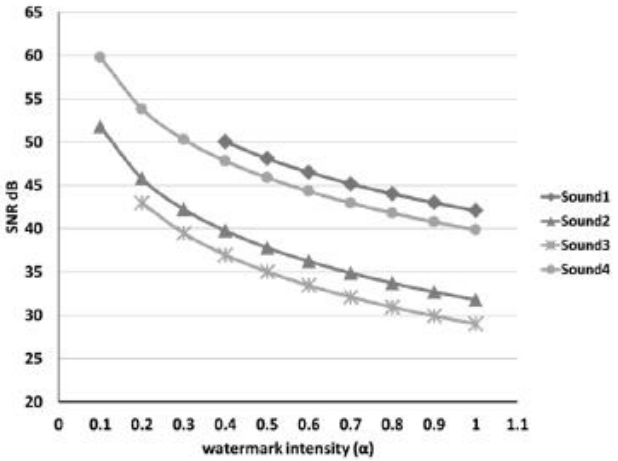


Figure 7. SNR for different audio samples.

Instead of applying the watermark after the SVD transform used in the old method [11], the proposed system suggests to embed the watermark only in the last low frequency coefficient after the dynamic n-level DWT transformation described earlier.

This new approach for embedding the watermark bits improved the quality of the watermarked audio considering that the watermark bits are not added to all frequency components of the signal but only to the lowest frequency coefficients obtained from the dynamic n-level DWT**.**

For the same audio signal, using different frame lengths for the proposed framing stage creates multiple SNR versus watermark intensity (**α**) curves, as shown in Figure 8.

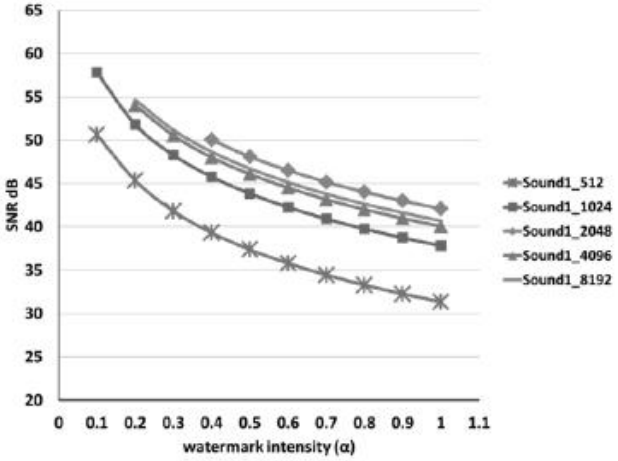


Figure 8. SNR curves for different frame lengths.

4. RESULTS AND DISCUSSION

In this section, the effect of the proposed improvements on the watermarking system under investigation is described and compared with the old systems [9], [10], [11]. As discussed earlier, applying the watermark in the proposed system to all frames after partitioning the signal caused a large increase in the bit rate and it improved the minimum audio-cover period remarkably. Using a 100×100 watermark image as an example, the improvements caused by the new framing approach are shown in Table 1.

TABLE 1. The bit rates and minimum audio cover periods for different watermarking methods.

Methods	Bit Rate (bps)	Min. Cover Period (second)
[9] DWT-SVD,2010	0.882	11338
[10] Bio. DWT-SVD,2011	0.882	293
[11] Enh. DWT-SVD,2012	1.84	90
Suggested Method	43.07	8

The suggested watermark embedding procedure highly increased the quality of the watermarked audio. This improvement was a result of adding the watermark bits only to the last approximate coefficient after the dynamic DWT leveling process, as in Eqn. 5. Figure 9 shows SNR for embedding different watermark images using α = 0.1, which is the value used in [11].

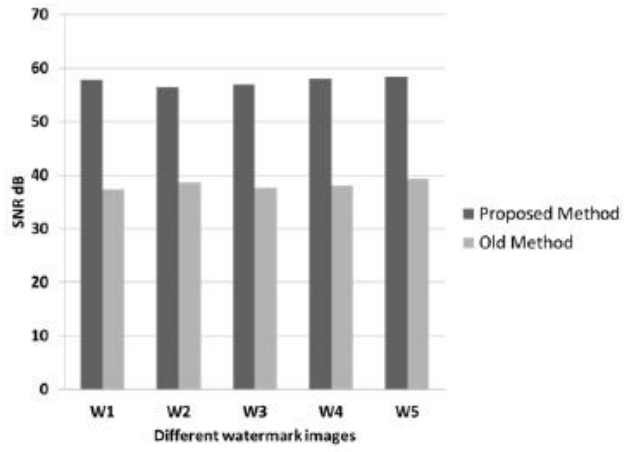


Figure 9. SNR for watermarked audio when using different watermark images.

The robustness of the system against different types of attacks was enhanced by using the extended Golay code for error correction. The experimental results obtained showed system robustness against bench mark attacks compared to the old systems [9] and [11], as shown in Table 2. The used attacks are defined by Stirmark® watermarking benchmark [13].

TABLE 2. System robustness test against various attacks.

Stirmark® Attack Name	[9] DWT-SVD ,2010 BER (%)	[11] Enh. DWT-SVD ,2012 BER (%)	Suggested Method BER (%)
AddBrumm	0	0	0
AddSinus	0	0	0
AddNoise	0	0	0
ZeroCross	0	N/A	3.1250
LSB Zero	0	0	0
Amplify	0	0	0
Stat1	45	1.25	0
Stat2	N/A	1.875	0
Invert	0	0	0
Smooth	0	1.875	0
Extra Stereo	0	0	0

Bit Error Rate (BER) is used to measure watermark robustness. It is defined as the ratio of incorrect bits to the total amount of bits, as expressed as [14]:

$$\mathbf{BER} = \frac{\mathbf{100}}{\mathbf{L}} \sum_{\mathbf{i=0}}^{\mathbf{L-1}} (\mathbf{1} : \mathbf{W}_{\boldsymbol{i}} \neq \mathbf{W'}_{\boldsymbol{i}}) \qquad \textbf{(6)}$$

where L is the watermark length, W_i is the i^{th} bit of the original watermark and W'_i is the i^{th} bit of the extracted watermark which has been decoded by the error detection-correction code.

5. CONCLUSIONS

This paper introduced extensions to a previously-reported digital audio watermarking system based on enhanced DWT and SVD techniques. The suggested extensions lead to a significant increase in the watermark bit rate. In addition, they improved the watermarked audio signal quality. The performance of the suggested system was evaluated for different audio signals. The experimental results obtained demonstrated better performance compared to the previously reported algorithms with 54% higher SNR, 23.4 times higher watermark bit rate, 92% reduction in the minimum required audio-cover period, and better robustness against watermark benchmark attacks.

REFERENCES

[1] Y. Liu, H. S Yun, J. S Sung, N. S. Kim, "A novel audio fingerprinting scheme based on sub band envelop hashing", *Proc. Asia-Pacific Signal and Information Processing Association*, pp. 813-816, Oct. 2009.

[2] A. Bouridane ,L. Ghouti ,M. Ibrahim and S. Boussakta" Digital Image Watermarking Using Balanced Multiwavelets", *IEEE Transactions on Signal Processing*, Vol. 54, No. 4, pp. 1519-1536, April 2006.

[3] D. Mathivadhani, C.Meena; "Performance Evaluation of key for watermarking using 2-D wavelet transformations", *International Journal of Engineering Science and Technology (IJEST)*, ISSN, 0975-5462 Vol. 3 No. 3 March 2011.

[4] S. Ramakrishnan, T. Gopalakrishnan, K. Balasamy," SVD Based Robust Digital Watermarking For Still Images Using Wavelet Transform", *CCSEA 2011*, CS & IT 02, pp. 155–167, 2011.

[5] N. Lalitha, G. Suresh, V. Sailaja, "Improved Audio Watermarking Using DWT-SVD", *International Journal of Scientific & Engineering Research*, Vol. 2, Issue 6, June, 2011,ISSN 2229-5518.

[6] N. Vejic, T. Seppnen, "Watermark Bit Rate in Diverse Signal Domains", *International Journal of Signal Processing*, Vol.1, 2004.

[7] W-N Lie and L-C. Chang, "Robust and High-Quality Time-Domain Audio Watermarking Based on Low-Frequency Amplitude Modification", *IEEE Transactions on Multimedia*, Vol. 8, No.1, February 2006.

[8] A. Al-Haj, A. Mohammad, L.Bata "DWT–Based Audio Watermarking", *The International Arab Journal of Information Technology*, Vol. 8, No. 3, July 2011.

[9] A. Al-Haj and A. Mohammad, "Digital Audio Watermarking Based on the Discrete Wavelets Transform and Singular Value Decomposition", *European Journal of Scientific Research*, ISSN 1450-216X Vol.39 No.1, pp.6-21, 2010.

[10] M. S. Al-Yaman, M. A. Al-Taee, A.T Shahrour, I.A. Al-Husseini; "Biometric Based Audio Ownership Verification Using Discrete Wavelet Transform and SVD Techniques", *Proc. 8th International Multi-Conference on Systems, Signals and Devices (SSD'11)*, Sousse-Tunisia, March 22-25, 2011.

[11] M. S. Al-Yaman, M. A. Al-Taee, H.A Alshammas; "Audio-Watermarking Based Ownership Verification System Using Enhanced DWT-SVD Technique", *Proc. 9th International Multi-Conference on Systems, Signals and Devices (SSD'12)*, Chemnitz-Germany, March 20-23, 2012.

[12] C. Lee, T. Truong, Y. Chen; "Convolutional Encoding of Some Binary Quadratic Residue Codes", *Proc. the International Multi Conference of Engineers and Computer Scientists*, Hong Kong, March 18 - 20, 2009.

[13] A. Lang, "Stirmark Benchmark for Audio (SMBA)" [Online]:http://amsl-smb.cs.uni-magdeburg.de/smfa/ main. php, Accessed on September 8, 2012.

[14] J. Grody and L. Brutun,"Performance Evaluation of Digital Audio Watermarking algorithms", *Proc. of the 43rd IEEE Midwest Symposium on Circuits and Systems*, 456-9, 2000.

SELF-VERIFIED DNS REVERSE RESOLUTION

Zheng Wang and Rui Wang

China Organizational Name Administration Center

ABSTRACT

*Domain Name System (DNS) reverse resolution is commonly relied on by anti-spam techniques to verify the e-mail origins and by measurements or applications to uncover the host information. But the current practice is not able to clarify the IP addresses with no reverse resolution response and the source verification process is not optimized in terms of network bandwidth and response latency. This paper proposes an explicit scheme to bind A/AAAA resource records (RRs) with their matching PTR RRs by introducing APTR/AAAAPTR RR types. The DNS cache server can automatically switch from forward resolution to reverse resolution when handling the APTR/AAAAPTR RR types. This scheme enables the negative verification if no reverse records are returned for APTR/AAAAPTR records. Furthermore, the analytical and numerical results show that the number of queries and response delay are significantly c*ut by the proposed scheme.

***Keywords*—**Domain Name System, reverse resolution, source verification

1. INTRODUCTION

The Domain Name System (DNS) is a fundamental component of the modern Internet [1], [2], providing a critical link between human users and Internet routing infrastructure by mapping host names to IP addresses. It is relied on by many Internet applications such as web browser and email.

It is generally acknowledged that the implementations and specifications of DNS should have their concepts optimized for the prevailing conditions on the Internet. The most common use of DNS by far is to translate a domain name into its IP address. This process is also known as forward DNS resolution. At times, however, it may be also useful to be able to determine the name of the host given a particular IP address. While sometimes this is required for diagnostic purposes or providing more human-usable data in system logging, more frequently these days it is used for security reasons to trace a hacker or spammer; indeed, many modern mailing systems use reverse mapping to provide simple authentication by using DNS lookup policies, for instance, IP-to-name and name-to-IP to confirm that the specified IP address does represent the indicated host rather than a forged one.

Analogous to normal domain name structure, the reverse DNS database of the Internet is rooted in the Address and Routing Parameter Area (arpa) top-level domain of the Internet. IPv4 uses the in-addr.arpa domain and the ip6.arpa domain is delegated for IPv6. The process of reverse resolving an IP address uses a DNS resource record (RR) type functioning as a pointer from an IP address to a domain name, namely, PTR RR type.

It is expected in [1] that every Internet-reachable host should have a name (thus matching PTR RR). This makes sense because the reverse resolution is viable merely for those IPs that have registered their domain names in DNS. More often than not, the registrant with security awareness would heed to reverse mapping registration bundled with normal domain name registration. Unfortunately, except for a few hosts, the majority of IPs in use has not put their PTR records in the root of in-addr.arpa or in-addr.arpa domains [2].

Due to the coexistence of reverse mapped domain names and non-reverse mapped ones, the validation clients would encounter ambiguity when using reverse resolution. While successful or unmatched reverse response is self evident for the authentication, it is hard to distinguish between non-reverse mapped domain names and forged ones which have no corresponding PTR RRs.

As not all active IP addresses have their matching PTR RRs, it is unsafe to identify the one with negative answers of reverse resolution as the unauthenticated one. Consider there are such A/AAAA RRs which do have their corresponding PTR RRs and do need the reverse resolution as the mandatory means of IP source validation. DNS specifications are supplemented in this paper to be enabled to accommodate the mandatory reverse resolution. For those specific types of A/AAAA RRs, forward resolution is linked automatically with a reverse resolution and only those having matching PTR RRs via reverse resolution may be accepted as the validated ones.

The rest of the paper is organized as follows: Section II discusses previous related work. The proposed scheme is introduced in III. The performance analysis and the numerical results are presented in Section IV and V respectively. Section VI concludes the paper.

2. RELATED WORK

Reverse DNS checks are largely used as one of the anti-spam techniques by mail transfer agents and other applications.

The function of "Report Spam" is implemented in many mail clients to send e-mail to abuse@example.com (where example.com is the sender's domain, possibly found through reverse DNS lookups, when available) [3]. In the "received" header of an email message, the information that a mail transfer agent logs often includes the reverse DNS entry for this IP [4]. The host-names of reverse DNS lookups have been shown to be a strong predictor of spamming likelihood [5].

Reverse DNS resolution is also leveraged to uncover the host related information directly via the IP address.

To obtain the list of authoritative servers to be probed, Pang et al. performed a "reverse-crawl" of the .in-addr.arpa domain, which reverse maps IP addresses to domain names [2]. The reverse DNS lookup can return the domain name associated with the given address, which is meaningful for identifying the application characteristics of the IP address owners, e.g. routers or Internet endpoints. Trestian et al. used reverse DNS lookup for profiling Internet endpoints [6]. However, they also show that the information found via other method is not revealed by the reverse DNS lookup, and this is partly due to the fact that a large amount of the servers do not even have a DNS record. This problem is identical to the source verification dilemma handled by this paper. Mao et al. relied on reverse DNS lookups to identify the AS responsible for the IP address in the traceroute results [7]. The same difficulty lies in the low success rate of reverse DNS lookups for locating all ASes on the path. To defend against DNS rebinding attacks, Jackson et al. proposed a new reverse delegation scheme to authorize a set of host names for an IP address [8]. But the scheme still has no way of advertising the expected existence of reverse mapping records especially in the case that the owner of the reverse delegation of an IP address, the ISP, might not be the owner of the machine at that IP address. To assess the geographic distribution of clients in the testing of the deployed internet source address validation filtering, Beverly et al. sought to identify the location of each client's IP address based on reverse DNS names and heuristics [9]. But the missing reverse entries are short of any explicit messages informing whether they are the expected results for the reverse DNS names. This information is indispensable for the validation of the reverse resolution. If a reverse DNS name is required to have the A/AAAA records and their corresponding PTR records, the unmatched or missing A/AAAA records for the reverse DNS name will invalidate the reverse resolution. But for a reverse DNS name with unenforced association between its A/AAAA records and its corresponding PTR records, the unmatched or missing A/AAAA records can only categorize it as "unknown". To obtain the data sets of legitimate domain names whose characteristics are extracted and compared with the malicious ones, Yadav et al. used reverse DNS crawl of the entire IPv4 address space [10]. Similar to [9], the domain names revealed by reverse resolution are quite weak in term of their authenticity unless the requirement for the matching of forward and reverse resolution is explicitly expressed. Leonard and Loguinov developed a high-performance, Internet-wide service discovery tool, whose main design objectives have been to maximize politeness at remote networks, allow scanning rates that achieve coverage of the Internet in minutes/hours (rather than weeks/months), and significantly reduce administrator complaints [11]. As a metric of intrusiveness of the Internet scan, they introduced DNS lookups to the assessment. This is based on the fact that many specialized tools augment IDS reports and firewall logs with DNS lookups on offending IPs to provide more information on the scanning host to the user. But with no compulsory association between reverse and forward lookups and no explicit advertising of it, the mismatching of reverse and forward resolution is not sufficient to prove the malicious or forged scanning hosts. Muir and Oorschot proposed to obtain the geographic information about an IP address by the lookup of public whois databases [12]. To locate a host with a given IP address, the domain name of the host should first be found by a reverse DNS lookup. This approach has its limitations: One is about completeness, that is, not all IP addresses map to a domain name; The other is about correctness, that is, the reverse DNS lookup response is very possible to be the forged one unless it is checked by the consistency between forward and reverse resolution.

3. SELF-VERIFIED REVERSE DNS RESOLUTION

3.1. APTR/AAAAPTR Resource Record Definition

Two new types of RR are defined: APTR is used for linking A RR with PTR and AAAAPTR is used for linking AAAA RR with PTR RR. Except for the TYPE field, the definition of the other fields of APTR/AAAAPTR is exactly the same as A/AAAA (as specified in [13]).

For each domain name, a zone MAY have either A/AAAA or APTR/AAAAPTR RR. But the presence of both A/AAAA and APTR/AAAAPTR records for a domain MUST be prohibited. This clears out the possibility of response ambiguity between A/AAAA and APTR/AAAAPTR when A/AAAA s requested.

The two newly defined RRs address the class of domain names which must satisfy the following two conditions simultaneously:

1) They have their IPs registered in the reverse mapping database. So the reverse resolution for verification should definitely get successful mapping results. But they need a scheme to inform the verification clients of the existence of PTR RRs for the authentic domain names. We allocate a specific RR type, APTR/AAAAPTR, for this purpose.

2) They require the mapping PTR RRs prefetched by the intermediate servers. Note that an alternative verification means is to send separate requests for A/AAAA records and then PTR records. But it is not an efficient way especially for the clients, as clients have to deliver two requests and maintain the unanswered state for each request. APTR/AAAAPTR response virtually incurs a consecutive

PTR lookup, and this is the intentional verification process of the clients.

The other class of domain names is covered by the original definition of A/AAAA RR types, whose properties can be summarized as follows:

1) They are not registered by the matching PTR RRs. So if the reverse resolution gets a negative answer, the domain name cannot be considered as the forged one.

2) Even though some domain names put their PTR RRs in the reverse mapping DNS tree, still they do not prefer to be verified on one lookup. Alternatively, it is decided by the clients whether to initiate the reverse resolution once receiving a A/AAAA response. This allows for sparing some apace for a relatively more flexible (but probably inefficient) solution in comparison to the mandatory binding by APTR/AAAAPTR.

3.2. Authoritative Server Implementation

Any records, A/AAAA or APTR/AAAAPTR, for a particular domain name in the authoritative zone should be returned to the A/AAAA request. This is based on the fact that the clients cannot foresee whether a domain name has A/AAAA or APTR/AAAAPTR records. So a client would send the A/AAAA request with uncertainty on the replied record type. Accordingly, the algorithm used by the name server to answer the query will be modified as follows:

1) We should search the available zones for the zone which is the nearest ancestor to QNAME. If the whole of QNAME as well as QTYPE is matched for the A/AAAA query, copy the RRs into the answer section. Otherwise, if we have not found the RRs matching the requested A/AAAA type, we then search for APTR/AAAAPTR type and return them if hit in the zone. The implication of APTR/AAAAPTR response also includes the negative answer for A/AAAA query, though it is not particularly signaled unless DNSSEC response is required.

2) If a DNSSEC answer is requested and available, appropriate DNNSEC records should be included in the answer message. It should be noted that the positive answer, with A/AAAA or APTR/AAAAPTR, virtually should notify the non-existent of each other due to mutual exclusivity of A/AAAA and APTR/AAAAPTR RRs. Therefore we have to use the NSEC RRs as the negative answer together with the positive answer, one for A/AAAA and one for APTR/AAAAPTR. The signed RRs for any of them should be added to explicitly indicate the authenticated existence and non-existence. There are also circumstances that clients do know the presence of APTR/AAAAPTR records for a domain name, or they are only interested in APTR/AAAAPTR records rather A/AAAA records. For such circumstances, the most appropriate request is clearly solely for the APTR/AAAAPTR records. The algorithm of the name server adheres to the one-to-one QTYPE matching policy. In other words, the answer should be constrained by the existence of APTR/AAAAPTR itself. e.g. if the question is

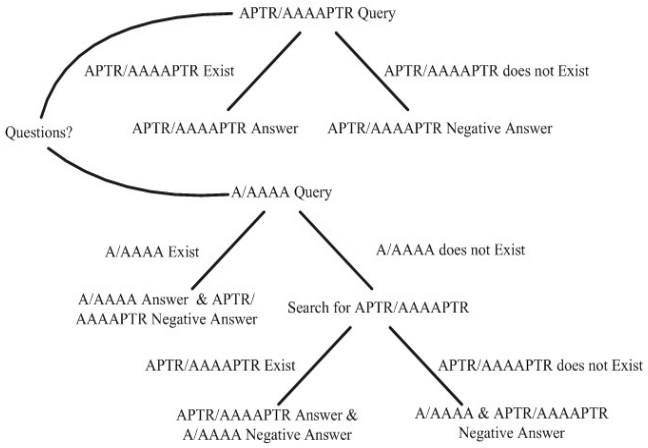


Figure 1. The possibility tree for the implementation of authoritative servers

for a domain name "example.example.", we can expect that the answer takes the NSEC RR format of "example.example. 3600 NSEC *.example. A RRSIG NSEC" for APTR/AAAAPTR QTYPE while the possible NSEC RR contained in the response to A/AAAA QTYPE should like "example.example. 3600 NSEC *.example. RRSIG NSEC".

The possibility tree for all responses is shown in Figure 1.

3.3. Recursive Server Implementation

For simplicity, we do not address the caching of recursive servers first. And it is to be specified in the end of this section.

If a recursive request for A/AAAA comes, the recursive server should first search the cache for the desired data. If the data is not in the cache, it should send a normal request for A/AAAA. The afterwards lookup process is dependent of the response:

1) If A/AAAA RR is returned in the response, the recursive server caches the A/AAAA RR and the NODATA error for APTR/AAAAPTR as well as returning the response back to the client. All DNSSEC related RRs should also be cached if necessary.

2) If the response has APTR/AAAAPT RR in the answer section, the recursive server caches the APTR/AAAAPT RR and the NODATA error for A/AAAA (DNSSEC related RRs should also be cached if necessary). And this is followed by a lookup for the matching PTR RR, emitting a request for the in-addr.arpa/ip6.arpa domain in line with the IPv4/IPv6 address in A/AAAA RR. When the response for the reverse resolution is received, it should be placed in the answer section of the DNS response for the client. It is to be noted that the DNS response thereby consists of at least three meaningful components: APTR/AAAAPT RR, NODATA error for A/AAAA, the answer of the reverse resolution. Here the last component deserves further analysis as follows:

2.1) If the answer of the reverse resolution is a matching PTR RR set, just put the PTR RR set in the answer section

of the response. The corresponding RRSIG RR set is, of course, to be included in DNSSEC scenarios.

2.2) If the answer of the reverse resolution is a NODATA error of PTR RR, it is expressed by the empty or NSEC RR (for DNSSEC). The corresponding RRSIG RR set should also be contained in the response.

2.3) If the answer of the reverse resolution is a NXDOMAIN error of PTR RR, the empty should also be allocated for it in the answer section of the response. Undoubtedly, a pair of NSEC RRs covering the reverse domain is to be added in the response. Interestingly, the only problem puzzles us is how to set the RCODE of the DNS response message. For the domain name wanted by the origin query, the RCODE should be set NOERROR for the domain name does exist in the authoritative zone. Whereas considering the reverse mapping automatically linked to APTR/AAAAPTR, the result clearly pronounces NXDOMAIN for the RCODE. However, the NXDOMAIN RCODE may be more appropriate than NOERROR since it is more informative. Because the NOERROR RCODE is virtually indicated by the existence of A/AAAA record, we should spare RCODE for PTR RR. Otherwise, the client would have no knowledge of the response code specific for PTR RR.

3) If a negative answer with the RCODE of NXDOMAIN is returned in the response, the recursive server does negative caching meanwhile returns the response back to the client.

If a recursive request for APTR/AAAAPTR arrives, the work of the recursive server is comparatively simple. It should first send a query for APTR/AAAAPTR. Then the following steps are determined by the response:

1) If APTR/AAAAPTR RR is returned in the response, the response message of the recursive server is analogy to that for A/AAAA request described in 2) except for the slightly different question section.

2) If the NOERROR error for APTR/AAAAPTR RR is replied, the negative caching is performed by the recursive server. And the simple NOERROR response goes back to the client.

3) If a negative answer with the RCODE of NXDOMAIN is returned in the response, the recursive server does negative caching meanwhile returns the response back to the client.

3.4. Client Implementation

The clients should interpret the response properly. We classify all cases according to the request and the response as follows:

1) If the request is for A/AAAA, the negative answer with NOERROR RCODE (with empty answer section) indicates non-existence of both A/AAAA and APTR/AAAAPTR.

2) If the request is for A/AAAA and the answer does replies A/AAAA, no particular actions are needed to be specified here.

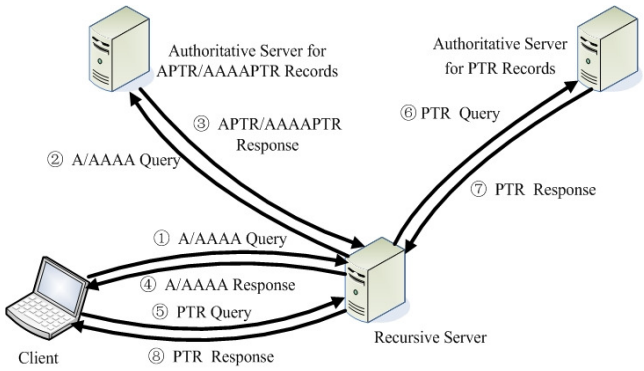


Figure 2. The data flow of the current practice for source verification

3) If the request is for A/AAAA and the response includes APTR/AAAAPTR in its answer section, the first information delivered to the client is that A/AAAA does not exist while APTR/AAAAPTR is found for the queried domain name. And if a proper PTR RR is seen in the answer section, the domain name can be taken as authenticated. Otherwise, negative answer on the reverse resolution informs the client of the failed validation on the domain name.

4) If the request is for APTR/AAAAPTR, the negative answer with NOERROR RCODE (with empty answer section) tells the client about nonexistence of APTR/AAAAPTR.

5) If the request is for APTR/AAAAPTR and the response has APTR-/AAAAPTR in its answer section, and if a proper PTR RR is found in the answer section, the domain name can be taken as authenticated. Otherwise, negative answer on the reverse resolution means the failed validation on the domain name.

4. PERFORMANCE ANALYSIS

Besides facilitating the source verification of domain names, the proposed scheme is also favorable for its efficient handling of the two adjacent DNS lookups when the source verification process is involved in one domain name lookup. In this section, we present thorough quantitative analysis as well as numerical results for the performance comparison between our proposed scheme and the current practice.

For simplicity, we neglect the resolution time cost of the upper level above the requested domain name in the DNS hierarchy. This means that the recursive server has known the DNS servers (or had them in its cache) for the requested domain name and its correspondent PTR record before issuing the request. This is a reasonable assumption based on the measurements that NS records tend to have much longer TTL values than A records [14].

4.1. Performance Analysis without Cache Modeling

To clearly elaborate the performance gain of the scheme, we first illustrate the dada flow of the current practice of source verification in Figure 2.

Let the network latencies between the client and recursive server, the recursive server and the authoritative server for A/AAAA records, the recursive server and the authoritative server for PTR records be d_c, d_a and d_p respectively. And we assume that the processing delays of any queries staying in the recursive server, the authoritative server for A/AAAA records and the authoritative server for PTR records are p_r, p_a and p_p respectively. In terms of the overall query latency, the performance of the client's source verification, D_c, can be written as

$$D_c = 4d_c + 2d_a + 2d_p + 4p_r + p_a + p_p \quad \textbf{(1)}$$

The number of queries sent by the client for the source verification, Q_c, yields

$$Q_c = 2 \quad \textbf{(2)}$$

Therefore, we can obtain the average query rate required for one domain name, R_c, as follows

$$R_c = Q_c / D_c \quad \textbf{(3)}$$

We also discuss the impacts on the load of the recursive server, since the proposed scheme and current practice make difference here. As long as the incoming query has not been ultimately answered, the recursive server has to maintain its unanswered state (mostly in its memory), waiting for the response, until time out. Apparently, system resources are consumed by the query processing and more sojourn time of queries in the recursive server results in more resource consumptions. The overall sojourn time of query in the recursive server is denoted by D_r and D_r yields

$$D_r = 2d_a + 2d_p + 4p_r + p_a + p_p \quad \textbf{(4)}$$

Eq.(1)-(4) do not take caching into consideration. If caching takes effects on the recursive server, the query latency as well as the query sojourn time will be cut off. We use D_c^A and D_c^P to denote the latency if A/AAAA and PTR hits cache respectively. They can be expressed as

$$D_c^A = 4d_c + 2d_p + 3p_r + p_p \quad \textbf{(5)}$$

and

$$D_c^P = 4d_c + 2d_a + 3p_r + p_a \quad \textbf{(6)}$$

Moreover, if both A/AAAA and PTR hit cache, we have the minimum latency, D_c^H, as follows

$$D_c^H = 4d_c + 2p_r \quad \textbf{(7)}$$

The overall sojourn time of query in the recursive server for A/AAAA, PTR and both of them cache hitting, D_r^A, D_r^P and D_r^H respectively, can be given as

$$D_r^A = 2d_p + 3p_r + p_p \quad \textbf{(8)}$$

$$D_r^P = 2d_a + 3p_r + p_a \quad \textbf{(9)}$$

and

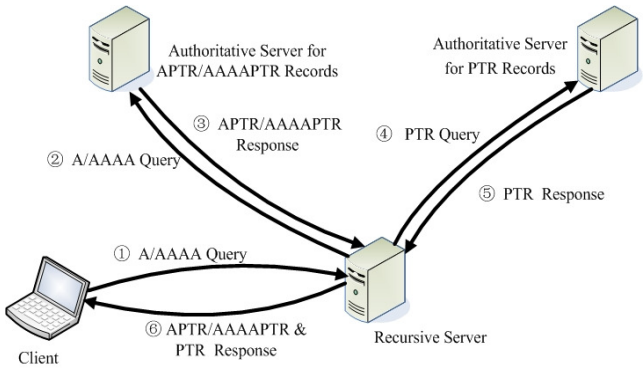


Figure 3. The data flow of the proposed scheme for source verification

$$D_r^H = 2p_r \quad \textbf{(10)}$$

The data flow of the proposed scheme for source verification is shown in Figure 3.

We use the same definitions of the intermediate latencies as described above. The overall query latency of the client, Γ_c, can be written as

$$\Gamma_c = 2d_c + 2d_a + 2d_p + 3p_r + p_a + p_p \quad \textbf{(11)}$$

The number of queries sent by the client for the source verification, N_c, yields

$$\mathrm{N}_c = 1 \quad \textbf{(12)}$$

The average query rate required for one domain name, $\Re_c$, is

$$\Re_c = \Gamma_c / \mathrm{N}_c \quad \textbf{(13)}$$

The overall sojourn time of query in the recursive server is denoted by Γ_r and Γ_r yields

$$\Gamma_r = 2d_a + 2d_p + 3p_r + p_a + p_p \quad \textbf{(14)}$$

Considering caching effects on the recursive server, we use Γ_c^A and Γ_c^P to denote the latency if APTR/AAAAPTR and PTR hits cache respectively. They can be written as

$$\Gamma_c^A = 2d_c + 2d_p + 2p_r + p_p \quad \textbf{(15)}$$

and

$$\Gamma_c^P = 2d_c + 2d_a + 2p_r + p_a \quad \textbf{(16)}$$

If both APTR/AAAAPTR and PTR hit cache, we have the minimum latency, Γ_c^H, as follows

$$\Gamma_c^H = 2d_c + p_r \quad \textbf{(17)}$$

The overall sojourn time of query in the recursive server for A/AAAA, PTR and both of them cache hitting, Γ_r^A, Γ_r^P and Γ_r^H respectively, can be given as

$$\Gamma_r^A = 2d_p + 2p_r + p_p \quad \textbf{(18)}$$

$$\Gamma_r^P = 2d_a + 2p_r + p_a \quad \textbf{(19)}$$

and

$$\Gamma_r^H = p_r \quad \textbf{(20)}$$

4.2. Overall Performance Analysis Integrated by Cache Modeling

Based on the above analysis in Subsection 4.1, the overall response delay is dependent of the cache hit rates of A/AAAA(or APTR/AAAAPTR) and PTR records. For further quantitive analysis, we take the factors relevant to the cache hit rate into accounts and examine their impacts on the performance improvement.

Let $N(T)$ equal the number of queries for the given record in the interval (0, t]. Let the value of the time-to-live (TTL) of the record be T. We have such formula for the cache hit rate as follows [15]:

Theorem 1. If the inter-query times to a given record are proper, nonnegative, independent and identically distributed random variables, whose mean may be infinite, then

$$H[T]=\frac{E[N(T)]}{E[N(T)]+1} \quad (21)$$

Where $E[N(T)]$ is the expected number of queries falling into the interval of TTL.

We denote the mean query rate for the given record by R. So $E[N(T)]$ is

$$E[N(T)]=RT \quad (22)$$

Let the TTLs of A/AAAA(or APTR/AAAAPTR) and PTR records be TA and TP respectively. Substitute T in Eq.(21) and (22) by TA and TP , we can obtain the cache hit rates of A/AAAA(or APTR/AAAAPTR) and PTR records, H(TA) and H(TP). Assume that the cache hit rates of A/AAAA(or APTR/AAAAPTR) and PTR records are independent of each other. So the expected overall lookup latency of client for the current practice can be expressed as

$$\begin{aligned}\bar{D}_c=&\left(1-H(T_A)\right)\left(1-H(T_P)\right)D_c+H(T_A)\left(1-H(T_P)\right)D_c^A\\&+\left(1-H(T_A)\right)H(T_P)D_c^P+H(T_A)H(T_P)D_c^H\end{aligned} \quad (23)$$

The expected overall query sojourn time in the recursive server for the current practice can be written as

$$\begin{aligned}\bar{D}_r=&\left(1-H(T_A)\right)\left(1-H(T_P)\right)D_r+H(T_A)\left(1-H(T_P)\right)D_r^A\\&+\left(1-H(T_A)\right)H(T_P)D_r^P+H(T_A)H(T_P)D_r^H\end{aligned} \quad (24)$$

Given the mean query rate for the given record as R, we can derive the mean number of queries maintained in the recursive server, W_r , as follows

$$W_r=R\bar{D}_r \quad (25)$$

Massive amounts of queries maintained in the recursive server virtually put a heavy burden on it. Therefore, the performance advantage of our proposed scheme can also be drawn from the comparison of W_r between them.

Similar to the results derived for the current practice, we can describe the performance of the proposed scheme by the following metrics: Γ_c or the expected overall lookup latency of client, Γ_r or the expected overall query sojourn time in the recursive server, Ψ_r or the mean number of queries maintained in the recursive server, which can be expressed respectively as

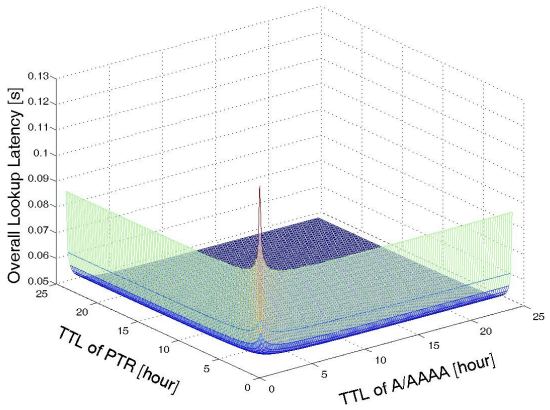


Figure 4. The overall lookup latency of the current practice vs. T_A and T_P

$$\begin{aligned}\bar{\Gamma}_c=&\left(1-H(T_A)\right)\left(1-H(T_P)\right)\Gamma_c+H(T_A)\left(1-H(T_P)\right)\Gamma_c^A\\&+\left(1-H(T_A)\right)H(T_P)\Gamma_c^P+H(T_A)H(T_P)\Gamma_c^H\end{aligned} \quad (26)$$

$$\begin{aligned}\bar{\Gamma}_r=&\left(1-H(T_A)\right)\left(1-H(T_P)\right)\Gamma_r+H(T_A)\left(1-H(T_P)\right)\Gamma_r^A\\&+\left(1-H(T_A)\right)H(T_P)\Gamma_r^P+H(T_A)H(T_P)\Gamma_r^H\end{aligned} \quad (27)$$

$$\Psi_r=R\bar{\Gamma}_r \quad (28)$$

4.3. Numerical Results

The expressions derived in Subsection 4.1 and 4.2, though suggestive and clear, are not so intuitive in terms of performance evaluation. So without loss of generality, we tend to provide more visual representation of the performance improvement by the approximated data in a simple model.

The data we used tends to be consistent with the measurements [14], which shows that a large proportion of query load comes from requests for some most popular domain names. Hence we quantify the query rate for these popular domain names as R = 1000/h, which is derived by the relevant data [14].

The network delays between the recursive server and the authoritative server, d_a and d_P , is both set as 200 ms. Since the client usually would choose its local recursive server rather than those in the wild, we set the network delay between the client and the recursive server, d_c , as a relatively small value, say, 10ms. Let the server processing delays for the two authoritative servers and recursive server be an equal value as $p_a = p_p = p_r$ = 5ms.

We first present the overall lookup latency as a function of T_A and T_P . We can see in [14] that over 90 percent of A or PTR records have their TTLs fall below one day. So we set the TTLs of A and PTR records range from 0.1 hour to 24 hour. Figure 4 and Figure 5 show the overall lookup latency for the current practice and for the proposed scheme respectively. It is clear that the proposed scheme achieves significantly lower lookup delay than the current practice.

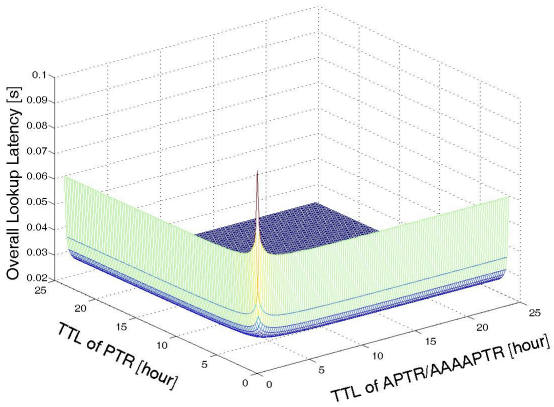


Figure 5. The overall lookup latency of the proposed scheme vs. T_A and T_P

We investigate the overall lookup latency as a function of d_a and d_p. According to the previous measurements [14], the overwhelmingly majority of network delays between the recursive server and the authoritative server fall into the range between 10 ms and 500 ms. So we vary d_a and d_p in the range and show lookup latency results. Figure 6 illustrates the comparative overall lookup latencies between the current practice and the proposed scheme. It shows that the proposed scheme decreases the overall lookup latency in comparison with the current practice.

To summarize the results presented above, we can observe that the proposed scheme out-performs the current practice in terms of providing shortened lookup latency for all values of T_A, T_P, d_a and d_p.

We then illustrate the performance gain of the proposed scheme as to the query sojourn time in the recursive server. The settings of the two pairs of parameters, T_A, T_P, d_a and d_p remain the same as above. Figure 7 and Figure 8 show the query sojourn time in the recursive server for the current practice and for the proposed scheme respectively. Figure 9 compares the query sojourn time in the recursive server for the current practice and for the proposed scheme. All of these results support the performance advantage of the proposed scheme over the current practice.

5. CONCLUSIONS

DNS reverse resolution is commonly relied on by the source IP verification as well as host information revealing applications, but there is no explicit way to inform the binding of A/AAAA and corresponding PTR records and perform automatic resolution switch. In this paper, we propose a glued scheme to supplement basic DNS specifications accommodating the chained forward and reverse resolution. Besides its function as binding notification, the proposed scheme is also efficient in terms of its capability of query delay cut and recursive server load relief. Performance analysis and numerical results show that the proposed scheme significantly outperforms the current practice.

ACKNOWLEDGMENT

This work was supported by the National Key Technology R&D Program of China (No. 2012BAH16B00) and the National Science Foundation for Distinguished Young Scholars of China (No. 61003239).

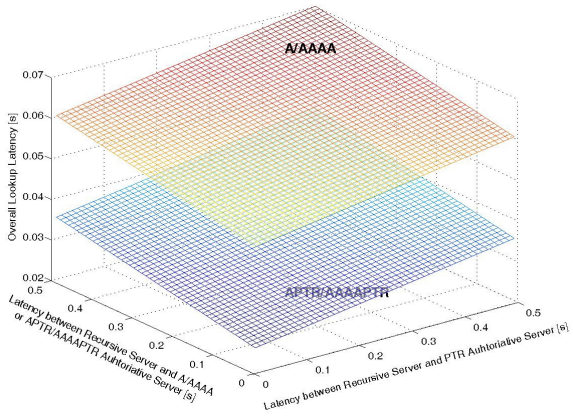


Figure 6. Comparison of the overall lookup latencies vs. d_a and d_p

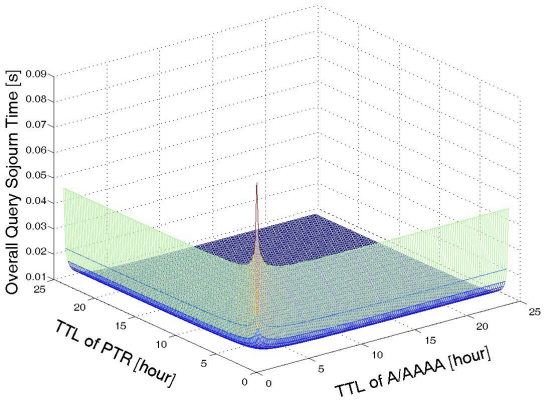


Figure 7. The query sojourn time of the current practice vs. T_A and T_P

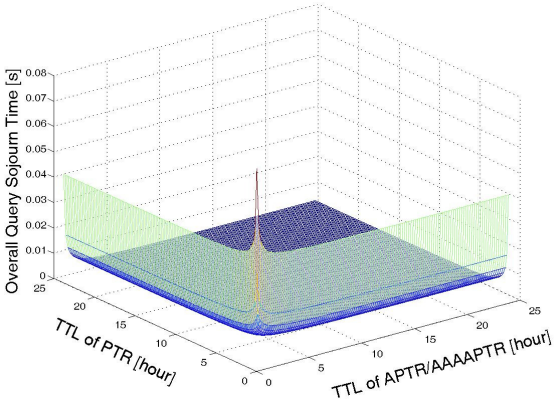


Figure 8. The query sojourn time of the proposed scheme vs. T_A and T_P

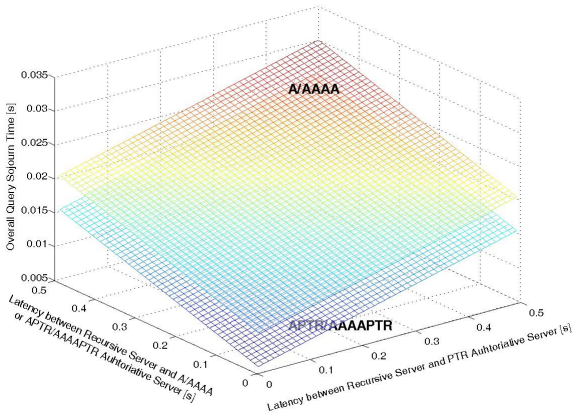


Figure 9. Comparison of the query sojourn time vs. d_a and d_p

REFERENCES

[1] Barr D. Common DNS practiceal and Configuration Errors. RFC 1912. 1996.

[2] Pang J, Hendricks J, Akella A, Prisco RD, Maggs B, Seshan S. Availability, usage, and deployment characteristics of the domain name system. In: Proc. ACM IMC'04, 2004. p. 1-14.

[3] Goodman JT, Rounthwaite R. Stopping outgoing spam. In: Proc. ACM Conference on Electronic commerce (EC'04), 2004. p. 30-39.

[4] Eggendorfer T. Methods to identify spammers. In: Proc. International Conference on Forensic Applications and Techniques in Telecommunications, Information, and Multimedia and Workshop (e-Forensics'08), 2008.

[5] Qian Z, Mao Z, Xie Y, Yu F. On network-level clusters for spam detection. In: Proc. NDSS'10, 2010.

[6] Trestian I, Ranjan S, Kuzmanovic A, Nucci A. Googling the internet: profiling internet endpoints via the world wide web. IEEE/ACM Trans Networking. 2010;18(2):666-679.

[7] Mao ZM, Rexford J, Wang J, Katz RH. Towards an accurate AS-level traceroute tool. In: Proc. ACM SIGCOMM'03, 2003. p. 365-378.

[8] Jackson C, Barth A, Bortz A, Shao W, Boneh D. Protecting browsers from DNS rebinding attacks. ACM Trans Web. 2009;3(1):201-226.

[9] Beverly R, Berger A, Hyun Y, Claffy K. Understanding the efficacy of deployed internet source address validation filtering. In: Proc. ACM IMC'09, 2009. p. 356-369.

[10] Yadav S, Reddy AKK, Reddy ALN, Ranjan S. Detecting algorithmically generated malicious domain names. In: Proc. ACM IMC'10, 2010. p. 356-369.

[11] Leonard D, Loguinov D. Demystifying service discovery: implementing an internet-wide scanner. In: Proc. ACM IMC'10, 2010. p. 48-61.

[12] Muir JA, Oorschot PCV. Internet geolocation: Evasion and counterevasion. Computing Surveys. 2009;42(1):4:1-4:23.

[13] Barr D. Common DNS practiceal and Configuration Errors. RFC 1912. 1996.

[14] Jung J, Sit E, Balakrishnan H, Morri R. DNS performance and the effectiveness of caching. IEEE/ACM Trans Networking. 2002;10(5):589-603.

[15] Jung J, Berger AW, Balakrishnan H. Model TTL-based Internet caches. In: Proc. IEEE INFOCOM'03, 2003. p. 417-426.

A PERIODIC COMBINED-CONTENT DISTRIBUTION MECHANISM IN PEER-ASSISTED CONTENT DELIVERY NETWORKS

Naoya Maki and Ryoichi Shinkuma

Graduate School of Informatics,
Kyoto University,
Yoshidahonmachi, Sakyo-ku,
Kyoto-shi, Kyoto, 606-8501 Japan

Tatsuya Mori, Noriaki Kamiyama and Ryoichi Kawahara

NTT Network Technology Laboratories,
NTT Corporation,
Midoricho 3-9-11, Musashino-shi,
Tokyo, 180-8585 Japan

ABSTRACT

The concept of peer-assisted content delivery networks (CDNs) lets other nearby altruistic clients forward requested content files instead of the source servers, which works to localize overall traffic. Our prior work proposed a traffic engineering scheme to localize traffic in peer-assisted CDNs. To induce altruistic clients to download content files that are most likely to contribute to localizing network traffic, this scheme combines the content files and allows them to obtain the combined content file while keeping the price unchanged from the single-content price. Although we have discussed how much traffic only a set of combined content files can theoretically reduce, we can expect further traffic localization by distributing combined content files to multiple altruistic clients. This paper proposes a periodic combined-content distribution mechanism based on our scheme. This mechanism determines when combined content files should be provided by considering the network cache state. Computer simulations confirmed that our new mechanism could double performance.

Keywords— Content distribution networks, Web service, Communication system traffic control

1. INTRODUCTION

Online content delivery services, which sell music, movies, and application software, have been widely used over the last decade [1][2]. However, with the rapid increase in content volume, the traffic generated by delivering requested content files has been increasing. Service providers and network operators are under pressure to minimize the amount of traffic in their transactions in order to lower the cost charged for bandwidth and the cost for network infrastructure, respectively. Peer-assisted content delivery networks (CDNs) have been proposed as a way of minimizing traffic and they have extensively been used [3]-[5]. CDNs distribute storage storing content files replicated from the source server and they direct client requests to the replicas nearest the clients [6][7]. However, from the service-providers' standpoint, CDNs do not always benefit because it costs too much to deploy and maintain distributed storage or to rent it. Peer-assisted CDNs direct client requests to the nearest replicas as in normal CDNs, but they do not need to deploy or borrow distributed storage since the replicas are stored in the cache of one of millions of clients. Such a distributed approach is essential not only in content delivery but also in other emerging network applications toward our sustainable communities.

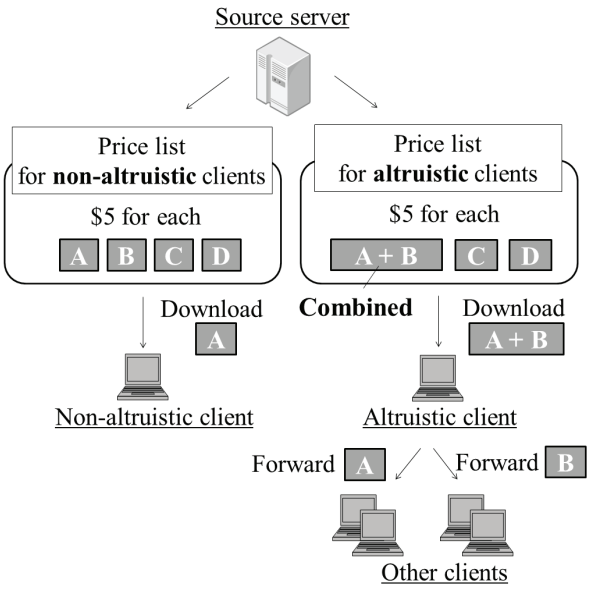


Figure 1. Example of content combination. Combination of content A and B is available for altruistic clients for $5 as content. Combination is likely to be requested in local networks.

Our prior work proposed a traffic engineering scheme with content combination for altruistic clients as outlined in Fig. 1. This scheme combined desired content files while keeping the price equal to that for single-content to induce altruistic clients to request them. The main advantage of our approach is that we can expect sustainable contributions from altruistic clients by using sustainable incentive. Our prior work discussed how much traffic only a set of combined content files could theoretically reduce through numerical analysis

[8]. This analysis confirmed a set of combined content files could reduce the overall amount of network traffic by about 10% comparing with the peer-assisted CDN model without our scheme. However, through the service, we could let multiple altruistic clients cache combined-content files to further localize traffic.

In this paper, we propose a periodic combined-content distribution mechanism to increase the gain in traffic localization. The main problem our mechanism solves is that although we may expect a large localization of traffic by distributing combined content files to multiple altruistic clients, a large amount of traffic is instantaneously generated when combined content files are transferred. Therefore, we have to optimize the period to increase gain by using periodic combined-content distribution.

The two main contributions of this paper are: i) we present the design of a periodic combined-content distribution mechanism that improves traffic localization, and ii) we report simulations that verified our distribution mechanism could double performance compared with when only a set of combined content files was considered.

2. PEER-ASSISTED CONTENT DELIVERY NETWORKS

2.1. RELATED WORK

Content placement in peer-to-peer (P2P) networks and CDNs are long-standing and well-studied problems [9]-[14]. In P2P networks, distributed approaches have been considered since each peer decides whether to cache the received content [9]-[11]. On the other hand, in CDNs, the problem falls under the policy and algorithm design since content-service providers manage caching networks and control content placement [12]-[14]. In peer-assisted CDNs, unlike P2P networks and CDNs, a central entity can attempt to control cache placement [15], but clients may refuse to cache the directed content files because peer-assisted CDNs owe clients storage resources. As we will explain in Section. 2.3, our approach does not directly control the cache in the clients but only induces altruistic clients to cache desired content for traffic localization.

It is generally well-known that most clients in P2P applications have non-altruistic attitudes concerning contributing to services [16][17]. Therefore, motivating free-riders has been the purpose of the previous work on incentive mechanisms [18]-[21]. Since we assume a paid service in this paper, all clients have good reason not to contribute to the service as content servers. Therefore, we should expect a limited number of altruistic clients to contribute to the service [22]. Unlike previous efforts, our purpose was simply to induce altruistic clients to request specific content that would likely be requested on local networks to reduce traffic. The form of the incentives was another factor in which we were interested. Some systems give incentives as service quality [18][19], and others provide monetary incentives [20][21]. However, it is mathematically unclear how much a certain level of improved service quality or money would increase the probability that free riders would contribute to networks. Our incentive by combining content, which will be explained in Sect. 2.3, is straightforward. As we will explain in Sect. 2.2, since clients are charged at a fixed rate to obtain coupons at every period and the combined content is just an electronic copy of the original file, unlike monetary incentives, we do not need to consider how much we gained or lost economically by giving incentives to clients.

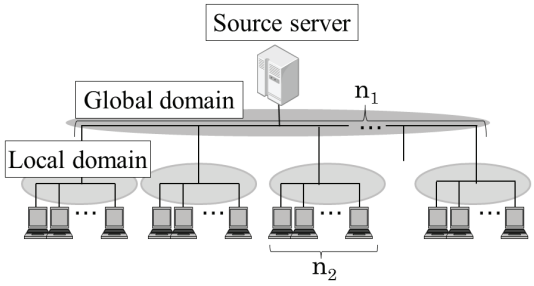


Figure 2. Network model.

2.2. Our service model

Here, we assume a content service only deals with such common entertainment as music and movies. We denote the set of all the content files that can be purchased from the service as $\mathbb{C}$. The content popularity in our model is preliminarily estimated by an existing method [15] and the request probability of content follows Zipf's law [23]. The request probability of content that has the i-th highest popularity, P_i, is:

$$P_{\mathrm{i}} = \frac{\frac{1}{i}}{\sum_{j \in \mathbb{C}} \frac{1}{j}}. \quad (1)$$

Figure 2 outlines the network model we assumed. This is a hierarchical model where the local domains, global domains, and source servers are in the bottom, middle and top layers, respectively. The n_1 and n_2 in Fig. 2 correspond to the number of local domains and the number of clients in each domain.

In general, there could be mainly two charging structures: pay-per-view (PPV) and fixed rate. In PPV, clients pay every time they view content, which does not suit our mechanism because it is not clear how much combined content files should be so as not to decrease the revenue of the service provider. Therefore, in our assumed service, clients are charged at a fixed rate and given a fixed number of coupons at every period. They can purchase a content file or a set of combined content files in exchange for a coupon. Even when clients retrieve their purchased content from their own caches, they must use a coupon. The fixed charge is essential in our system so that providing combined content files at a single content price does not reduce the revenues of the service provider. We assumed coupons would be provided frequently enough compared with the average interval between client requests.

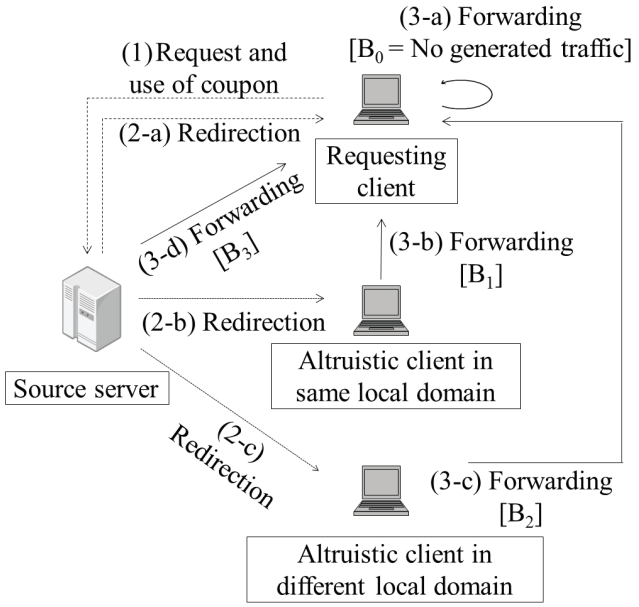


Figure 3. Peer-assisted CDN model.

Figure 3 has the flow for our assumed peer-assisted CDN model. (1) The client first requests content and uses a coupon. (2) Then, the source server makes a transaction on the content charge and redirects the request to the cached location that minimizes traffic; the request is redirected with four priorities: a) the client's own cache, b) the cache at other altruistic clients in the local domain to which the requester belongs, c) the cache at other altruistic clients in the global domain, and d) the source server. (3) When the requester retrieves the required content from the redirected locations of a), b), c), or d), traffic B_0, B_1, B_2, and B_3 are generated ($B_0(=0) < B_1 < B_2 < B_3$). Each client declares that s/he works as an altruistic or non-altruistic client before joining the service. If altruistic clients are redirected the requests, they have to forward the requested files.

In the flow of transaction, since all client requests are handled by the service provider in a centralized manner in peer-assisted CDNs, we can assume that the system is time-slotted. Therefore, only a client is permitted to request and download a content file at each unit-time and the unit-time can be considered to be the average interval between client requests.

2.3. Our incentive mechanism

As explained in Sect. 1, our key idea was combined content that would induce altruistic clients to cache content files that were likely to be requested in local networks. In this paper, we simply consider the request probability of a set of combined content files, P_{comb}, to be:

$$P_{\mathrm{comb}} = \sum_{i \in \mathbb{B}} P_{\mathrm{i}}, \tag{2}$$

where $\mathbb{B}$ is a set of the combined content files. Figure 4 has an example of the request probability for combined content

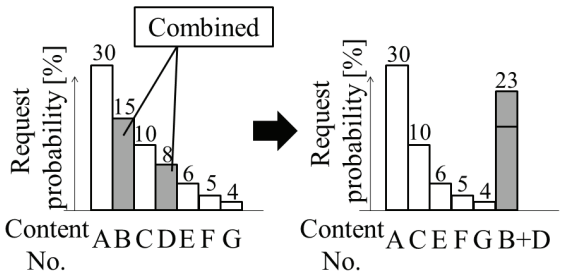


Figure 4. Example of request probability of combined content.

files. The request probability of the combination of content B and D equals the sum of the request probabilities of those files. As we will explain in Sect. 4, we have left it to future work as to how we will model the request probability of combined content files.

2.4. Our content selection algorithm

2.4.1. Problem formulation

When the requesting client is an altruistic client, the content file could be a set of combined files. Combined content files under the time-slotted system are not simultaneously provided for multiple altruistic clients.

Suppose that a client makes a request in a unit time and altruistic client u is going to request a content file at time t. Altruistic client u obtains a new bundle of combined content files $\mathbb{B}_{\mathrm{u}}$ and removes a set of content files $\mathbb{D}_{\mathrm{u}}$ from its cache $\mathbb{C}_{\mathrm{u}}^{\mathrm{t}}$ to make space for $\mathbb{B}_{\mathrm{u}}$. To determine the combination of content files, the service provider solves the optimization problem written as:

$$\max_{\mathbb{B}_{\mathrm{u}}, \mathbb{D}_{\mathrm{u}}} T\left(\zeta\left(S_{\mathrm{t}}\right) - \zeta\left(S_{\mathrm{t+1}}\right)\right) - \eta\left(\mathbb{B}_{\mathrm{u}}\right) \tag{3}$$

$$\begin{aligned}
\text{s.t.} \quad S_{\mathrm{t}} \cap S_{\mathrm{t+1}} &= \left(\mathbb{C}_1, \cdots, \mathbb{C}_{\mathrm{u-1}}, \mathbb{C}_{\mathrm{u+1}} \cdots, \mathbb{C}_{\mathrm{N}}\right) \\
S_{\mathrm{t}} &= \left(\mathbb{C}_1, \cdots, \mathbb{C}_{\mathrm{u}}^{\mathrm{t}}, \cdots, \mathbb{C}_{\mathrm{N}}\right) \\
S_{\mathrm{t+1}} &= \left(\mathbb{C}_1, \cdots, \mathbb{C}_{\mathrm{u}}^{\mathrm{t+1}}, \cdots, \mathbb{C}_{\mathrm{N}}\right) \\
\mathbb{C}_{\mathrm{u}}^{\mathrm{t+1}} &= \mathbb{C}_{\mathrm{u}}^{\mathrm{t}} + \mathbb{B}_{\mathrm{u}} - \mathbb{D}_{\mathrm{u}} \\
\zeta\left(S_{\mathrm{t}}\right) &= \sum_{i \in \mathbb{N}} \sum_{j \in \mathbb{C}} \frac{B_{\mathrm{ij}} P_{\mathrm{j}}}{N} \\
\eta\left(\mathbb{B}_{\mathrm{u}}\right) &= \sum_{j \in \mathbb{B}_{\mathrm{u}}} B_{\mathrm{uj}},
\end{aligned}$$

where N and $\mathbb{N}$ indicate the total number of clients and a set of those clients, respectively; S_{t} is the state of the cache in the entire network at time t; $\zeta(S_{\mathrm{t}})$ indicates the traffic generated when the cache state is S_{t}; $t+1$ means the time just after the cache of client u has been replaced; $\eta\left(\mathbb{B}_{\mathrm{u}}\right)$ indicates how much traffic is increased by downloading a set of content files $\mathbb{B}_{\mathrm{u}}$ compared with downloading a single file; B_{ij} represents the traffic generated when client i requests and retrieves content j. As we can see from the definition in Eq. (3), the

difference between S_{t} and $S_{\mathrm{t+1}}$ is $\mathbb{B}_{\mathrm{u}} - \mathbb{D}_{\mathrm{u}}$. $\zeta\left(S_{\mathrm{t}}\right) - \zeta\left(S_{\mathrm{t+1}}\right)$ is how much traffic will be reduced from t to $t+1$ as a result of caching and discarding $\mathbb{B}_{\mathrm{u}}$ and $\mathbb{D}_{\mathrm{u}}$. If the cached content files at clients in the network do not change during period T, $T(\zeta\left(S_{\mathrm{t}}\right) - \zeta\left(S_{\mathrm{t+1}}\right))$ indicates the amount of reduced traffic during T; T equals T unit times while time t means t-th slot. This is an approximation because, in fact, cached files in the network change during T. The approximation will be described in detail in Section. 2.5. However, downloading $\mathbb{B}_{\mathrm{u}}$ instantaneously generates a large amount of traffic, which is represented as $\eta\left(\mathbb{B}_{\mathrm{u}}\right)$ in Eq. (3).

2.4.2. Selection algorithm

This section briefly describes how we find the sets of combined and discarded content files $\mathbb{B}_{\mathrm{u}}$ and $\mathbb{D}_{\mathrm{u}}$ that satisfy Eq. (3) using three steps. We assume that every client has already cached a sufficiently large number of content files and there is no empty space in their cache capacity.

Step 1 : Optimization of $\mathbb{B}_{\mathrm{u}}$
Step 1-(a) : We calculate the expected traffic reduction by every content j ($j \in \mathbb{C}$) given by:

$$E_{\mathrm{j}} = T \cdot \Delta^{-}_{\mathrm{uj}} - B_{\mathrm{uj}}, \tag{4}$$

where Δ^{-}_{uj} is the amount of traffic that is expected to be reduced at time $t+1$ if altruistic client u requests and retrieves content j at time t. In Eq. (4), as we discussed in Sect. 1, we consider the fact that, as we increase the number of combined content files, more instantaneous traffic is generated.

Step 1-(b) : We score every content $P_{\mathrm{j}}E_{\mathrm{j}}$ and sort them in descending order. This is because P_{comb} defined in Eq. (2) should also be considered because content j will not be effective if it is not actually requested and cached by altruistic client u. This step works for increasing the request probability of combined content files that will reduce a large amount of traffic at time $t+1$.

Step 2 : Optimization of $\mathbb{D}_{\mathrm{u}}$
We score the cached content of altruistic client u $P_{\mathrm{k}}\Delta^{+}_{\mathrm{k}}$ and sort them in ascending order. Δ^{+}_{k} is the traffic increased by discarding content k ($k \in \mathbb{C}^{\mathrm{t}}_{\mathrm{u}}$). P_{k} needs to be considered because, in our model described in Sect. 2.2, altruistic client u can request the content cached in his or her cache space; we can increase P_{comb} by attaching content already cached at client u with larger P_{k} to the combined files while discarding content with smaller P_{k}.

Step 3 : Optimization of Eq. (3)
Under the supposition that there is no empty space in every client's cache capacity, we can simplify the optimization of Eq. (3) to the following discrete optimization problem as a function of x, which represents the number of content files included in $\mathbb{B}_{\mathrm{u}}$:

$$\max_{x} G_{\mathrm{comb}} P_{\mathrm{comb}} \tag{5}$$

$$\begin{aligned} \text{s.t.} \quad G_{\mathrm{comb}} &= \sum_{g=1}^{x} \left(E_{\mathrm{b_g}} - T\Delta^{+}_{\mathrm{d_g}} \right) + B_{\mathrm{ub_1}} \\ P_{\mathrm{comb}} &= \sum_{g=1}^{x} P_{\mathrm{b_g}} + \sum_{h=C-x}^{C} P_{\mathrm{d_h}}, \end{aligned}$$

where G_{comb} represents the amount of traffic reduced by combined content files; b_{g} is the identification number of the content with the g-th largest $P_{\mathrm{j}}E_{\mathrm{j}}$; d_{h} is the identification number of a content with the h-th smallest $P_{\mathrm{k}}\Delta^{+}_{\mathrm{k}}$; C is the cache capacity of altruistic client u. We determine the combination of $\mathbb{B}_{\mathrm{u}}$ and $\mathbb{D}_{\mathrm{u}}$ on the basis of Eq. (5) as:

$$\begin{cases} \mathbb{B}_{\mathrm{u}} = \mathbb{D}_{\mathrm{u}} = \phi \quad (if \;\; G_{\mathrm{comb}} P_{\mathrm{comb}} < 0) \\ \mathbb{B}_{\mathrm{u}} = (b_1, b_2, \cdots b_{\mathrm{x}}), \mathbb{D}_{\mathrm{u}} = (d_1, d_2, \cdots d_{\mathrm{x}}) \quad (else). \end{cases} \tag{6}$$

Since b_{g} and d_{h} are sorted, $G_{\mathrm{comb}}P_{\mathrm{comb}}$ becomes a convex function. Therefore, we can easily solve the discrete optimization problem and obtain the optimal number of content files for combination.

You see the detail of our selection algorithm in [8].

2.5. Numerical analysis

We previously observed how much traffic could be reduced by a set of combined content files based on Section. 2.4. In our evaluation, the next question was how long period T was. To determine this, we introduced the expected traffic reduction as a metric. Suppose that a set of combined content files is chosen for client u at t_0 in accordance with Eq. (3). We can calculate the expected traffic reduction by the set of content files at t_0, $E(t_0)$ given by:

$$E(\mathrm{t}) = \sum_{v \in \mathbb{N}} \sum_{k \in \mathbb{B}_{\mathrm{u}} \cap \mathbb{C}^{\mathrm{t}}_{\mathrm{u}}} \frac{P_k \cdot \epsilon^{+}_{\mathrm{vk}}}{N} \tag{7}$$

$$\text{s.t.} \quad \epsilon^{+}_{\mathrm{vk}} = B'_{\mathrm{vk}} - B_{\mathrm{vk}},$$

where $\epsilon^{+}_{\mathrm{vk}}$ and B'_{vk} mean how much generated traffic is increased and how much traffic is generated when client v requests and obtains content k at time $t+1$ if altruistic client u discards content k at time t, respectively. Then, cached files are replaced every time clients request content. After a certain time, Δt, the expected traffic reduction by the set of content files chosen at t_0 is $E(t_0 + \Delta t)$. We define approximated period T as a period during which $E(t_0 + \Delta t)/E(t_0)$ is maintained larger than a threshold α. During period Δt, the sum of requests by all clients is Δt.

Figure 5 plots the gain obtained by a set of combined content files during approximated period T. The gain means how much traffic is reduced on average in our scheme when a set of combined content files is stochastically downloaded by an altruistic client compared with the peer-assisted CDN without our scheme. The main parameters are listed in Table 1 that we used in our previous work [8]. The dashed line in Fig. 5 plots the upper-bound gain when no cache replacement is done. Furthermore, we have discussed the

Table 1. Variable analysis parameters.

No. of content files	1000
No. of local domains (n_1)	50
No. of clients in each local domain (n_2)	40
Total no. of clients (N)	2000
Ratio of altruistic clients	10%
Cache capacity at each client (C)	50
Traffic weight B_3,B_2,B_1,B_0	2000,50,1,0

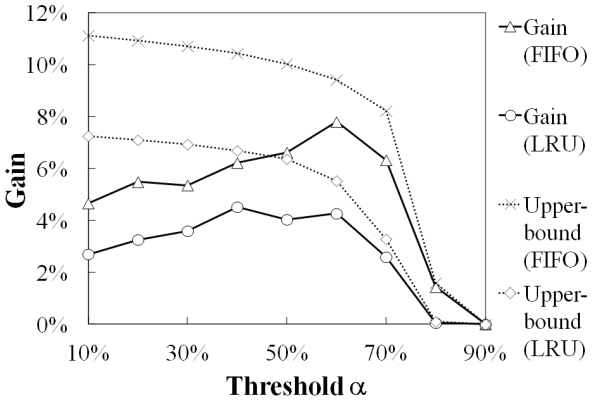


Figure 5. Theoretic gain by a set of combined content files during approximated period T.

actual gain obtained by considering the cache replacement algorithm, which is plotted by the solid line. The gain with considering of the cache replacement algorithm had peaks. As we discussed in Sect. 2.4.1, Eqs. (3) to (6) were derived on the assumption that cached content files in the network did not change during approximated period T. Therefore, the upper-bound is simply larger as α is larger. In reality, our scheme cannot reduce the amount of network traffic as much as the upper-bound because every time a client requests content, the cache can be replaced. Therefore, the actual gain becomes away from the upper-bound as α becomes larger.

3. PROPOSED PERIODIC DISTRIBUTION MECHANISM

3.1. Mechanism design

Here, we propose a new periodic distribution mechanism that uses our content combination algorithm described in Sect. 2.4, which was designed to improve traffic localization.

Figure 6 is a flowchart of our mechanism, where n and $t_{\rm n}$ are the counter for identifying a retrieved set of combined content files and the time when the set was retrieved, respectively. n is initialized as 0. A client makes a content request at time t. Then, this mechanism calculates the expected traffic reduction defined in Eq. (7) and assesses whether $E(t)/E(t_{\rm n})$ is still larger than α. If $E(t)/E(t_{\rm n})$ is larger, the service provider does not offer combined content files since we can expect the n-th retrieved combined-content to still be

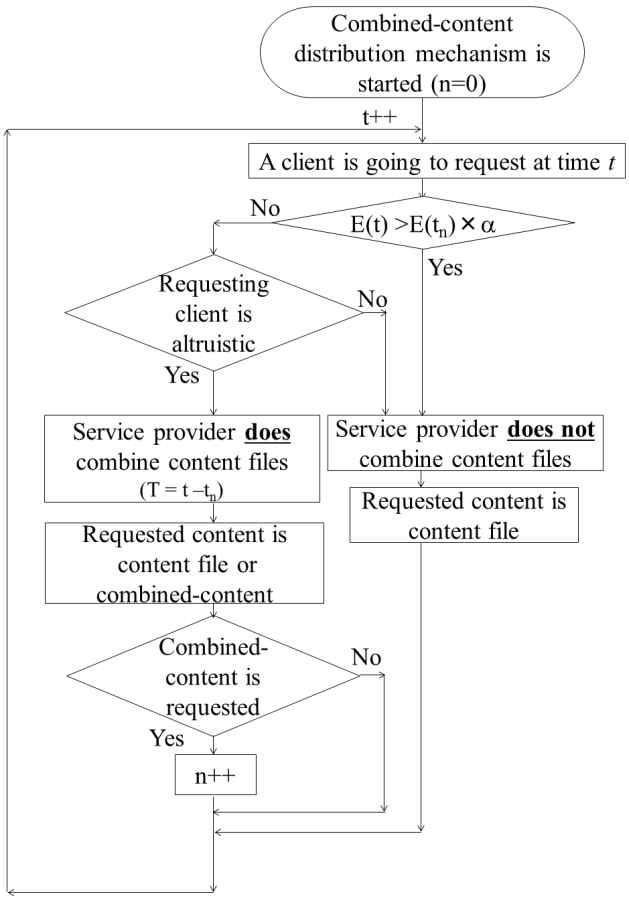


Figure 6. Flowchart for proposed distribution mechanism.

working to reduce traffic. If $E(t)/E(t_{\rm n})$ is smaller, the expected traffic reduction by the n-th downloaded combined-content is small because the cache state has been already changed from $t_{\rm n}$. Therefore, to further localize traffic, the service provider offers the next set of combined content files by considering the current state of the network cache when a requesting client is altruistic. T is given by:

$$T = t - t_{\rm n}, \tag{8}$$

where T means the period at which a new set of combined content files is offered.

If the offered combined-content is requested by the requesting client, n is incremented and $t_{\rm n}$ is updated. If the offered combined-content is not requested, T is incremented. The average period of offering combined content should be obtained from T by using an averaging function.

Why we designed our mechanism according to the flow in Fig. 6 is because: if the service provider sets T too small and greedily offers combined-content, a large amount of traffic is instantaneously generated when combined content files are retrieved by clients; however, if the service provider offers combined-content less frequently with longer T, they would not do anything even though the combined-content cached previously did not work anymore to localize traffic because the network cache state had been changed. Thus, unless

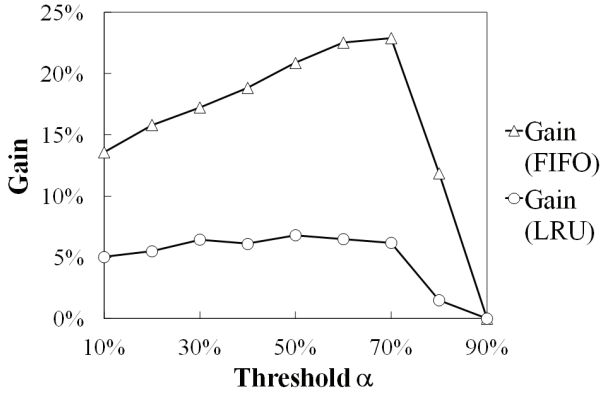


Figure 7. Gain by our distribution mechanism as a function of threshold α during $400,000$ requests.

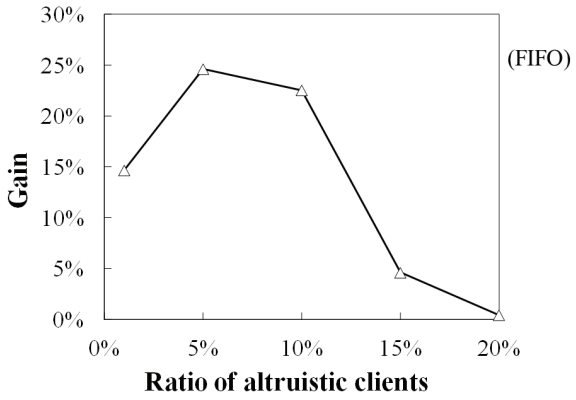


Figure 8. Gain as a function of ratio of altruistic clients at $\alpha = 60\%$ during $400,000$ requests. Cache replacement algorithm was FIFO.

the period is set appropriately, we cannot effectively localize the network traffic. We have to optimize the period where the service provider offers combined-content. Our proposed mechanism automatically optimizes the distribution period by using how long we can expect the previous downloaded combined-content to localize traffic.

3.2. Simulation evaluation

Here, we discuss how our new periodic mechanism described in Sect. 3 increases the gain in traffic localization. We used the parameters listed in Table 1 for our computer simulation. We observed performance during $400,000$ requests. We use the evaluation metric defined as:

$$\Psi = \frac{\tau^{\overline{C}} - \tau^{C}}{\tau^{\overline{C}}}, \tag{9}$$

where τ^{C} and $\tau^{\overline{C}}$ are how much traffic is generated in the entire network during $400,000$ requests when we use our mechanism and peer-assisted CDN without content combination, respectively. T was initialized to a sufficiently large value to combine content files even though we confirmed that performance was independent of the value of the initial T.

3.2.1. vs. threshold α

Figure 7 plots the gain defined in Eq. (9) as a function of threshold α. Let us first compare how much traffic our mechanism proposed in Sect. 3 generated with the previous numerical results presented in Sect. 2.5 where we only considered a single set of combined content files. Comparing Figs. 5 and 7, we can see that our mechanism can double performance; the peak with our mechanism was 22.9%, while the peak with the numerical results was 7.8% when we used first-in/first-out (FIFO) as the cache replacement algorithm. This reveals that, as stated in Sect. 3, we can obtain further traffic localization by distributing combined content files to multiple altruistic clients as long as the distribution period is set appropriately.

Next, we will discuss what we obtained by varying threshold α when we used FIFO. We can see that the gain had a peak at $\alpha = 70\%$. When α exceeded 80%, the gain became smaller. This is because our mechanism frequently combined content files since the cache state in the entire network was considered to be significantly changed even if cached files were just partly replaced. However, when α was less than 70%, the gain simply decreased since our mechanism did not combine content files while the expected traffic reduction of previous downloaded combined-content largely decreased.

Finally, let us discuss the difference between FIFO and the least-recently used (LRU) algorithm. In Fig. 7, basically, the gain becomes smaller when we use LRU because it performs so that popular content files are cached more likely even if our scheme is not used. However, in reality, new content files are published on a daily basis, which means that hardly any popular content files have already been cached locally. Let us consider an extreme case where many content files are published every day. In such cases, LRU would work similarly to FIFO.

3.2.2. Results in other scenarios

Figure 8 plots the dependence of our mechanism on the ratio of altruistic clients. Our mechanism successfully reduced a large amount of traffic when there was a small ratio of altruistic clients. This is because when there is a large ratio of altruistic clients, traffic has already been localized without using our mechanism since many content files are locally cached by altruistic clients.

Figure 9 plots the dependence of our scheme on the total number of content files. We can see here that our mechanism reduced traffic by 27.7% when there were 1200 content files. When there were a small number of total content files, most requested content files had already been cached locally even without using our mechanism. Furthermore, when there were a large number of total content files, it was less likely altruistic clients would request combined content files because they had lots of other options when they chose content. This is why we can see a peak in Fig. 9.

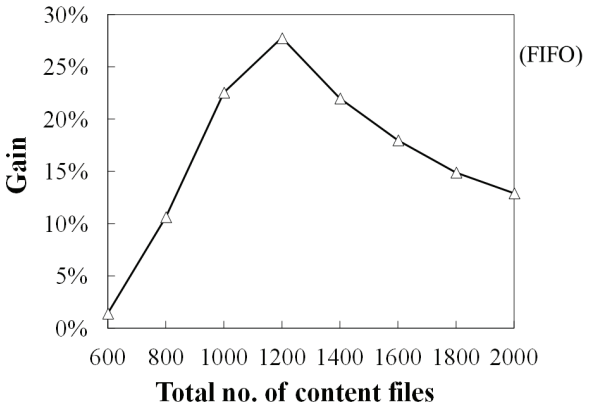


Figure 9. Gain as a function of total number of content files at $\alpha = 60\%$ during $400,000$ requests. Cache replacement algorithm was FIFO.

4. CONCLUSION

We proposed a periodic combined-content distribution mechanism to improve traffic localization by inducing multiple altruistic clients to cache combined-content in peer-assisted CDN models. Our mechanism suggested how to set period for a combined content distribution appropriately because it is not always better to distribute combined content files to more altruistic clients. Our computer simulations confirmed that our mechanism doubled performance compared with the numerical results we previously presented.

We should mention a couple of remaining issues. We simplified the content request probability in Eq. (2). We could make it more realistic using a utility function like that introduced by [24]. We could also make our network model more realistic in terms of diversity and scalability.

REFERENCES

[1] "Hulu," http://www.hulu.com/.

[2] "iTunes," http://www.apple.com/itunes.

[3] T. Mori, N. Kamiyama, S. Harada, H. Hasegawa and R. Kawahara, "Improving deployability of peer-assisted CDN platform with incentive," in Proc. of IEEE GLOBECOM'09, pp.1–7, Honolulu, Nov. 2009.

[4] D. Xu, S.Kulkarni, C. Rosenberg and H. Chai, "Analysis of a CDN-P2P hybrid architecture for cost-effective streaming media distribution," Mutimedia Systems, vol.11, no.4, pp.383–399, Mar. 2006.

[5] C. Huang, A. Wang, J. Li and K. W. Ross, "Understanding hybrid CDN-P2P: why Limelight needs its own red swoosh," in Proc. of the 18th International Workshop on Network and Operating Systems Support for Digital Audio and Video (NOSSDAV'08), pp.75 – 80, Braunschweig, Germany, May 2008.

[6] G.Peng, "CDN: Content distribution network," Dept. Computer Science, State Univ. of New York, New York, Tech. Rep. TR-125, 2003.

[7] G. Pallis and A. Vakali, "Insight and perspectives for content delivery networks," Commum. ACM, vol.49, no.1, pp.101–106, 2006.

[8] N. Maki, T. Nishio, R. Shinkuma, T. Mori, N. Kamiyama, R. Kawahara and T. Takahashi, "Traffic engineering of peer-assisted content delivery network with content-oriented incentive mechanism," IEICE Trans. on Inf. & Syst., vol.E95-D, no.12, Dec. 2012.

[9] E.Cohen and S. Shenker, "Replication strategies in unstructured peer-to-peer networks," in Proc. of ACM SIGCOMM'02, Pittsburgh, Aug. 2002.

[10] K. Sripanidkulchai, B. Maggs and H. Zhang, "Efficient content location using interest-based locality in peer-to-peer system," in Proc. of IEEE INFOCOM'03 vol.3 pp.2166-2176, Pittsburgh, April 2003.

[11] J. Kangasharju, K. W. Ross and D. A. Turner, "Optimizing file availability in peer-to-peer content distribution," in Proc. of IEEE INFOCOM'07, pp.1973-1981, Anchorage, May 2007.

[12] X. Tang and S. T. Chanson, "Coordinated en-route web caching," IEEE Transactions on Computers, vol. 51, no. 6, pp. 595–607, June 2002.

[13] A. Nakaniwa, H. Ebara and H. Okada, "File allocation designs for distributed multimedia information networks," in Proc. of IEEE GLOBECOM '98, vol.2, pp.740–745, Sydney, Nov. 1998.

[14] F. L. Presti, N. Bartolini, and C. Petrioli, "Dynamic replica placement and user request redirection in content delivery networks," in Proc. of IEEE ICC, vol.3, pp. 1495–1501, Seoul, May. 2005.

[15] N. Kamiyama, R. Kawahara, T.Mori, and H. Hasegawa, "Multicast pre-distribution in VoD service," in Proc. of IEEE CQR, Naples, May 2011.

[16] E. Adar and B. Huberman, "Free riding on gnutella," First Monday, vol.5, no.10, Oct. 2000.

[17] D. Hughes, G. Coulson and J. Walkerdine, "Freeriding on gnutella revisited: the bell tolls?" in IEEE DS Online, June 2005.

[18] M. Yamada, K. Sato, R. Shinkuma and T. Takahashi, "Incentive service differentiation for p2p content sharing by wireless users," IEICE Trans. Commum., vol.90-B, no.12, pp.3561–3571, 2007.

[19] R. Cheng and J. Vassileva, "User motivation and persuasion strategy for peer-to-peer communities," in Proc. of the 38th Annual Hawaii International Conference on System Science (HICSS'05), pp.193–202, Hawaii, Jan.2005.

[20] K. Sato, R. Hashimoto, M. Yoshino, R. Shinkuma and T. Takahashi, "Incentive mechanism for P2P content sharing over heterogeneous access networks," IEICE Trans. Commum., vol.91-B, no.12, pp.3821–3830, 2008.

[21] C. Wang, H. Wang, Y. Lin and S. Chen, "A currency-based p2p incentive mechanism friendly with ISP," in Proc. of the Computer Design and Applications (ICCDA), vol.5, pp.403–407, Qinhuangdao, Jun. 2010.

[22] R. Cuevas, M. Kryczka, A.Cuevas, S. Kaune, A. Guerrero and R. Rejaie, "Is content publishing in Bittorrent altruistic or prifit-driven?," in ACM CoNEXT, 2010.

[23] L. Breslau, P. Cao, L. Fan, G. Phillips and S. Shenker, "Web caching and zipf-like distributions: Evidence and implications," in Proc. of INFOCOM'99, vol.1, pp.126–134, New York, Mar. 2002.

[24] W. Hanson and R.K.Martin, "Optimal bundle pricing," Management Sci. vol.36, pp.155-174, 1990.

MEDICATION ERROR PROTECTION SYSTEM WITH A BODY AREA COMMUNICATION TAG

Yoshitoshi Murata†, Shuji Ikuta††, Nobuyoshi Sato†, Tsuyoshi Takayama†

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

†† NTT ComTechnology Corporation
Sarugakucho, Chiyoda-ku, Tokyo, 101-8347 Japan

ABSTRACT

Errors in administering medication are serious problems. Most errors are due to some confusion between the patient and the medication. The bar-code has been used to deal with this problem. However, since a nurse has to put a reader over a bar-code tag, there is still room for improvement when, say, a nurse is affected by stress due to too heavy a workload. As a safer alternative to bar-codes and RFIDs, we propose Touch-tag, a body area communication tag. The concept is that the patient wears a tag and the nurse has a Touch-tag reader that reads the ID on the tag by a nurse just by touching the patient. Since a nurse usually touches a patient during administration of medication, the medical error protection system using the Touch tag does not involve any additional work. We describe the medical error protection system with the Touch-tag and experiments to confirm whether the Touch tag works well or not.

***Keywords*—** bar-code, RFID, Touch tag, medication management, medication error, body area communication network

1. INTRODUCTION

Medical errors occur frequently and are a very significant problem. Patient safety is a fundamental principle of health care. But "to Err is human" [1], and preventable adverse events are a leading cause of death in the United States. When extrapolated to the over 33.6 million admissions to U.S. hospitals in 1997, the results of studies imply that at least 44,000 and perhaps as many as 98,000 Americans die in hospitals each year as a result of medical errors. The Institute of Medicine reported that about 400,000 preventable medication-related injuries occurred in US hospitals in 2006 [2].

Every point in the process of care-giving contains a certain degree of inherent unsafeness. The most extensive study of adverse events is the Harvard Medical Practice Study [3]. According to this study, 58% of adverse events were "preventable adverse events" and 27.6% were due to negligence. Some of them are work flow errors such as supplying the wrong patient with the wrong drugs. Vincent et al. presents a framework that aims to encompass the many factors that cause adverse events in clinical practice [4]. One of the seven factors that influence the clinical practice is the work environment. Staffing levels, skills mix, workload, shift patterns, etc., are included in this factor. Fu-In Tang et al. report that the personal neglect, heavy workload, and new inexperienced staff are three main factors contributing to medication errors [5].

A comprehensive mediation management system using bar-codes or RFIDs and wireless communication systems is very useful for mitigating adverse events due to negligence in hospitals [6][7]. In particular, there are two types of RFID: active-tag and passive-tag. There are several critical issues such as electrical interference, impacting RFID usage in a hospital [8], and these have meant that only the passive sort of tag can be used safely. In this case, nurses or other caregivers have to put a reader over an RFID tag when they want to read it.

In case of the bar-code system as well, nurses have to put a reader over a tag similar to the RFID. This additional action probably is not problem at the desk. However, it increases a nurse's workload, and it must be a source of stress for a nurse at the bedside. 34% of medication errors occur in the administration of medication [9]. Moreover, one of major factors contributing to medication errors is heavy workload. These two factors mean that devices which increase workload are not suitable for use at the bedside.

To solve this problem, we developed a prototype medication error protection system using the Touch-tag [10] instead of the bar-code or the passive RFID. This tag was developed by our partner, Ad-Sol Nissin Corp. Since it uses a body area communication system [11], the tag-reader can simply read data written in a tag by a nurse who has a tag-reader touching a patient who has a tag. Hence, there is no additional work or stress involved in using it.

After explaining the related work, we describe the configuration of our medical error protection system in Section 3, our prototype system in Section 4, and our experimental evaluation in Section 5. Future work is

*Touch tags and a touch tag reader were supplied from Ad-Sol Nissin Corp.

mentioned in Section 6, and the key points are summarized in Section 7.

2. RELATED WORK

2.1. Medication management system

Medication management involves the following four stages;

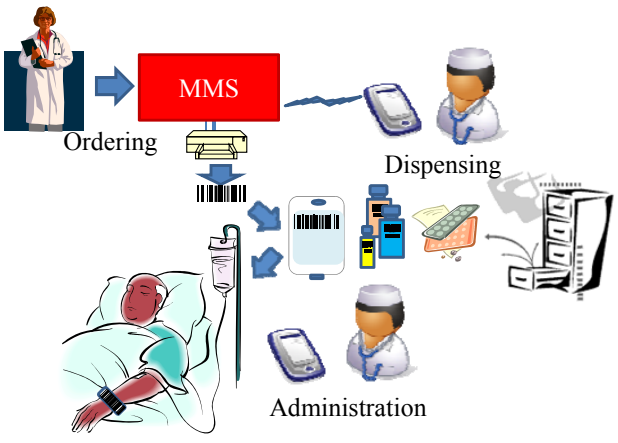
- Ordering
- Transcribing
- Dispensing
- Administration

D. W. Bates indicates that among preventable events, the primary error tends to occur in the ordering stage 56% of the time, but only 6% of the time in transcribing stage, and 4% in the dispensing stage. Moreover, 34% of the time it occurs during medication administration [9].

Charles Vincent proposed a framework of risk factors in clinical medicine that allows us to take a systematic approach to safety and error reduction [4]. This framework is useful for guiding investigations into incidents, for generating ways of assessing risk, and for focusing research on the causes and prevention of adverse outcomes. However, it cannot be used to intercept preventable adverse events.

Five or more "rights" of medication administration such as right patient, right medication, right dose, right route, and right time, have been identified and put in practice in many hospitals [12]. This approach is very useful for reducing preventable medication errors. But, as they depend on the humans involved, it is difficult to guarantee them perfectly. Preventable errors are incidents in unusual cases such as a big accident. We believe that combining a human approach with a systematic approach would be a more effective way to intercept preventable medication errors before they occur.

Eric Paul is promoting a point care medication management system using bar-code and wireless technology [6]. The wireless bar-code terminal is used during the dispensing and administration of medication to identify a patient and a medication, as shown in Fig. 1. A physician inputs a medicine order to the MMS. The order is clinically screened by a pharmacist. A nurse or a pharmacist picks up the correct medications from a smart drawer according to a medication list by reading a bar-code. After dispensing but before administrating a medication, the nurse identifies the patient by reading the bar-code on his or her bracelet.



MMS: Medication Management System

Figure 1. System configuration of the medication management system with the bar-code

A nurse has to carry a handy terminal that includes a bar-code reader function and put it over the patient's bracelet. This action increases the nurse's workload and must be stressful for a nurse at the bedside. Remember that 34% of medication errors occur during administration, and one of major factors that contribute to medication errors is the hard workload. These issues mean that a bar-code is not the most suitable for use at the bedside.

To deal with this problem, we propose using a Touch-tag that uses a body area communication system instead of a bar-code.

2.2. Body area communication systems

There are two types of body area communication system [13]. One is that a body is used as a data-bus between devices on a body. The other is that a radio system is used as a communication system between devices on a body. The first type of system was invented by T. G. Zimmerman in 1996 [11]. His body area communication system used the variable of electric-field on a body of which the frequency range was from 0.1 to 1 MHz. He built a prototype operating at 330 kHz. However, since he chose a frequency range from 0.1 to 1.0 MHz to suppress electric-field emissions, probably it was difficult to achieve stable communication because of the significant ambient noise. Ambient noise usually presents itself at frequencies below 1 MHz. When the transmission signal is strong enough to maintain stable communication, there is no difference between his system and existing near-field communication systems such as WiFi and Bluetooth. Ultimately, Zimmerman stopped developing this technology and chose the second scheme. These days, wireless body area networks are actively under development for applications to improve health care and the Quality of Life [14].

Panasonic Electric Works Co. Ltd. developed a touch communication system using the variable of the electric current in a body instead of the electric field to avoid the ambient noise in 2004 [15].

Yuichi Kado evolved Zimmerman's body area communication system [16]. He selected the frequency band from 5 to 10 MHz to avoid the ambient noise and developed a modulation scheme to efficiently modulate the electric field near the body and receiver circuits which reduce ambient noise such as electrical hum.

Our partner, Ad-Sol Nissin Corp., has introduced a semi-active scheme to improve the battery life cycle of tags used in the evolved electric field schemes [10]. It is possible to use a tag for a few years without replacement of a battery.

We chose Ad-Sol Nissin's Touch tag for the following reasons:

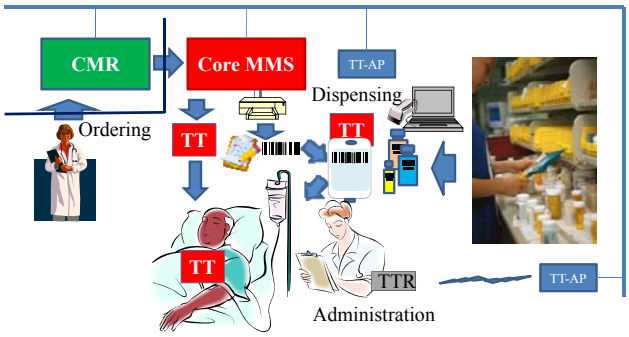
- It is possible for the electric field scheme to communicate over clothes or shoes. On the other hand, the user has to touch the tag directly in the electric current scheme.
- The life cycle of the tag is very long.

3. PROPOSED SYSTEM

We propose a medication management system (MMS) combining the bar-code and the Touch tag ideas (see Fig. 2). An intravenous drip injection is a typical inpatient treatment in Japan. We designed the MMS for intravenous drip injections. The MMS comprises the following units:

- Core MMS: This unit is for taking and storing ordering lists of medication and patient's information from a computerized medical records system (CMR), transcribing from an ordering list to actual medications, printing a bar-code to be attached to an intravenous bag, and managing patient, intravenous bag, injected medications and nurse using both Touch tags and bar-codes.
- PC with a bar-code reader: This unit is for reading a bar-code on liquid medications and a printed bar-code attached to an intravenous bag.
- Touch tag: This unit is attached to an inpatient and an intravenous bag. The relationship between the Touch tag's ID and inpatient or intravenous bag is managed by the core MMS.
- Touch tag access point (TTAP): This unit is for connecting a Touch tag reader to the core MMS; it is a ZigBee or Bluetooth or WiFi router.
- A list of five or more "rights" of medication administration.

The order data and information for each inpatient are stored and managed in the CMR. After a physician puts an order for medications for an inpatient into the CMR, the MMS takes the patient's information and list of orders, checks for problems that may be between the patient and the ordered medications by using the knowledge system, transcribes



CMR: Computerized Medical Records System
MMS: Medication Management
TT: Touch-tag
TTR: Touch-tag Reader
TT-AP: Touch-tag Access Point

Figure 2. System configuration of the proposed system

and presents the order on the display on the PC. Since screening by a pharmacist is probably very useful to eliminate wrong medications, this error protection scheme must be used together with such a knowledge system.

Since a bar-code such as EAC, UPC or JAN is attached to a package or a bottle of medications, the bar-code system is a very efficient way of comparing medications with an order list and linking them with an intravenous bag. Hence, we plan to use the Touch tag for making links between the patient, nurse, and intravenous bag during administration of the medication.

When a nurse injects liquid medication into the intravenous bag according to the order list, she or he attaches a Touch tag and a bar-code label printed out from the core MMS, and reads the Touch tag and the bar-code to link the Touch tag to the intravenous bag.

During administration of medication, the nurse carries a Touch tag reader and a checklist for five or more "rights" of medication administration. It is possible to combine a Touch tag reader and a checklist to a smart terminal. The nurse can verify the right patient, right medication, and right time simply by touching the patient and intravenous bag.

We developed a prototype system for checking the intravenous bag and patient (instead of using the MMS) and confirmed whether the Touch tag can be used to verify the right patient, right medication, and right time.

4. PROTOTYPE SYSTEM

The Touch tag has mainly been used for security doors. In this case, the user puts a Touch tag around his or her neck and the Touch tag reader is set on a door or on the floor in front of a door. There is not a use case in which someone carries a Touch tag reader. In the case of a person carrying a Touch tag reader, he or she may touch someone else who has a Touch tag. We have to develop a reading scheme to reliably read a Touch tag ID that the right person has. In addition, the start and end times of drip medication have to be managed. We need to establish a scheme to distinguish its start and end timings.

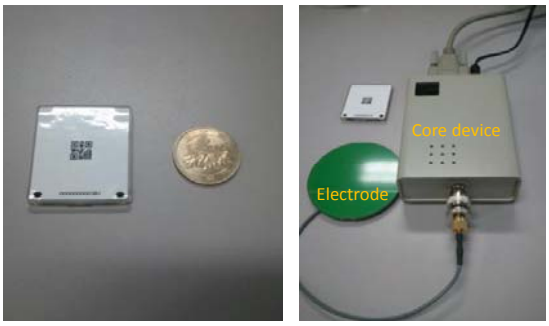
We developed a prototype system to examine the following issues:

- Places to put a Touch tag reader on a nurse.
- Places to put a Touch tag on a patient.
- Places to put a Touch tag on an intravenous bag.
- The algorithm to read only the right Touch tag.
- Scheme to distinguish the start and end times of the drip medication.

The Touch tag and the Touch tag reader are shown in Fig. 3. The Touch tag reader is a model used for security doors. We implemented the following functions on a PC instead of the core MMC:

- Verifying a relationship between a Touch tag's ID for the patient and the Touch tag's ID for the intravenous bag.
- Setting the reading rate of the Touch tag reader.

- Reading and storing the Touch tag's ID and its received signal level.



(a) Touch tag (b) Touch tag reader

Figure 3. Touch tag and Touch tag reader

5. EXPERIMENTS

5.1. Places to put Touch tag reader on a nurse and Touch tag on a patient

The Touch tag was attached to several points on the subject, as shown in Fig. 4. The Touch tag reader was set on the waist, right lower arm, and right upper arm of members of our research groups instead of a nurse (see Fig. 5). The number of subjects was two in this case, and we measured whether it was possible to read the ID of the Touch tag five times for each position.

When the Touch tag was on the lower or upper arm, the Touch tag reader could always reads the ID. We think the upper arm is more suitable than the lower arm because it it's a better place to avoid an unexpected touch.

5.2. Places to put Touch tag on intravenous bag

We measured five times whether it was possible to read the ID on the Touch tag on a pet bottle instead of an intravenous bag. We measured the bottle filled with water and the bottle empty. When the bottle was filled with water, the ID was readable regardless of where the tag was on the bottle. But when a bottle was empty, it was impossible to reliably read the ID.

5.3. Algorithm to read only the "right" Touch tag

In case of unexpected touches, the duration of the touch must be short. We devised an algorithm that can read an ID on a Touch tag a prescribed number of times within a prescribed duration. In this case, we have to consider the valance between reliability and usability. Verifying many times improves reliability but it may take too long.

We did experiments varying the number of verifications and duration. As a result, we decided that the prescribed times should be 2, and the prescribed duration should be 10 seconds.

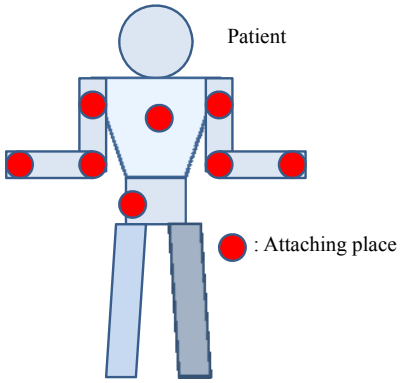
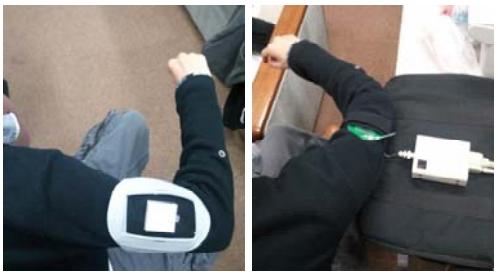


Figure 4. Places to attach a Touch tag



(a) Touch tag (b) Touch tag reader

Figure 5. Touch tag and Touch tag reader attached to an arm

5.4. Scheme to distinguish the start and end times of drip medications

We noticed that the received signal level when the Touch tag was touched directly was different from the level when the tag was touched through an intravenous bag. We decided to use this difference to distinguish the start and end times of a drip. We did an experiment to decide the threshold level. In this paper, touching a Touch tag directly is called "direct touch", and touching a Touch tag through an intravenous bag is called "indirect touch."

[Experiment]

We evaluated the following direct touch and indirect touch situations:

- Direct touch: The subject holds a bag with his or her arm without a tag reader on it and touches the tag directly with his or her arm with the tag reader on it (see Fig. 6).
- Indirect touch: The subject holds a bag using his or her arm without a Touch tag reader on it and touches some portion of a bag except the tag itself by using the arm on which the tag reader is set (see Fig. 7).

A polyethylene bag filled with water was used instead of an intravenous bag. We used different polyethylene bags to simulate the start and end conditions:

- A polyethylene bag filled with water
- A polyethylene bag emptied of water.

We measured the received signal level in the following four cases:

- Case 1: Directly touching the filled bag.
- Case 2: Indirectly touching the filled bag.
- Case 3: Directly touching the empty bag.
- Case 4: Indirectly touching the empty bag.

Three subjects participated in the experiment. Each subject touched a bag 30 times in each case. Hence, the number of measurements was 90 in each case.

As the result, the average signal levels in each case were: Lc1=431, Lc2=298, Lc3=389 and Lc4=280. The difference between Lc1 and Lc2 is about 1.4 times, the difference between Lc3 and Lc4 is also about 1.4 times. This means that the difference between the direct touch and the indirect touch is roughly constant under the same conditions of the bag. Since the difference between Lc1 and Lc4 is about 1.6 times, we assigned case 1 to the start of intravenous drip injection and case 4 to the end. The frequency distributions corresponding to the received signal level in case 1 and case 4 are shown in Fig. 8. It is difficult to distinguish the direct touch from the indirect touch at received signal levels of 301 to 400. Therefore, we decided two threshold levels; one for distinguishing the direct touch from the indirect touch, the other for distinguishing the indirect touch from the direct touch.

We did an experiment to confirm whether this distinguishing function worked well or not. Four polyethylene bags filled with water were put on a desk. A Touch tag was attached to each polyethylene bag. A subject chose a polyethylene bag and did the direct touch procedure. After verifying its Touch tag's ID, the subject released the bag. A subject did the same for the other three bags. The whole trial was repeated five times by each subject. The number of subjects was four.

The results are shown in Table I. In total, four mis-verifications occurred. The reason is that a subject took a long time to pick up a bag, and the verification was done without touching the tag directly.

Table 1. Number of mis-verifications in direct touch case

# of Set	Subject 1	Subject 2	Subject 3	Subject 4
1st	0	0	0	1
2nd	2	0	0	0
3rd	0	0	0	0
4th	0	1	0	0
5th	0	0	0	0
Total	2	1	0	1

Figure 6. Direct touch

Figure 7. Indirect touch

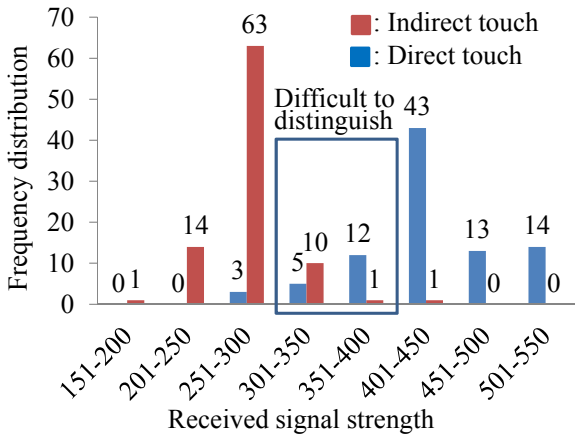


Figure 8. Frequency distribution of received signal levels

5.5. Practicability test

In some hospitals, medications including intravenous bags are sorted to each inpatient. We did an experiment simulating this sort of work to confirm whether our error protection system would work well or not in this circumstance. Four subjects repeated the following sorting operation five times.

[Setup]

We prepared two trays in addition to the four polyethylene bags. Polyethylene bags were put on these trays. A "right" tag was attached to three of four polyethylene bags and one tray; a wrong tag was attached on the other bag and tray as shown in Fig. 9.

[Operation]

The subject repeated the following procedure five times:

(1) Touch the tag on a tray to determine if it is the right one.

(2) Pick up the bag to determine if it is the right one.

(3) Put the right bag in the right tray.

(4) Repeat the above operations for every bag.

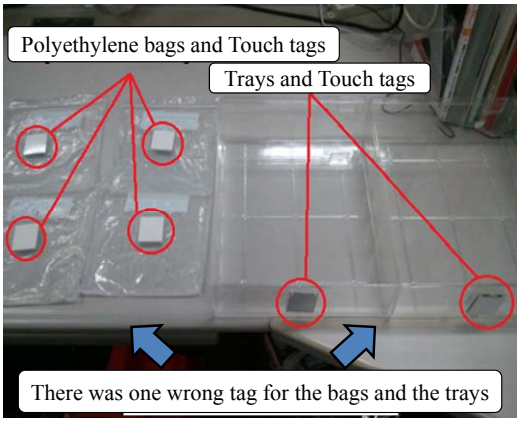


Figure 9. Tools for the practicability test

Table 2. Results of practicability test

# of set	Subject A		Subject B	
	Time	Number	Time	Number
1	18	4	18	5
2	18	4	27	6
3	31	6	22	5
4	15	5	20	4
5	31	6	28	5
Average	22.6	5	23	5
# of set	Subject C		Subject D	
	Time	Number	Time	Number
1	44	6	27	5
2	38	5	26	5
3	28	5	29	5
4	27	5	23	4
5	21	4	22	5
Average	31.6	5	25.4	4.8

[Results]

We measured the operation time and number of verifications for each set. The results are shown in Table II. The average time for one procedure was about 25 seconds. It took about 5.2 seconds to pick up and verify one bag. About 50% of this time was used for the verification; the rest of the time was used to pick up a bag and put it on a tray.

The ID was misread sometimes. As a result, it took a long time for the subject to complete a set. We recommend verifying two times to avoid the misreading problem.

6. FUTURE WORK

In this paper, we used a Touch tag reader intended for a security door. This device is too big for a nurse to carry, as can be seen in Fig. 5. It must be connected to a PC directly, so it is impossible to use an MMS. Our partner, Ad-Sol Nissin Corp., has already developed a smaller wireless Touch tag reader (Fig. 10). This prototype device comprises a Touch tag reader unit and a ZigBee radio unit. The electrode is set on back of the reader case. It can connect a server PC through the ZigBee router and intranet. Therefore, it is very easy for a nurse to carry and connect it to the core MMC. And, since its reading cycle of a Touch tag becomes short, the prescribed duration of the touch becomes a few seconds.

We plan to incorporate the findings of this study in this device and develop an MMC as shown in Fig. 2 in cooperation with a hospital.

Fig**ure 10.** Wireless Touch tag reader

7. CONCLUSION

Our medication error protection system combines the bar-code and Touch tag concepts. We developed a prototype system and performed an experiment demonstrating that the Touch tag was useful for verifying the relationship between an intravenous bag, a patient, and a nurse. We also devised a reading scheme to avoid unexpected touches and the scheme to distinguish the start and end times of drip medications.

Our system will not increase the workload or stress of a nurse and we believe it will help eliminate most of the preventable medication errors.

REFERENCES

[1] L. T. Kohn, J. M. Corrigan, M. S. Donaldson, “To Err is Human: Building a Safer Health System,” National Academy Press, WashingtonDC, 2000.

[2] P. Aspden, J. A. Wolcott, J. L. Bootman, and L. R. Cronenwett, Committee on Identifying and Preventing Medication Errors., Preventing Medication Errors, National Academies Press, Washington, DC, 2007.

[3] Leape, Lucian L.; Brennan, Troyen A.; Laird, Nan M., et al., “The nature of adverse events in hospitalized patients,” Results of the Harvard Medical Practice Study II. N Engl J Med. 324(6):377-384, 1991.

[4] Charles Vincent, Sally Taylor-Adams, Nicola Stanhope, “Framework for analyzing risk and safety in clinical medicine,” BMJ. 316, pp. 1154-1157, 1998.

[5] Fu-In Tang et al., “Nurses relate the contributing factors involved in medication errors,” Journal of Clinical Nursing, Vol. 16, Issues 3, pp.447-457, 2007.

[6] E. Paul, “Reengineering medication management from the bedside using bar-coding and wireless technology,” HIMSS 2000, Volume 1, Session 12, pp61-69, 2000.

[7] F. Wu, F. Kuo and L. W. Liu, “The Application of RFID on Drug safety of Inpatient Nursing Healthcare,” ACM, ICEC2005, pp.85-92, 2005.

[8] C. H. Kuo, H. G. Chen, "The Critical Issues about Deploying RFID in Healthcare Industry by service perspective," IEEE, 41st Hawaii International Conference on System Science 2008, 2008.
[9] David W. Bates et al., "Incidence of Adverse Drug Events and Potential Adverse Drug Events," JAMA, July 5, 1995 Vol. 274, No. 1, 1995.
[10] Touch Tag, http://www.adniss.jp/en/archives/59, August, 2012
[11] T. G. Zimmerman, "Personal Area Networks: Near-field intrabody communication," IBM Systems Journal, Vol.35, Nos. 3&4, pp.609-617, 1996
[12] 8-rights-of-medication-administration, http://www.nursingcenter.com/Blog/post/2011/05/27/8-rights-of-medication-administration.aspx
[13] Intrabody communication, Nikkei Electronics, June, 30th, 2008. (in Japanese)
[14] Benoit Latre et al., "A survey on wireless body area networks," Springer Science+Business Media, Wireless Netw (2011) 17:1-18.
[15] Remarkable ubiquitous technologies - Body Area Communication – (in Japanese), http://itpro.nikkeibp.co.jp/prembk/NBY/techsquare/20050304/157032/
[16] Yuichi Kado, Mitsuru Shinagawa, "RedTacton Near-body Electric-field Communications Technology and Its Applications," NTT Technical Review, Vol. 8 No. 3 Mar. 2010.

INTRA-CITY DIGITAL DIVIDE MEASUREMENTS THROUGH CLUSTERING

Tuğra Sahiner[(1)], *Ayşegül Ozbakır*[(2)], *Güneş Karabulut Kurt*[(1)]

[(1)] İstanbul Technical University,
Department of Electronics and Communications,
34469 İstanbul, Turkey
{sahinert,gkurt}@itu.edu.tr

[(2)] Yildiz Technical University,
Department of City and Regional Planning
34349 Yildiz-Istanbul, Turkey
aozbakir@yildiz.edu.tr

ABSTRACT

With latest development of telecommunication technologies and end user's increased bandwidth and mobility demand reachability of information and communication technologies (ICT) became more critical. In this paper, we approach to end user behaviors and the reachability to ICT by tackling digital divide concept along with clustering analysis. To the best of our knowledge, this research is a unique case study that attempts to analyze digital divide at intra-city level by neighborhoods. While governments and institutions, such as ITU, are in question of whether the global divide is widening or narrowing, there are no studies, neither in the literature nor in practice to understand the gap between ICT users in a city. With this goal, Istanbul habitants were asked to fill a questionnaire, in order to be classified in terms of their technology reachability and reasons of using ICT. Then, clustering analysis was performed to questionnaire results. Respondents have been clustered into sub groups from digital divide perspective. The required steps and suitable clustering techniques during this process are discussed with determination of questions at the end which are commonly answered by respondents with different ICT knowledge, which may lead us determine precise reasons of digital gap later on.

Keywords **—** clustering, digital divide, characteristics of digital gap, network usage questionnaires.

1. INTRODUCTION

Digital divide, with respect to Organisation for Economic Co-operation and Development's (OECD) definition, is the term that refers to the gap between individuals, households, businesses and geographic areas at different socio-economic levels with regard both to their opportunities to access information and communication technologies (ICT) and to their use of the Internet for a wide variety of activities [1]. Along with this information, the main motivation in our work is to determine the digital gap between the habitants of city Istanbul and common questions that were answered by respondents that are scientifically separated into digital divide groups. This will be done through quantification of digital gap within neighborhoods and using clustering analyses during this process. We also need to note that ICT usage patterns among different city segments play a significant role when setting up and managing an efficient network from a network service provider's perspective. Whilst, understanding and measuring the digital divide types among city neighborhoods are relevant research questions to be answered.

In the literature, addressing the gaps between information society by providing global ICT developments through quantitative measurements has gained importance since the late 2000s. As a result of global calls for these developments, World Summit on the Information Society (WSIS) has come up with strategy documents that underline future needs in the measurement of worldwide digital gaps [2], [3].

As a reply to the request of composite and comparable statistical measurements, ITU has developed some indices. Among them are: Digital Access Index (DAI), the Digital Opportunity Index (DOI) and the ICT Opportunity Index (ICTOI). The final index of ITU, ICT Development Index (IDI), released in 2009 attempts to incorporate different aspects of the previous indices. However, still most of the recent research indicates that *one size does not fit all* due to the geographic, social, economical and cultural differences among countries. For this reason, when ranked according to the results of these indices, countries or regions might reveal misleading performance results [4].

The novelty of this research is to analyze local and intra city level digital divide categories. To our knowledge, there is no study in the literature that attempts to signify the digital gap categories by neighborhoods. To analyze the digital gap categories, clustering algorithms are applied to a set of questionnaire responses submitted to 1140 individuals with different socio-economic profile from different neighborhoods of Istanbul. The methodology chosen for the research is to apply different clustering algorithms in order to differentiate digital literates, immigrants and illiterates.

This work is supported by Yıldız Technical University, 2011-03-02-GEP02.

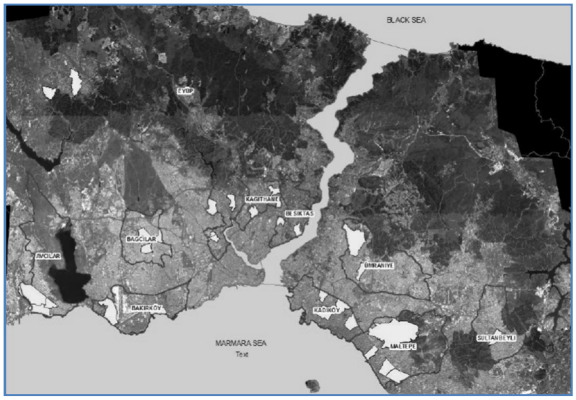


Figure 1. 31 Selected neighborhoods, labeled white

Data clustering analyses are empirical steps of classifying various subjects into different clusters with respect to the properties of the corresponding subjects. These techniques, referred to as unsupervised classification, aim to create groups of clusters, in such a way that objects in one cluster are very similar and objects in different clusters are quite distinct [5]. In our work, we also propose a cluster validation technique based on pre-determined user expectations, after determination of the number of distinct user patterns.

In this paper, a guideline for clustering questionnaires will be presented along with significant findings about literacy levels of cellular network users. In Section 2, details on questionnaire and preparation of retrieved dataset for clustering steps will be explained. In Section 3, detailed steps in the structure of clustering analyses and some improvements and new proposals on stability and validity of clusters will be explained. Finally, in Section 4 and 5, results are presented and conclusions are drawn, respectively.

2. QUESTIONNAIRE DETAILS AND DATA PRE-PROCESSING

Our questionnaire is filled by 1140 individuals from 10 districts with 31 neighborhoods (Figure 1) of Istanbul Metropolitan Area and they have been requested to answer 100 questions in total. The amount of participants of this questionnaire has been calculated statistically by random sampling method, to reflect an approximate result of all Istanbul regions (with a %95 reliability, ± 0.055 error rate). 10 districts (in a total of 39) were chosen, and 1140 respondents from 310 houses in different 31 neighborhoods (10 houses/neighborhood), in these 10 districts are included in this questionnaire. Table 1 and Table 2 show the ICT service types that are asked to the responders and main categories of questions.

In this paper, a general clustering approach will be proposed for a data set that contains a mixture of data types. Because of the respondents' randomly selective characteristic, and variety of data types; data preparation, clustering analyses and the selected methodology of each step may differ from one questionnaire type to another [6]. Hence, we need to prepare questionnaire responses for further clustering steps, accordingly.

Table 1. Ict service types that are asked to respondents

Internet	Fixed Line	Mobile Phone
3G	Home Line (PSTN)	3G
Wi-Fi hot spots	Mobile: 2G	2G
xDSL	Mobile: 3G	WiFi hot spots
Fiber	Mobile: VoIP application	
WiMAX		

Table 2. Main questionnaire categories for digital divide analysis

Main categories and answers for each of the households
Demographic structure:
- Gender, - Age, - Place of birth, - Mother tongue, - Literacy.
Economic Structure:
- Occupation status, - Monthly income.
ICT ownership and use:
- Number of mobile phones, - Network type, - Invoice type, - Computer usage and frequency (hour/day) - Internet usage and frequency (hour/day) - Place of internet accessibility - Mobile phone usage, - Mobile services, - Reasons of computer use, - Reasons of internet use, - Mobile phone applications.
ICT Education:
- Date of learning computer skills, - Date of learning internet skills, - Place of learning computer skills, - Place of learning internet skills.
Expenditure for ICT services:
- Monthly expenditure for cellular phone.

2.1. Data Preparation

2.1.1. Conversions, Re-assignments and Data Standardization

Data set from a questionnaire should not be used directly in clustering process and some transformation, conversion and preparation techniques must be used [5].

An unprocessed data set may contain some anomalies that should be summarized. Hence, a decision rule is required about whether all inputs are necessary or not. Furthermore, all questions, which were asked, may not be in accordance to numeric answers. As in the question "reasons of usingcomputer" text data should be converted into numeric interval data by categorizing them within the range of total answers given, instead of every single reason, total amount of reasons is more valuable in this concept. For numeric inputs, some data conversion techniques may need to be used. But data type conversions can be skipped due to usage of data standardization or similarity metric. Finally, unprocessed data set may have outliers (such as 9,99,999 for special conditions) or missing values, which create inconsistencies during similarity measurements and causes subjects to be measured as similar or dissimilar with each other. In these cases these values should be converted to an ineffective value (such as 0 to make corresponding mean value 0). In the data set, variables 9, 99 and 999 which were

referred to "not answered" and/or "not known", are re-assigned to 0 in order to prevent any probable inconsistencies. With such conversions and re-assignments, object amount in the data set decreased to 48 clear answers, without any loss of information, from 100 optional answers which contains missing, improper, separated or outlier answers because of the nature of a questionnaire.

2.1.3. Similarity Measures

Similarity metrics in clustering refer to the quantification of how similar subjects are. All answers to different questions in a questionnaire, may contain different data types (such as binary data for respondent's gender), in different ranges (such as max 13 spoken languages or max 30 reasons of internet usage).For this purpose, although there are several methods as mentioned in [5], Euclidian distance which directly differentiates the same objects (answers) is used as a simple and realistic approach:

$$D_1(x_1, x_2) = \sqrt{\sum_{j=1}^{P}(y_{1j} - y_{2j})^2}, \quad (1)$$

where x_1 and x_2 are 1st and 2nd respondents, y_{1j} and y_{2j} are standardized j^{th} responses of 1st and 2nd respondents and P is the total number of questions to be considered.
This metric will also lead us to obtain more compact clusters, which may help to prove the clusters' validity, as will be described in following sections.

3. CLUSTERING

A clustering analysis has different steps depending on the selected clustering method as in [7]

1. Selection of clustering elements and variables,
2. Variable standardization,
3. Measure of association,
4. Clustering method,
5. Number of clusters,
6. Interpretation,
7. Testing and replication.

Clustering methods can be classified into two groups, one is hierarchical clustering methods and the other is the iterative clustering methods. Within iterative clustering methods two different types; k-means clustering and fuzzy c-means clustering methods are the most frequently used ones.

Hierarchical clustering is a classification method that separates clusters with respect to the first and last measured distance metrics. Mainly, there are two hierarchical clustering methods in use; agglomerative and divisive methods. In our work, we have selected divisive methods and Ward's method in hierarchical clustering since it is the most frequently used method in the literature [8].

The k–means clustering algorithm is a clustering technique that determines new cluster members according to first cluster center and calculates new cluster centroids, iteratively.–At the k^{th} iterative step, subjects (respondents) are distributed among K cluster domains, using

$$S_i^{(t)} = \left\{x_p : \left\|x_p - z_i^{(t)}\right\| \leq \left\|x_p - z_j^{(t)}\right\| \; \forall\, 1 \leq j \leq k\right\}, \quad (2)$$

where, x_p is a new member which will be assigned to a clustered group S_i , according to the distance to i^{th} cluster's center z_i and j^{th} cluster's center z_i that is measured on t^{th} iteration step.

Then new cluster centroids z_j are computed at $(t+1)^{th}$ iteration

$$z_j^{(t+1)} = \frac{1}{\left|S_i^{(t)}\right|}\sum_{x_j \in S_i^{(t)}} x_j. \quad (3)$$

Where S_i is the clustered group particularly at t^{th} iteration with its x_j members (respondents).

We need to note that k-means clustering is differentiated from hierarchical clustering, due to its iterative strategy. In k-means new cluster centroids are calculated at each step. This illustrates a dynamic approach at creating clusters.

Fuzzy clustering is the clustering method that is derived from fuzzy logic. According to fuzzy logic, true and false expand into an interval [0, 1]. This characteristic lets fuzzy sets to have an opportunity each subject be an element of more than one cluster.

Because of the mentioned characteristic above, a so-called Cluster membership grade (u_{kj} – membership value of j^{th} subject to k^{th} cluster) of each subject is calculated. This grade is a value in the interval [0,1], and the goal is to optimize the value of the objective function

$$J_m(U, Z\,; X) = \sum_{k=1}^{c}\sum_{j=1}^{M}\left(u_{kj}\right)^m D_{kj}, \quad (4)$$

where m is defined as the fuzzifier [9] which is mostly set to 2 and effects the final membership distribution. D_{kj}, is the distance measure which was described as in similarity measurements [10].

With fuzzy c-means clustering, a subject may be assigned to more than one cluster, and this obviously can be used as an advantage during clustering data sets such as questionnaires. In this work, all of the above-mentioned clustering techniques are considered in combination with standardizations and different similarity metrics.

3.1.1. Number of Clusters

In the literature, there is no identified rule about determining exact amount of clusters. The most frequently used method is to examine both the agglomeration schedule and the dendogram [11]. An inconsistent increase in the dissimilarity measure would indicate that the clusters joined at the corresponding stage were quite distinct.

Although we manage to find optimum cluster amounts, for specific clustering methods, by determining minimum value of "compactness and separation criteria" which will be described in next section: Determining the number of clusters a-priori will provide more necessary and meaningful output then expecting a proper amount of clusters by empirical methods. Because it is planned to determine the level of digital literacy and set, we expect to see three different groups that state their cluster members as:

1) Digital literates,

2) Digital immigrants, and

3) Digital illiterates.

Therefore the data set is clustered into three classes.

3.1.2. Testing the Stability and Validity of Clusters

Stability of a cluster implies a high possibility of re-creating the same clusters as the outcome of similar clustering processes. Hence, stability is a measure of robustness of the selected clustering method.

According to [11] the most frequently used clustering strategy is to randomly divide the study sample into two halves and repeat cluster analyses for each. If the clusters are stable, a similar cluster structure should be obtained in each half of the sample. Furthermore, applying different clustering methods to the same data set and comparing them may be another approach, as we investigate in detail in our work.

Based on the obtained results, we have observed that different clustering techniques can result with very similar cluster labels. [12] States there should be at least four or five times as many observations as the variables to be analyzed for our data set, which is based on 48 objects/answers, considering 5 as the reliability constant, that 48×5= 240 subjects are sufficient to properly cluster and analyze. Based on this, stability of our cluster analysis can be measured.

On the other hand even after careful analysis of a data set and the determination of a final cluster solution, we have no assurance of having arrived at a meaningful and useful set of clusters. A cluster solution will be reached even when there are no natural groupings in the data [13]. Hence there must be some tests to prove validity of the cluster solutions.

Tests may need to be adjusted according to the clustering method as well. For instance, for k-means a validity measure which measures the compactness of the clusters can be defined. In [5] it is named as "compactness-separation criteria" and defined as

$$V = \frac{M_{intra}}{M_{inter}}. \tag{5}$$

Here M_{intra} is intra cluster distance, which means distance sum of subjects located in that cluster, and calculated as

$$M_{intra} = \frac{1}{n} \sum_{i=1}^{k} \sum_{x \in C_i} \|x - z_i\|^2 , \tag{6}$$

and M_{inter} is inter cluster measure which means minimum distance from one cluster to another which can be obtained according to

$$M_{inter} = \min_{1 \le i < j \le k} \|z_i - z_j\|^2 . \tag{7}$$

Our goal is to minimize the value of V since the most compact and separated cluster outputs are desired as the outcome. This objective can also be used to find the optimum number of clusters as well.

Table 3. Chosen objects for validation

	Question	'Yes'	'No'
1	Having a cell phone ?	1	0
2	Did you chose the model and technical properties of your cell phone ?	1	0
3	Usage of 3G ?	1	0
4	Mobile internet connectivity ?	1	0
5	Computer Usage ?	1	0
6	Fixed Internet Usage ?	1	0

During empirical tests of clustering, it is also desired to determine the approximate amount of clusters for later comparisons. This criterion can be applied to any clustering technique. Furthermore, because fuzzy c-means let subjects to be included in different clusters, specific validity indices for fuzzy c-means are given in various resources [14].

In our context, we propose a different validation method that uses specifically chosen objects within studied questionnaire. Those selected objects that are listed in Table 3 are related with and can reflect the digital literacy (digital gap) of participants to determine and quantify digital divide.

As for the proposed validation algorithm, the first step is to calculate all cluster centroids and their Q values with respect to specified objects (Table 3) of the studied questionnaire. Here Q is the predisposition of clusters and can be calculated as

$$Q_i = \frac{1}{L} \sum_1^L c_{il} \ , 1 \le i \le k , 1 \le l \le L . \tag{8a}$$

Where we consider six objects (L=6) and (k=3) clusters. Following this step, according to Q values, clusters satisfying the conditions (8b), (8c) and (8d), are assigned to the expected clusters: digital literates, digital immigrants and digital illiterates,

$$S_{\text{Digital Literates}} = \left\{ \begin{matrix} S_i : max\left\{\frac{1}{L}\sum_1^L c_{il}\right\} \\ \forall\ 1 \le i \le k , 1 \le l \le L \end{matrix} \right\}, \tag{8b}$$

$$S_{\text{Digital Illiterates}} = \left\{ \begin{matrix} S_i : min\left\{\frac{1}{L}\sum_1^L c_{il}\right\} \\ \forall\ 1 \le i \le k , 1 \le l \le L \end{matrix} \right\}, \tag{8c}$$

$$S_{\text{Digital Immigrants}} = \left\{ \begin{matrix} S_i : min\left\{\frac{1}{L}\sum_1^L c_{jl}\right\} < \left\{\frac{1}{L}\sum_1^L c_{il}\right\} < max\left\{\frac{1}{L}\sum_1^L c_{jl}\right\} \\ \forall\ 1 \le i \le k , 1 \le j \le k, 1 \le l \le L \end{matrix} \right\}. \tag{8d}$$

Assigning cluster labels and determining the amount of respondents within a cluster, will enable us to calculate a validation ratio by exposing how much a subject is really suitable to be included in the corresponding cluster. For the digital divide concept in this paper, with respect to the objects and the variables, the assigned cluster label may be considered "correct" if one of the conditions given below is satisfied

$$\frac{1}{L}\sum_{1}^{L} x_{jl} > 0.66 \ \Lambda \ S_i = S_{\text{Digital Literates}} ; \quad 1 \le i \le k, 1 \le l \le L, \tag{9a}$$

$$\frac{1}{L}\sum_{1}^{L} x_{jl} < 0.33 \ \Lambda \ S_i = S_{\text{Digital Illiterates}} ; \quad 1 \le i \le k, 1 \le l \le L, \tag{9b}$$

$$0.66 > \frac{1}{L}\sum_{1}^{L} x_{jl} > 0.33 \ \Lambda \ S_i = S_{\text{Digital Immigrants}} ; \quad 1 \le i \le k, 1 \le l \le L. \tag{9c}$$

3.1.2. Determining questions commonly answered

After clustering steps were concluded, common characteristics of clusters' respondents were determined. Our goal here was determining main reasons of digital gap by deriving them from the common answers given by respondents. Therefore within all clusters, for top three cluster techniques, standard deviation of standardized objects was calculated (10).

$$\sigma_j = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \left(y_{ij} - \mu\right)^2} \tag{10}$$

Where σ_j is standard deviation, N is the total number of respondents (1140), y_{ij} is the j^{th} object (answer) of i^{th} respondent and μ is the average of whole answers (objects) of j^{th} question.

Smallest standard deviations (weighted %10 of total) of j^{th} questions, such as "Date of learning computer skills", "Internet usage and frequency", "Reasons of internet use" so on, show: Answers to those questions are so similar within all along the cluster.

4. RESULTS

MATLAB software is used for clustering analyses. Actual questionnaire data, collected from 1140 distinct respondents from different neighborhoods in Istanbul is standardized and clustered. 6 clustering techniques; pure k-means (C1), z-score k-means (C2), USTD k-means (C3), hierarchical clustering (C4), controlled k-means (C5) and fuzzy c-means (C6) techniques have been used. Respondent percentages of clusters per cluster technique are given in Table 4.

Stability of the outcomes has also been compared and shown in Table 5. Not surprisingly, all clustering techniques applied to standardized data set and all clustering techniques applied to non-standardized data set were found out to be similar. Also, non-standardized clustering methods have given similar outputs.

Comparison of clustering methods was the way of measuring stability of the clustering analysis. But even though stability test has given proper values, validation of analysis outcomes was still needed. For this purpose, Q-valued validation method proposed in previous sections and inverse "compactness and separation criteria" value (V^{-1}) were used by calculating them and given in Table 6. Participants, who were clustered by C2 and C3 clustering methods, proved that they belong to the cluster they are naturally supposed to be with over 70 % of validity. On the other hand, most compact and separated clusters were determined as C4 and C6 with their values above "1".

By these validation values, to decide which cluster methods are best to separate the most compact and correctly allocated clusters; we multiplied our validation ratio with inverse V with same weighting as an acceptance value. According to the observed results, C4 and C6 were the best clustering methods to use in our concept. Those clusters were also analyzed to find out low variety answers and so the possible reasons of digital gap.

Furthermore, Table 7 shows the distribution of digital literacy through C4 and C6 clustering methods by neighborhoods in Istanbul. According to these results, one can notice the spatial distribution of digital literacy differs with respect to the clustering methodology.

5. CONCLUSION

The application of knowledge is a primary source of growth in the knowledge economy and sustainable communities. There are many models and methodologies proposed both in the literature and practice to determine the indicators of the ICT use performance of the communities. However, much of this work is carried out at regional or national scales. The focus of this paper is on the development of a methodology to monitor progress towards a knowledge economy at intra-city level through clustering algorithms.

To this aim, we quantify the digital gap between ICT users intra-city level and show that although users may be located within a close proximity, the digital gap between them may be non-negligible (as shown through clustering analysis results). Hence, we propose that considering the spatial resolution of digital divide quantification level in nation-wide scale is definitely insufficient to analyze and assess the ICT usage status. Additionally, a generalized framework for questionnaire clustering is proposed in this paper. Necessary steps were described and a novel validation technique is proposed. According to the results, top three cluster methods were determined by over % 70 validations

In future work, two possible next steps were outcome as opportunities.

First, the opportunity to deeply analyze low variety answers within clusters. Analyses will possibly lead us to reasons of digital gap from micro-scale level to nation-wide. Therefore new telecom policies may be proposed to close these gap which was investigated at intra-city level.

Second, the opportunity to develop a novel mobile network strategy based on digital gap findings. Correlating digital literacy with data rate requirements may lead to demand oriented network topology, exact hardware and energy needs by using telecommunication and optimization knowledge.

Table 4. Assigned clusters and their sizes

	Digital Literates (%)	Digital Imigrants (%)	Digital Illiterates (%)
C1	23	35	42
C2	51	26	23
C3	51	26	23
C4	31	35	34
C5	23	35	42
C6	32	24	44

Table 5. Comparison of clustering methods

	C1	C2	C3	C4	C5	C6
C1	1	0	0	0	0	0
C2	0.58	1	0	0	0	0
C3	0.58	1.00	1	0	0	0
C4	0.84	0.56	0.56	1	0	0
C5	0.99	0.57	0.57	0.84	1	0
C6	0.85	0.59	0.59	0.78	0.84	1

Table 6. Clusters' validations

	Validation Ratio	V^{-1}	Acceptance
C1	0.58	0.0428	0.02500
C2	0.76	0.6950	0.53041
C3	0.76	0.6950	0.53040
C4	0.66	1.3285	0.87285
C5	0.58	0.0428	0.02490
C6	0.56	1.2625	0.71099

Table 7. Digital Literacy for the first 5 neighborhoods

C4			
Neighbor. No	**Digital Illiterates**	**Digital Immigrants**	**Digital Literates**
1	2	16	16
2	3	12	14
3	4	17	11
4	11	10	13
5	14	14	11
C6			
Neighbor. No	**Digital Illiterates**	**Digital Immigrants**	**Digital Literates**
1	4	15	15
2	6	7	16
3	9	9	14
4	14	9	11
5	19	7	13

REFERENCES

[1] Organisation for Economic Co-operation and Development – OECD, Understanding the digital divide. OECD Publications, France, 2001.

]2] WSIS (2003). Geneva Plan of Action. Geneva. (WSIS-03/GENEVA/DOC/5-E.)

[3] WSIS (2005). Tunis Agenda for the Information Society, §§ 28, 113 – 119. Tunis. (WSIS-05/TUNIS/DOC/6(Rev. 1)-E.)

[4] Measuring the Information Society: Information Development Index, International Telecommunication Union (ITU), 2009

[5] Gan, Goujun, Chaoqun Ma and Jianhong Wu, Data Clustering: Theory, Algorithms, and Applications, ASA-SIAM Series on Statistics and Applied Probability, SIAM, Philadelphia, ASA, Alexandria, VA, 2007.

[6] Busse, L. M., Orbanz, P., Buhmann, J. M., "Cluster Analysis of Heterogeneous Rank Data", *ICML 2007 proceedings*, 2007.

[7] Milligan, G.W. Clustering validation: results and implications for applied analyses, in Clustering and Classification pp 341-375. World Scientific, Singapore 1996.

[8] Ward, J. H., Jr., Hierarchical Grouping to Optimize an Objective Function, J*ournal of the American Statistical Association*, 48, 236–244, 1963.

[9] Everitt, Brian S., Sabine Landau, Morven Leese and Daniel Stahl: Cluster Analysis, Wiley Series in Probability and Statistics, WILEY, 2011.

[10] M.S. Yang, "A survey of fuzzy clustering", *Math. Computational Modeling*, 18 (11) 1–16, 1993.

[11] Clatworthy, J., Buick, D., Hankins, M., Weinman, J. and Horne, R., "The use and reporting of cluster analysis in health psychology: A review", *British Journal of Health Psychology*, 10 pp 329–358, 2005.

[12] Hair, J.F., Anderson, R.E. and Tatham, R.L., "Multivariate Data Analysis with Readings", 2nd ed., Macmillan, New York, 1987.

[13] Punj, G., Stewart, D. "Cluster analysis in marketing research: Review and suggestions for application", *Journal of Marketing* Research 20 pp 134–148, 1983

[14] Sadaaki, Miyamoto, Ichihashi Hidetomo and Honda Katsuhiro, Algorithms for fuzzy clustering : methods in c-means clustering with applications, Springer , Berlin, c2008.

ICT INNOVATION IN SOUTH AFRICA: LESSONS LEARNT FROM MXIT

Michael Kahn
Centre for Research on Evaluation, Science and Technology
University of Stellenbosch, South Africa

ABSTRACT

In the last decade South African innovators have produced two game changing innovations: Thawte and Mxit, the former for Internet security and the latter for mobile messaging. It is of interest to understand what it is in the local system of innovation that has enabled such innovations to emerge. Mxit is one of the first instant messaging systems to be freely available for almost all mobile phones. From a simple text service it has now evolved into a multimedia platform that offers gaming, education, community support services, and that is poised to enter the money transfer and payment market. The paper locates this development in the context of the telemetry sectoral system of innovation of South Africa, and the innovation ecology of the city of Stellenbosch that is designated as 'Silicon Vineyard.' Mxit has particular importance as a means of bridging the various divides that continue to characterize post-Apartheid South Africa.

Keywords— Mxit mobile messaging, telemetry sectoral system of innovation, R&D spillovers, mathematics education, payment system

1. INTRODUCTION

It is an old adage that's there's always something new out of Africa. Recent history would attest to this: the world's first heart transplant (1967) and the most effective Internet security algorithm (2000) serve as two examples. This paper draws attention to a third, the successful development and dissemination of a social networking technology on mobile platforms, known as Mxit (pronounced 'mix it.').

Mxit stands alongside Kenyan M-Pesa money transfer by mobile telephony as important African-originated game changers that to the surprise of many have emerged from two middle-income countries. Mxit has become the dominant player in the local mobile messaging environment.

This paper is structured as follows: the second section provides an overview of the South African system of innovation; next follows discussion of the coincident emergence of mobile telephony with the new democratic order of 1994. This provides the background to the origins of Mxit as an affordable means of electronic communication. The fifth section considers Mxit today, while the sixth examines the particular innovation environment in Stellenbosch, 'Silicon Vineyard.' The final section offers perspectives on mobile platforms going forward.

2. THE SOUTH AFRICAN SYSTEM OF INNOVATION

The exploitation of diamonds and gold catalyzed the agglomeration of the Cape and Natal Colonies, the Boer Republics and a range of African Kingdoms into the Union of South Africa. The mining-led industrial revolution then transformed the agricultural economy into an important adjunct of the world financial system. Universities were founded, public research organizations and technical institutes were established, and industry diversified. Today three South African universities are included in the Shanghai Jiao Tong world top 500 higher education institutes; the Times Higher Education Supplement rates the University of Cape Town 103rd, and its Business School 53rd.

Considerable wealth was generated in a short space of time, and in the rush to extract diamonds and gold, technologies were imported as and when they were needed. Accordingly the country was an early adopter of technologies – electric street lighting being one of such. (This tradition of early adoption persisted over time with the largest mainframe computer in the Southern hemisphere being installed in Cape Town as early as 1959.)

Alongside the importation of technology local scientific and technological capacity led to the emergence of a local innovation system. The First World War saw further diversification of industry, and at the end of hostilities the modernizing state established public monopolies in transport, electricity generation and distribution, bulk water supply, telecommunications and iron and steel. At this time state-sponsored industrial research was located in the Department of Industry and Mines. Institutes for medical, agricultural and animal research were all put in place, including the renowned Onderstepoort Veterinary Laboratory. By the 1930s a dedicated state laboratory for mining and metallurgy functioned alongside those within the mining houses. Local scientists undertook doctoral

studies abroad, gaining recognition in nuclear physics, electronics and crystallography.

The decisive moment for the innovation system came through World War 2 that saw this local capability harnessed to the Allied war effort for military vehciles, artillery, munitions and supplies, and in particular the development of radar technology. The competence to work on vacuum tube radars was based on the prior university research in electronics and crystallography that involved glassblowing techniques, applications of electromagnetism and knowledge of instrumentation. This competence, together with basic information on the principles of radio ranging allowed local scientists to develop an indigenous version of radar specially suited to tracking shipping and low flying aircraft [1].

After 1945 these capabilities in physics, chemistry, building science and psychometrics were housed in a new public research institute, CSIR. Some ten years later the radar specialists adapted microwave ranging technology to carry out geodetic surveys to an accuracy of 1 part in 100 000 over 50km, a technology commercialized as the Tellurometer. Next came civil war and isolation during which the innovation system concentrated on self-sufficiency with innovations in telemetry, encryption, military vehicles, missiles, artillery, fuels, chemicals, pulp and paper, agriculture and mining. By 1994 the system of innovation comprised a network of thirty-six higher education institutions, eight science research councils, some fifty department based research institutes, and scores of R&D labs in industry, supported by extensive scientific and technological services. By design Apartheid excluded the majority of the population from the innovation system proper, thereby creating a persistent skills backlog. Even so the economy was (and still is) the most advanced in Africa, producing more than 40% of the continent's scientific outputs and patents.

These efforts laid the basis of a number of sectoral systems of innovation [2] that linked companies, enterprises and government laboratories in specific industrial sectors. The telemetry sectoral system of innovation (SSI) has two wings. The first and older is mainly located in the Western Cape, with its major node in the city of Stellenbosch. This SSI stretches from the missile test ground at Arniston through Stellenbosch to Cape Town and south to the Simon's Town naval base. It includes the University of Cape Town, The Cape Peninsula University of Technology, the Institute of Maritime Technology, the Naval Warfare School, and a range of private firms.

The second wing of the telemetry SSI developed to meet the aerospace needs of the army and air force and is centered in Gauteng, the mining, industrial and financial hub of the country. Pretoria University, the Tshwane University of Technology, University of the Witwatersrand, CSIR, state arms group Denel, and industry partners stretch over a 200km axis at one end of which are the radio telescopes at Hartbeeshoek, and at the other the aircraft industries of Kempton Park.

Returning to Stellenbosch, one also finds companies offering late stage venture capital, of which PSG and Remgro, and latterly World of Avatar are examples. The city is renowned for its natural beauty, good food, fine wine and music. The ingredients for fostering innovation are thus in place. The telemetry capacity, plus the considerable software engineering that supports the local financial services sector, especially insurance, which is also centered in the Western Cape, provide the financial, intellectual and human capital for I would call 'Silicon Vineyard.'

3. MANDELA, MONOPOLIES AND MOBILES

The new democracy under Mandela's 1994 government faced massive challenges: dismantling Apartheid; globalization and the end of trade isolation; accession to the World Trade Organization and deregulation of agriculture; state modernization; poverty; the ongoing ICT revolution. The challenge came down to willpower, funding and capability.

Some of the funding for this agenda of transformation could come from the 'restructuring of state assets,' privatization in all but name. Two of the monopolies of the Apartheid state, Sasol and ISCOR had already already privatized, now, in the quest for local sources of finance other privatizations were mooted, with telecommunications, in the person of state monopoly Telkom high on the list.

The privatization of Telkom and deregulation of broadcasting and networks has been a long and difficult process [3], with many about turns. A case in point is the early and ill-fated quest for universal access on the Australian model that failed, with rows of fixed line coin boxes standing unused.

The reason was simple. Mobile arrived at exactly the time when restructuring was in vogue, and then took off, far beyond expectation. Coin boxes were expensive, in the wrong place, and inconvenient. Today, based on SIM users, the 2011 mobile penetration rate for South Africa stands at 117.8% with fixed lines in the order of 4.4 million (2008 value). By comparison the mobile penetration rate for Kenya is around 75%, with an unknown but very small fixed line rollout. In both countries the mobile Internet penetration rate is much higher than that for landline access.

In 1994 however, this was still to come. With the advent of portable, lower cost mobile telephony, two local companies gained licences to enter the market - Vodacom, a joint venture between Telkom and UK Vodafone, and a local company, MTN. They were later joined by a third provider Cell-C that is part owned by Saudi Oger. Telkom initially provided the network; today Vodafone and MTN host their own networks, but most critically all mobile companies are

subject to the Telkom 'last mile' monopoly with its high price tag.

Vodacom and MTN were the market pioneers that sought to increase their user base and network utilization. Though there was local manufacturing capability in transformers and power infrastructure, the routers and repeaters for mobile were imported. Local innovation went into consumer products, such as the innovation 'please call me' that appeared on both networks in 2001 allowing a user without airtime to request another user to call them, thereby ensuring greater network use. This was the first point where wealth asymmetry was mitigated through mobile telephony.

At the turn of the millennium entrepreneur Mark Shuttleworth was synonymous with 'Silicon Cape' as the inventor of Thawte Internet security. He sold Thawte to Verisign for USD 575 millions at the peak of the dotcom boom and went on to engineer the Ubuntu operating system, LibreOffice free productivity software, and maintains a strong interest in promoting science and mathematics education. Shuttleworth studied business science (finance and information systems) at the University of Cape Town and was drawn into Internet service provision whilst a university student.

4. ENTER MXIT

Mxit, as for Thawte, is associated with an entrepreneurial individual, Herman Heunis. The biography of Heunis bears some resemblance with that of Shuttleworth in that he too graduated with a business degree (from Stellenbosch University) rather than one in the natural sciences or engineering.

What followed Heunis' graduation was critical: he spent two years doing national service in the SA Navy electronic warfare division (in Simon's Town) at the heart of the telemetry sectoral system of innovation. There he received intense training as a software engineer, becoming a highly skilled programmer and analyst. His next five years were spent at the Stellenbosch University computer center. This gave him exposure to, and insight into the academic research environment. Heunis is a product of the telemetry SSI – he worked in a state defense laboratory, at one of the leading research universities, and is now in business.

In 1990 Heunis left Stellenbosch University and started his own software company in that city. With the onset of mobile telephony he founded Swist that began providing software for Vodacom and MTN, thereby gaining intimate knowledge of how mobile telephony technology worked. This, with his skills as a programmer provided an inside track for his later development of mobile applications.

Heunis' first attempt to enter the mass mobile market was in 2000 via mobile gaming technology when Swist morphed into 'a lifestyle company' offering mobile services, clothing, music and eventually its own currency. Swist then set up Clockspeed Mobile that launched Alaya, an SMS based game that though well received, failed to achieve wider uptake because of the high cost of texting.

It was recognition of this cost barrier that persuaded Heunis and his research team to re-think. In effect he set out to emulate Internet protocol for mobile data transfer, much like VOIP had already done. This led to the development of Mxit. The app went live in 2003 and its take off was exponential as cash-strapped teenagers discovered an affordable way of contacting one another.

Swist's approach to intellectual property rights was simply 'to out innovate the opposition.' It has been able to stay ahead of the game by its willingness to pay a salary premium for the best staff. By the time of Heunis' exit from Mxit Lifestyle in 2011 for the sum of ZAR 330 millions, they had not registered a single patent – the cost of filing and defending patent rights was viewed as simply too expensive. Indeed up to that point Swist had yet to turn a profit on Mxit.

A need for further capital led to Swist selling a 30% stake to South African media house Naspers, a company listed in Johannesburg and New York. Eventually Heunis in 2011 sold his interest in Mxit to local investment company World of Avatar, headed by Alan Knott-Craig Jnr. World of Avatar describes itself as a company that is committed to `Africa's development by funding the creation of apps that 'that enable the people of Africa to make a better living using their mobile phones' [4]

It may be noted that Alan Knott-Craig Jnr is the son of the outgoing CEO of local mobile telephony giant Vodacom. Like Shuttleworth and Heunis, Knott-Craig Jnr also has a business background, being a chartered accountant with a BCom (Hons) from the University of Port Elizabeth.

The implication is that, at least in the South African environment, being a leader in the world of connectivity does not require a degree in computer science, let alone postgraduate study. This resonates with R&D Survey data [5] that suggest that large firms in the services sector – retail, banking and insurance do not to employ R&D staff with doctoral degrees. They are able to meet company needs with research staff holding undergraduate and masters degrees.

MXit now stands alongside a number of other instant messaging systems such as BBM, Viber, Whatsapp, Kik, Google Talk, Line and Kakao. It permits text and multimedia messaging that runs according to its own open standard on XMPP. MXit, like other parties engaged in public electronic communication is subject to regulatory approval by the Independent Communications Authority of South Africa. Mxit has captured a gap in the market for low-cost communication. This gap persists because of the pricing policy of the Telkom monopolist - the country is

ranked 63rd of the 64 countries for cost of data in the Ookla survey, and according to Netindex is behind Rwanda, Kenya and Ethiopia for speed of download.

What sets MXit apart is that it is a free app that has been engineered to run on no less than 3000 different handsets, both basic and smartphone. These two factors explain its considerable market penetration that places it in the same league as the market leaders such as Whatsapp. MXit has particular appeal to teenagers who often make use of hand-me-down telephones that are running slow operating systems. MXit has 50 million users that generate some 23 billion messages each month.

The app runs on Microsoft Windows, Microsoft Mobile, Mac OS X, Android, iOS, Java ME, Linux and Blackberry OS and is developed and maintained by a software engineering team of sixty staff, based in Stellenbosch. MXit is officially supported across the European Union, the USA, Nigeria, Kenya, Brazil and Malaysia and has a large server farm in Germany. The MXit web site hosts a developer zone allowing access to the MXit client protocol.

Having successfully developed a free IM service, Swist re-launched the games platform. Gaming that sends data via GPRS can now run at close to zero cost, so to speak. This is the reason why MXit is so attractive to children and teenagers.

There are another three important innovations that rely on the MXit platform. The first is a telematic student support program known as Dr Math that links school students with remote volunteer tutors who assist them to deal with difficulties they encounter with primary and secondary school mathematics [6].

The quality of education generally, and science and mathematics in particular, is very weak in South Africa, with the school system being essentially three tier – a very high quality fee paying private sector; a quality fee paying component of the public sector; poor quality in the bulk of the public school sector. The top universities, public research institutes, state weapons labs, and private sector R&D labs constitute a small innovation system within a larger sea of underperforming institutions. The Global Competitiveness Index data for 2011 thus reports that South Africa's financial market development was 9th in the world while its primary education was ranked at 129th [7]. Dr Math targets the public school sector.

Dr Math went live in 2007 and is operated by the ICT Meraka Institute of CSIR in collaboration with the University of Pretoria, a pioneer in telematic education. The University has considerable experience in various forms of tutor-student telematics functioning, via telephone as well as the Internet.

In order to protect minors from possible abuse and exploitation, Dr Math is designed to work under the guarantee of student-tutor anonymity. Since minor children are in effect handing over their mobile contact number to a stranger this is a necessary precaution. Accordingly all text conversations are recorded for security auditing but also to generate a research database allowing the later study of linguistic styles and cognitive difficulty.

In addition an in-house software package termed MXit Understander has been developed to assist the Dr Math tutors to decode the highly abbreviated way that MXit text users communicate with one another. For times when tutors are unavailable, the Dr Math portal allows users to play games that enhance mathematical problem solving skills such as single-user text adventure games, multi-user arithmetic, algebra and multiple choice quiz competitions, as well as math encyclopedia functions. Dr Math thus seeks to build sustainability by strengthening the school system.

The second innovation is collaboration with Rlabs [8] that emerged in 2009 out of the work of Impact Direct Ministries, a non-governmental organization dedicated to community-based upliftment through counseling drug addicts and supporting family life. Reconstructed Living Labs is headquartered in Athlone, Cape Town and has branches in the United Kingdom, Europe, Asia and Central Africa. It offers consulting services for the use of social media to address social problems, with MXit providing the mobile connectivity for Rlabs and its deployment of the JamiiX platform for low cost call center services [9].

The third is the role of MXit in financial services. An early innovation was the introduction of a micro payment system MXit Moola (moola is a local slang term for money). Moola is MXit's own 'currency' and is available for purchase in South Africa, Kenya, Namibia, Lesotho, the United Kingdom and Indonesia. One Moola unit is approximately one South African cent, and the system conforms to local banking regulations. Moola currency purchases may be made from mobile phones or via the web site of a high street bank. Moola may be used to buy games, mobile skins, wallpaper and music.

The Moola experience with micro payments has now led MXit to explore the feasibility of an open payment system. This innovation has considerable potential in South Africa and other African countries where many people function outside the conventional banking system and rely on informal means to transfer money point to point. Such money transfer systems often rely on trust, but are slow and involve high overhead costs.

To enter this market MXit has partnered with UCS subsidiary wiWallet to develop a rival to M-Pesa and other money transfer systems. UCS is a company with extensive experience in encryption technology, being the main provider on the continent of digital subscriber television technology. UCS skills in encryption came through

spillovers from the cybernetic R&D programs of South Africa's military efforts in the 1980s.

At the time of writing MXit Money was still under in-house development. It is described as a platform, not a bank or wallet. Trials with several thousand users, including employees of MXit have been undertaken to understand their expectations and to engineer features that users want.

5. INNOVATION IN SILICON VINEYARD

Mxit is embedded in Stellenbosch city, with its University and associated Techno Park in which many of its graduates are running their own businesses. A number of these businesses are part of the telemetry SSI. So for example, Sunspace a spinout company of the University produces microsatellites. The Sunspace team makes use of the prior investment of the 1980s when state weapons company Armscor was developing ballistic missiles for which it required clean rooms, test facilities, and command and control installations, all of which are in the immediate locality of Stellenbosch.

Another innovative company in the Techno Park is EMSS that was started by two freshly minted electrical engineering doctoral graduates. Upon graduation their expected positions in the defence force did not materialize, and faced with the prospect of unemployment they began to seek other opportunities. Their understanding of electromagnetic theory then enabled them to spot a promising research finding in the literature. They believed this could have commercial application and entered into discussion with the article author to exploit his intellectual property and then developed a system for mapping radar electromagnetic emissions.

As in the case of Heunis' company Swist, it was the advent of mobile telephony that opened up a new market for EMSS, in the latter case to meet the need for measuring cellphone tower emissions near to human habitats. EMSS now employs a hundred staff and has clients across the OECD member states. EMSS is also the designer of the detection system for the Karoo Array Telescope, the precursor to the massive Square Kilometre Array (SKA) telescope.

A fourth company that was attracted to the Techno park is Reutech Radar, part of listed company Reunert, with a history in radar and military communications equipment long preceding the founding of the Techno Park.

Stellenbosch University Electronic Engineering has played a pivotal role in enabling the careers of the individuals behind these companies. Its course offerings include electrical energy, electronics, electromagnetic systems, computer systems, control systems, telecommunication systems, informatics, robotics and signal processing, which have relevance to radar and mobile communications, to telemetry in general.

The weak state of public education in South Africa remains a significant bottleneck for all high-level skills production. Indeed there is a national shortfall that has led to significant salary escalation.

A matter of considerable importance is to understand the sources of information for innovations such as Mxit. Evidence from many countries in which OECD/Eurostat innovation surveys have been carried [10] out shows that the prime source information for innovation is mainly from internal sources, followed by interaction with customers, suppliers and competitors, in other words from the company value chain and its competitors. Non-market knowledge producing institutions, namely universities and public research organizations are not directly involved in innovation, and innovation survey respondents generally rank such non-market players as less important as sources of information for innovation.

Mxit practice conforms to the 'standard' pattern for sourcing information for innovation [12], with most of its ideas being generated internally. The local university thus is a source of skilled staff, rather than ideas. The cases of Sunspace and EMSS were initially quite different as their ideas came from the academic environment with its strong links to military R&D. One might note similarity with Skype that originated in Estonia, also a middle-income country with a small innovation system, and a history of military cybernetics.

There is a now a strong movement in South Africa's universities to seek commercial benefit from publicly funded research. This has come about through the promulgation of an Act [11] to regulate such intellectual property and to allow for benefit sharing. The implementation of the legislation includes the establishment of a National Intellectual Property Management Office with counterpart technology transfer offices in some of the large research universities, Stellenbosch included. It is too soon to tell if these bureaucratic steps will stimulate or stifle commercialization.

It is the intense interaction of people and the MXit platform that then generates further ideas for exploration and possible implementation. One of the most difficult, if not painful tasks of management is to be able to identify those ideas that have already taken root as projects that must be culled in order to channel investment toward more promising solutions. The decision to terminate such project is based on market potential and cost factors. Careful cost accounting may assist in the decision, but ultimately the termination of a project is a management call. No matter how freewheeling or *avant-garde* the day-to-day management style may be such decisions have to be made.

The informal work environment that characterized the early days of Mxit has been retained, and if anything has evolved even further, with the new owner Knott-Craig intent on ensuring that creativity and innovation are fostered. Mxit appears to have found a comfortable niche in the ecology of Silicon Vineyard centered on Stellenbosch.

6. CONCLUDING REMARKS

This article speaks to the creation of the mobile messaging platform Mxit in South Africa, middle-income country of extremes – technological innovation, high quality universities, and wealth, alongside a weak public education system and poverty.

Despite a regulatory regime that has promoted the state telecommunications monopoly that has restricted access to value added services and maintained a high cost of the last mile, there has been rapid uptake of mobile telephony.

This uptake was driven by the availability of affordable handsets, the passing on of handsets upon renewal of contract services every two years, and innovations such as 'please call me,' that brought tens of thousands on line. It was the recognition that there was space for a low cost instant messaging alternative to SMS texting that stimulated Herman Heunis to fund the development of Mxit through his company Swist.

The development of Mxit in turn depended on the availability of the necessary skills in project management, business modeling and software engineering, all of which were available in the Greater Cape Town area that is at the heart of the southern wing of the telemetry sectoral system of innovation.

It is contended that the prior investment of the 1980s in military cybernetic research and development paid off in the long term through the unintended spillovers that resulted from that R&D. Other cases of local spillover include the Sunspace microsatellite company and EMSS that produces software for mapping antenna electromagnetic emissions.

Elsewhere in the country a number of unique industries based on spillovers from military research have emerged – encrypted pre-payment technologies; digital subscriber TV; electronic detonators, to note but three.

These observations are not intended to encourage military R&D *per se.* Such R&D, like other fields, is often wasteful, and is subject to the uncertainties inherent in any R&D. One cannot precisely foretell the future outcome of R&D, let alone predict its timeline. Given the large budgetary flows to military R&D platforms of expertise arise from which civil applications may emerge. The Internet arose from a project of the US Defense Advanced Research Projects Agency that sought to ensure the resilience of distributed information systems. This is a perfect example of a military need spilling over into the civilian domain. On the other hand the World Wide Web was a serendipitous product that arose from the data handling needs of scientists at the CERN atomic particle accelerator. R&D spills over; R&D outcomes are unpredictable. The bulk of the Square Kilometre Array

Mobile telephony in South Africa, as elsewhere in Africa, has 'connected' the masses that were previously excluded from landline connectivity. Whilst 'connected,' mobile voice telephony remains an expensive luxury for most users. The high cost is essentially a consequence of telecommunications policy that seeks to maintain the Telkom monopoly. Text SMS is a more affordable option, but is still out of reach for many. Mxit, as has been shown, not only straddles the worlds of the haves and have nots, but serves to narrow some of the divide between them. It fills the connectivity gap, being cheap, intuitive and readily available, and has improved the quality of life of millions of users, not just through basic messaging but also through allowing a range of human-human and human-machine interactions. It is likely to remain a forerunner of things to come as the original Short Message Services fades into obsolescence.

REFERENCES

[1] B. Austin, 'Radar in World War II, the South African contribution," *Engineering science and education journal*, Vol. 1 No. 3, pp121-130, 1992.

[2] F. Malerba, "Sectoral systems of innovation and production," *Research Policy*, Vol. 31, pp247-264, 2002.

[3] S. Esselaar, A. Gillwald, M. Moyo and K. Naidoo, South African ICT Sector Performance Review 2009/2010. Towards Evidence-based ICT Policy and Regulation, Volume Two, Policy Paper 6, 2010

[4] http://www.worldofavatar.com/about-us/

[5] http://www.hsrc.ac.za/CCUP-RnD-7.ph

[6] L. Butgereit, "Math on MXit: Using MXit as a Medium for Mathematics Education," http://researchspace.csir.co.za/dspace/handle/10204/1614, accessed 8 September 2012.

[7] World Economic Forum, *The Global Competitiveness Report 2010-2011*, Geneva, pp515.

[8] http://www.rlabs.org

[9] http://www.jaamix.biz

[10] W. Blankley and C. Moses, *Main results of the South African Innovation Survey 2005*. HSRC Press, Cape Town, South Africa, pp200.

[11] Intellectual property from publicly funded research Act No 51 of 2008, Pretoria, The Presidency.

[12] Interview with H Heunis in Stellenbosch, 4 September 2012.

REVIEW OF CHALLENGES IN NATIONAL ICT POLICY PROCESS FOR AFRICAN COUNTRIES

Frank Makoza and Wallace Chigona

Department of Information Systems
University of Cape Town
Private Bag X, Rondebosch 7700
Cape Town, South Africa

Email: Frank.Makoza@uct.ac.za

ABSTRACT

National Information and Communication Technology (ICT) policies are vital in supporting socioeconomic development agendas. However, the formulation and implementation of national ICT policies is often beset with myriad of challenges which lender the policies ineffective in most developing countries. While there is substantial body of knowledge on the challenges for national ICT policies, no study has not yet dealt with the challenges holistically. This paper reports on the results of review of literature on the challenges for national ICT policy process in African countries using Grounded Theory method. The review categorised the challenges related to agenda setting, policy formulation, legal frameworks, implementation and evaluation. To mitigate some of the challenges, it is suggested that stakeholders' participation should be encouraged; monitoring and evaluation with mechanisms for learning should be integrated in all the phases of the policy process.

Keywords— National ICT Policy, Grounded Theory

1. INTRODUCTION

Most developing countries are incorporating Information and Communication Technologies (ICT) in their development agendas with the aim of supporting achievement of socio-economic development and alleviation of poverty [10, 32]. A national ICT policy embodies the plans of how ICTs may be used to support development [4, 33]. It is estimated that 85% of African countries have initiated the process of formulating and implementing national ICT policies [24]. Despite the impressive number, the impact of the policies is not evident [1, 20]. This may be indicative that national ICT policy process is still problematic in Africa.

There is a growing and substantial literature on the challenges for national ICT policy process in African countries [2, 13, 23, 28]. The studies highlight diverse problems in the policy process phases. However, to our knowledge there a few studies [3, 10, 52] that have holistically synthesized previous studies on the challenges for national ICT. For instance, a review of literature on Public IT polices between 1994 and 2002 [52]. The study focused on less developed countries which at that time included: Ireland, Singapore, China, South Africa, Mozambique, Nigeria etc. In their call for further research, the paper suggested consideration of work outside the field of Information Technology and concentrating on less developed countries [52]. In response to this call, we focused on studies on national ICT policy conducted in African countries. Although African countries may have disparities in political, economic, social and technological development but they share similar legacy of colonization and may be facing similar challenges [36]. Holistically synthesising literature on the challenges for national ICT policy process may improve the understanding of the problems and pave paths for future research [16, 45]. Additionally, such a study may be instrumental for practice and researchers in transforming national ICT policies [52]. The intention of this paper is to review previous studies conducted in Africa on the challenges in the process for national ICT policies and offer new perspective on how to mitigate some of the challenges. The study is guided by the research questions: What challenges or barriers face national ICT policy process? And how can the challenges in national ICT policy process be addressed?

The study employed Grounded Theory method to review studies on national ICT policy in Africa. Reviewing literature using Grounded Theory was considered ideal because the method support in-depth analysis of data and development of concepts or themes that are grounded on data which may be useful in explaining the challenges for national ICT policy process [18, 47].

The rest of the document is presented as follows. Section 2 outlines the background to the study and theoretical underpinning of the literature review. Section 3 highlights the research methodology. Section 4 summarises the results

of data analysis. Section 5 discusses the results and conclusion with implications for future research.

2. BACKGROUND TO THE STUDY

2.1. Policy process

The public policy process begins with identification of social problems which lead to policy agenda setting. Policy goals and objectives are formulated to address the identified social challenges [9, 25]. Legislations and regulations are then enacted to support the implementation of formulated policies. Implementation involves administrative, legal and regulatory arrangements where the actions are taken to achieve the policy goals [10]. The policy objectives are examined to establish their effectiveness, identify new needs and establish course of actions to address the policy process challenges [9]. The phases in policy cycle are shown in Figure 1.

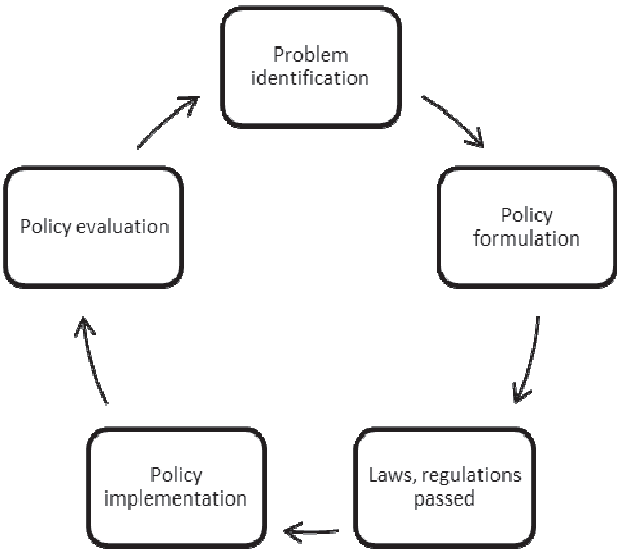


Figure 1. The policy cycle [8]

The policy process is iterative and may vary based of contexts [15]. Although the process may be perceived as having a structured sequences of carrying out policy activities that can be easily followed, in practice there are many challenges that may lead to failure to achieve the intended policy outcomes [25].

2.2. National ICT policy

National ICT policy is described as "*any public-sector action taken to advance the development of ICT or to promote their use by constituents for the benefit of society*" [15: 36]. National ICT policies are characterised as being dynamic, wide in scope and future oriented [15, 33]. The benefits of national ICT policies may be categorised into social and economic benefits [4, 15]. The social benefits include improved communication leading to more inclusion of communities, supporting the production and use of applications that lead to improved well-being of communities such as ICT public access interventions, health services support systems and commodity brokerage systems [4, 33]. The economic benefits include promoting ICT infrastructure development, supporting development of competitive telecommunication sector and enhancing effective operations of sectors such as finance and manufacturing. ICT sector as segment in an economy may contribute to economic growth of a country through trade and investment [19, 22]. Nonetheless, some of the African countries are yet to attain the benefits for national ICT policies. It is, therefore, critical to understand the problems impeding successful outcomes [14, 21].

2.3. Use of Grounded Theory in literature review

Grounded Theory is an inductive technique of analysing data with the aim of generating concepts, themes or theory that is grounded in data [41]. It can be used as method or approach to systematically investigate a social phenomenon and develop a theory. As a method, it has a set of principles and techniques such as principles of emergence, constant comparative analysis and theoretical sampling [22]. The approach employs four coding procedures in analysis of data to develop concepts, themes and a theory. The four systematic and iterative procedures are open, axial, selective and theoretical coding [11, 21]. These are summarized as follows:

- ***Open coding:*** Identifying concepts, categories and their properties in a corpus [11].
- ***Axial coding:*** Establishing relationships of identified categories and sub-categories focusing on answering "*who, where, when, why and how about the categories*" [41 cited by 35:3].
- ***Selective coding:*** Relating variable to core variables of categories [18].
- ***Theoretical coding:*** Integrating the codes into the emerging theory and producing descriptions or set of propositions [21, 22].

Grounded Theory can be used to review literature following five steps summarized as follows [47]:

1. Definition of a criterion for inclusion and exclusion of articles to be reviewed
2. Searching for the articles
3. Selection of articles fitting the set criteria.
4. Analysis of the articles using open, axial, selecting and theoretical coding
5. Presentation of the results

3. METHODOLOGY

The study employed qualitative approach using Grounded Theory method. The criterion for inclusion and exclusion of data for the study was studies that focused on African countries. The sample consisted of secondary data drawn from articles sourced from Information Systems centric publications. Other disciplines related to Information Systems such as Information Science, Telecommunications

Policy studies and Library studies were included in the corpus to improve the diversity of the views on the topic. Electronic databases were searched for peer-reviewed journal articles, book chapters and conference proceedings. The terms that were used in searching of articles were "ICT policy", "IT policy", "national ICT policy", "national IT policy", "e-strategies" and "national e-strategies". The results of the search yielded 103 articles which were trimmed to 30 articles meeting the criteria for context of Africa. Open coding process generated 141 concepts which were grouped to 29 sub-categories together with their dimensions. During axial coding the sub-categories were grouped to seven categories. The categories were refined during the selective and theoretical coding phases. The process for analysis was iterative. Memo writing technique was also used throughout the coding process. The presentation phase focused on reporting the findings.

4. RESULTS

The first part of the results of data analysis summarises the context of the study. The second part highlights the challenges for national ICT policy in the relation to the phases in the policy process and other challenges.

4.1. Context of the study

Table 1. Summary of context of study

Country	Human development category	Methodology of studies & Citation
Africa in general	-	Conceptual study [20, 39, 43, 46] and case study [14, 27]
Cameroon	Low	Document analysis [50]
Congo (DRC)	Low	Document analysis [50]
Cote d'Ivoire	Low	Document analysis [50]
Egypt	Medium	Document analysis and interviews [37]
Ethiopia	Low	Document analysis [50]
Ghana	Low	Conceptual study [48] and survey [29, 30]
Kenya	Low	Document analysis [50] and case study [28, 38]
Madagascar	Low	Document analysis [50]
Malawi	Low	Case study [26]
Mozambique	Low	Document analysis [33, 51]
Nigeria	Low	Document analysis [50], conceptual study [1, 5, 17, 49] and survey [2, 23]
Senegal	Low	Document analysis [32, 50]
South Africa	Medium	Conceptual study [9, 19], case study [34] and document analysis [6, 9, 33, 40, 44, 50]
Tanzania	Low	Document analysis [50] and case study [38, 42]
Uganda	Low	Case study [38], interviews and survey [31] and document analysis [33]
Zambia	Low	Document analysis [51]
Zimbabwe	Low	Document analysis [50, 51]

As illustrated in Table 1, the sample consisted of studies conducted in African countries with differences in human development and were classifies as very high, high, middle income and low income economies [57].The corpus of the study covered 17 countries and six articles were not country-specific and hence were considered as covering Africa in general. The methodologies employed in the papers included conceptual studies, case studies, surveys, documentary reviews and interviews.

The majority of the countries were in the category for low human development except for Egypt and South Africa which were in the medium human development category. The challenges for the low human development countries were similar to those of the medium human development particularly in constraints for participation in policy process and power struggles among policy actors [34, 37].

4.2. Challenges for national ICT policy

The challenges for the national ICT policy were categorised into seven: agenda setting, policy formulation, policy implementation, policy monitoring, policy evaluation, legal framework and power imbalance challenges. The challenges can be grouped into two: those directly related to each phase of the policy cycle and others challenges.

4.2.1. Agenda setting

For a number of cases, the process for identifying social problems and setting priorities on goals to address the challenges was problematic [1, 29, 42]. Some of the constraints were absence of information for making decisions, lack of consideration of local context and limited participation of stakeholders [1, 37]. The challenges arose from lack of capacity for research which might have generated information that could be used to support decision-making on policies [5, 28] as well as from technical determinism where policies where developed with more focus on technology and less focus on social factors [43, 48]. Consequently, policies did not fully address the problems of all policy stakeholders [1, 29]. Table 2 summarises the category of agenda setting.

Table 2. Challenges for agenda setting

Properties	Dimensions	Causes
Information for decision making	Available or not available	Limited research and capabilities
Decision challenges on policy priorities	Simple or complex	Lack of negotiating skills for stakeholders
Needs driven policy goals	Inclusive and exclusive	Technical determinism
Local context policies	Narrow or wide	Lack of knowledge transfer and learning
Participation of stakeholders	Formal or informal	Lack of leadership skills
Unrealistic policy objectives	Adequate or inadequate	Lack of skills for local stakeholders

4.2.2. Policy formulation problems

The activities for formulation policy were also beset with challenges which included limited comprehensiveness of policies, lack of skills in designing of policies especially among local stakeholders, lack of leadership to support formulation of realistic policy goals and problems of conflict of interest among the policy stakeholders [1, 13, 34]. Table 2 summarises the category for policy formulation challenges.

Table 3. Challenges for policy formulation

Properties	Dimensions	Causes
Comprehensiveness of policies	Narrow focus or wide focus	Limited participation of stakeholders
Stakeholders participation	Inclusive or exclusion	Top-down approach top policy process
Capabilities	Skilled or unskilled	Lack of knowledge transfer and learning
Implementation strategies	Included or omitted	Limited participation of stakeholders
Balance of power	Centralised or decentralised	Conflicts of interest and priorities
Leadership	Competent or incompetent	Lack of skills and expertise
Technical determinism	Narrow or holistic	Lack of understanding of local context

For Tanzania [42] and Nigeria [49] the national ICT policies were formulated using a top-down approach where politicians and government officials dominated the process and other policy stakeholders did not influence the policy decisions. Lack of policy design skills among local stakeholders also affected the formulation of policies. It was reported that national ICT policies for Uganda [31] and Nigeria [1, 17] were formulated without consideration of other government plans and strategies. This omission might have been attributed to limited skills in design of national ICT policies. As a result, the policy goals and strategies were narrow in scope and without plans for implementation to address the social problems.

4.2.3. Legal frameworks challenges

Legal frameworks enable institutions and organisations to regulate and enforce laws related to national ICT policy [11]. The analysis showed that the activities for establishing legal and regulatory frameworks that could support national ICT policies were also challenging [1, 4, 17, 26]. Legislators, politicians and government officials had a leading role in this process. Lack of political will among the legislators contributed to absence of laws and regulations for supporting the activities of institutions and implementation agents [1, 40]. Another challenge was the design of policies which did not include the strategies on implementation of legal frameworks to support policy goals [17, 26]. This was evident in cases of Kenya [28], Malawi [26], Tanzania [42], Nigeria [1] and South Africa [6, 40, 44]. As a result, institutions and implementation agents were not able to effectively carry out policy strategies and lacked the mandate to enforce regulations and laws [28]. Table 4 summarises the category for legal frameworks challenges.

Table 4. Legal frameworks challenges

Properties	Dimensions	Causes
Strategies for legal frameworks	Absent or present	Limited skills in development of policies
Laws and regulatory support	Limited or adequate	Incomprehensiveness of policy
Political will	Low or high	Lack of leadership for legislators

4.2.4. Policy implementation challenges

The analysis showed that the execution of policy strategies was also beset with challenges. Some of the constraints were lack of capabilities of policy actors, limited resources, poor coordination of activities, resistance to change among policy actors in the implementation institutions, limited delegation of responsibilities, and lack of mandate to enforce laws and regulations [5, 18, 20, 54]. Table 5 summarises implementation challenges.

Table 5. Policy implementation challenges

Properties	Dimensions	Causes
Capabilities of actors	Limited or adequate	Lack of knowledge transfer and learning
Participation of stakeholders	Inclusive or exclusion	Lack of leadership
Delegation of roles	Formal or informal	Lack of resources
Scope of activities	Narrow or broad	Budget constraints
Legal framework support	Limited or adequate	Absence of laws and regulations
Transparent process	More or less	Lack of accountability and transparency
Coordination of activities	Collaborative or isolated	Lack of stakeholders involvement
Resources	Limited or adequate	Over dependence on donors
Conflicts among stakeholders	Agreements or disagreements	Limited negotiating skills
Resistance to change	Flexible or rigid	Limited skills for implementing actors

Policy actors lacked the necessary skills to implement the policies partly because most of policies are designed by foreign experts with limited knowledge of local context [19, 50]. Lack of leadership also affected the coordination of implementation of activities [33, 34]. Limited participation of stakeholders also hindered progress in the implementation of policies [28, 29, 30]. In cases where there was participation, conflicts among stakeholders affected the implementation activities [13, 28].

4.2.5. Policy evaluation challenges

The studies noted constraints in the assessment of national ICT policy outcomes and the impact of policy activities. The challenges included poor coordination of activities, lack of feedback and poor design of policies. The cases of Senegal, Mozambique, Uganda and South Africa reported limitations of participation of stakeholders which made it difficult to assess the outcomes of the policy goals [34, 35, 40]. Lack of skills also affected the design of policies where policy evaluation did not include processes and description of mechanisms for assessing consequences for policies [23, 42]. Table 6 summarises policy evaluation challenges.

Table 6. Policy evaluation challenges

Properties	Dimensions	Causes
Capabilities of actors	Limited or adequate	Lack of knowledge transfer and learning
Participation of stakeholders	Inclusive or exclusion	Lack of leadership
Delegation of roles	Formal or informal	Lack of resources
Scope of activities	Narrow or broad	Budget constraints
Legal framework support	Limited or adequate	Absence of laws and regulations
Transparent process	More or less	Lack of accountability and transparency
Coordination of activities	Collaborative or isolated	Lack of stakeholders involvement
Resources	Limited or adequate	Over dependence on donors
Conflicts among stakeholders	Agreements or disagreements	Limited negotiating skills
Resistance to change	Flexible or rigid	Limited skills for implementing actors

4.2.6. Policy monitoring challenges

The studies conducted in Nigeria [1] and South Africa [6, 35], reported that the process for observing the activities for plans and strategies in the policy process were limited and in some cases absent. This was partly due to lack of resources and expertise in assessing progress on strategies and activities [1, 6, 34]. Implementing institutions lacked financial and human resources to effectively monitor the progress of policy activities [26, 50]. Table 7 summarises the category for policy monitoring challenges.

Table 7. Policy monitoring challenges

Properties	Dimensions	Causes
Activities	Structured or unstructured	Lack of resources and skills
Strategies	Present or absent	Poor design of policies
Resources	Limited or adequate	Budget constraints

4.2.7. Power asymmetry

Authority and control over decision and activities for national ICT policy in essential to ensure that policy goals are achieved [33]. This process requires a balance of power between the stakeholders with resources and those without resources to influence decisions and interests on policy goals [1]. The studies reported challenges for power relations among institutions and organisations involved in the national ICT policy process such as government departments and private sectors firms. Table 8 summarises power imbalance challenges.

Table 8. Power imbalance challenges

Properties	Dimensions	Causes
Control of policy activities	Gain or loose	Limited autonomy and control over policies
Domination	Inclusion or exclusion	Limited participation of some stakeholders i.e. women
Stakeholders contributions	Participation or involvement	Poor coordination in policy implementation in policy process
Exploitation	Beneficial or loss	Lack of resources and leadership skills

The problems were control over the activities of policy process, domination of some stakeholders leading to limited contribution of ideas and decisions on policy process [33, 51]. Consequently, there was loss of control over policy activities especially for government departments to more powerful institutions which have resource (human, financial and political) [1]. Another challenge was consideration of stakeholder's contributions to policy goals and objectives. The cases of Senegal, Cameroon, Ethiopia and Zimbabwe reported limited consideration of gender in policies [54]. Similarly, the cases of Zambia and Mozambique also reported the limitations of gender consideration in the policies [51]. This was attributed to domination of more powerful stakeholders where their ideas dominated the agendas of the policy and lack of participation of women in the national ICT policy process [14, 34].

5. DISCUSSION AND CONCLUSION

The key question guiding the research was on the challenges or barriers affecting national ICT policy process and how they can addressed. The analysis showed cross-cutting challenges in the different phases of policy process. The challenges were lack of resources, limited skills designing policies, lack of information to support policy decisions, lack of understanding needs, lack of legal frameworks to support policy implementation activities, limited participation of stakeholders and power imbalance among stakeholders. The challenges were categorised in relation to

the phases for the policy process. It was also noted that the effect of challenges affecting one phase had implications in the subsequent phases. While the corpus was restricted to studies focusing on Africa the results were consistent with similar studies conducted in other parts of the world such as Asia Pacific [10], Greece [12] and Malaysia [3]. This may be attributed to similarities in the policy process since introduction of national ICT polices was influenced by the international development agencies such as United Nations Economic Commission for Africa and African Information Society Initiative [15]. These organisations provided national ICT policy development frameworks and technical support to African countries.

The study suggests enhancement of participation of stakeholders in the process of so that national ICT policies are needs-driven, comprehensive and coordinated to mitigate some of the challenges [20]. National policies should be integrated with other development policies. Further, the study recommends monitoring and evaluation to be integrated in all the phases the policy process [8]. Participation of stakeholders in the policy process may also support feedback on progress of policy activities and learning from the success or failure of policy process activities. Capacity building should be incorporated in the policy activities [25].

The study recognised the challenges for the size of corpus and use of secondary data which may not fully reflect the realities on the ground. Nevertheless, the current study provides a useful basis for further research in validating the concepts and categories identified in the policy process and use of initial policy declarations for data analysis. For instance, the issue of power among stakeholders in national ICT policy and how the problem may be mitigated [13]. The national ICT policy process may be viewed from local, regional, national and global perspectives because of the diversity of stakeholders in the policy process [14, 32] Insights from such a study many be useful for policy makers and research community.

REFERENCES

[1] M. Adeyeye and C. Iweha, "Towards an effective national policy on Information and Communication Technologies for Nigeria," *Information Development*, Vol. 21, pp. 202-210, 2005.

[2] P. Akpan-Obong, "Unintended outcomes in Information and Communication Technology adoption: A micro-level analysis of usage in context" *Journal of Asian and African Studies*, vol. 45 no. 2, pp.181-195, 2010.

[3] R. Alinaghian, A. Rahman, and R. Ibrahim, "Information and Communication Technology (ICT) Policy: Significances, challenges, issues and future research framework," *Information Journal of Basic and Applied Sciences*, vol. 5 no. 12, pp. 963-969, 2011.

[4] S. Anie, "The economic and social benefits of ICT policies in Nigeria," Library Philosophy and Practice, Paper 457, 2011.

[5] E. Baro, "A critical examination of Information and Communication Technology policies: Effects on Library services in Nigeria," *Library Philosophy and Practice*, Paper 464, 2011.

[6] C. Benner, "Digital development and disruption in South Africa: Balancing growth and equity in National ICT policies," *Perspectives on Global Development and Technology*, vol. 2 no. 1, pp.1-26, 2003.

[7] P. Bridgeman and G. Davis, "What use is a policy cycle? Plenty, if the Aim is clear," *Australian Journal of Public Administration*, vol. 62 no. 3, pp. 98 -102, 2003.

[8] S. Brooks, "Public Policy in Canada: An introduction," *Oxford University Press: Toronto*, 1998.

[9] W. Brown and I. Brown, "The next generation of ICT policy in South Africa: Towards human development based ICT policy," In (eds.) Avgerou, C., Smith, M., & van den Besselaar, P. *Social Dimensions of Information and Communication Technology Policy*; IFIP, Volume 282. Boston: Springer, pp. 109–123, 2008.

[10] J. Chacko, "Paradise lost? Reinstating the human development agenda in ICT policies and strategies," *Information Technology for Development*, vol. 11 no.1, pp. 97-99, 2005.

[11] H. Chen and J. Boore, "Using a synthesised technique for grounded theory in nursing research," *Journal for Clinical Nursing*, vol. 18, pp. 2251-2260, 2009.

[12] I. Chini, "The saga of ICT policy in Greece*," Proceedings of the Third Observatory PhD Symposium on contemporary Greece: Structures, contexts and challenges, London*, 14 -15 June, pp. 1-18, 2007.

[13] S. Chiumbu, "Understanding the role and influence of External Actors and Ideas in African Information, Communication and Technologies Policies: The African Information Society Initiative," *PhD Thesis, University of Oslo*, 2008.

[14] S. Chiumbu, "Elites and donors: Interrogating the African ICT agenda," In (eds.) K. Kondlo and C. Ejiogu, Africa in focus: Governance in the 21st Century. *Cape Town: HSRC Press,* 2009.

[15] G. Cohen, I. Salomo and P. Nijkam, "Information Communication Technologies (ICT) and transport: does knowledge underpin Policy?," *Telecommunication Policy*, vol. 26, pp. 31-52, 2002.

[16] P. Cronin, F. Ryan and M. Coughlan, "Undertaking a literature review: a step-by-step approach," *British Nursing*, vol.17 no.1, pp. 39-43, 2007.

[17] L. Diso, "Information Technology policy formulation in Nigeria: Answers without questions," *The International Information and Library Review*, vol. 37, pp. 295-302, 2005.

[18] Y. Eaves, "A synthesis technique for grounded theory data analysis," *Journal of Advanced Nursing*, vol. 35 no.5, pp. 654-663, 2001.

[19] A. Gillwald, "Good intentions, poor outcomes: Telecommunications reform in South Africa,". *Telecommunications Policy*, vol. 29, pp. 469-491, 2005.

[20] A. Gillwald, "The poverty of ICT, Research, and Practice in Africa," *Information Technologies and International Development*, vol. 6, pp. 79-88, 2010.

[21] B. Glaser, "Basics of Grounded Theory Analysis". Mill Valley: Sociology Press, 1992.

[22] B. Glaser and A. Strauss, "The Discovery of Grounded Theory: Strategies for Qualitative Research," *New York: Aldine*, 1967.

[23] O. Hassan, W. Siyanbola and T. Oyebisi, "An evaluation of implementation mechanism of Nigerias Information Technology Policy" *International Journal of Technology, Policy and Management*, vol. 11 no. 1, pp.57-67, 2011.

[24] ITU WSIS Stocktaking, "ITU Word Summit on Information Society 2012 Report,". *International Telecommunications Union,* Switzerland, 2012.
[25] W. Jann and K. Wegrich, "Theories of the policy cycle," In (eds.) F. Fischer, G. Miller and M. Sidney, *Handbook of Public Policy Analysis: Theory, Politics and Methods*. London: Taylor & Francis Group, 2007.
[26] C. Kanjo, "Going beyond diagnostics and planning in ICT initiatives: limitations in the context of Malawi," *Proceeding of CIRN 2008: ICTs for social inclusion: what is the reality? Prato, Itary,* 2008.
[27] K. Kendall, J. Kendall and M. Kah, "Formulating Information and Communication Technology (ICT) Policy through Discourse: How internet Discussions shape policies on ICT for Development," *Information Technology for Development,* vol. 12 no. 1, pp. 25-43, 2006.
[28] M. Kerrette, "ICT regulation and policy at crossroads: a case study of the licensing process in Kenya. Southern Africa," *Journal of Information and Communication*, vol. 5, pp. 49-63, 2004.
[29] O. Kwapong, "Problems of policy formulation and implementation: The case of ICT use in rural womens empowerment in Ghana*," International Journal of Education and development using Information and Communication Technology*, vol. 3 no. 2, pp. 68-88, 2007.
[30] O. Kwapong, "Policy implications for using ICT for empowerment of rural women in Ghana," *The Turkish Online Journal of Educational Technology*, Vol. 7 no. 3, pp.35-45, 2008.
[31] A. Madanda, D. Okello and G. Bantebye-Kyomuhendo, "A gender critique of Uganda's rural ICT access policy: opportunities and challenges," *International Journal of Computing and ICT research*, vol. 3 no.1, pp. 42-52, 2009.
[32] R. Mansell, "Communication, Information and ICT Policy: Towards enabling research frameworks," In (ed.), C. Avgerou, M. Smith and P. van den Basselaar, *Social Dimensions of Information and Communication Technology*, Boston: Springer, pp. 15-28, 2008.
[33] G. Marcelle, "Getting Gender into African ICT Policy: A strategic review," In Rathgeber, E., & Adera, E. (eds.), Gender and Information revolution in Africa, *IDRC, Canada*, 2000.
[34] M. Mashinini, "Challenges for rural ICT policy for rural communities: A case study from South Africa," In (eds.) Avgerou, C., Smith, M., & van den Besselaar, P. *Social Dimensions of Information and Communication Technology Policy*, IFIP, Volume 282. Boston: Springer, pp. 125–137, 2008.
[35] R. Matavire and I. Brown, "Profiling Grounded Theory approaches in Information Systems research," *European Journal of Information Systems*, pp.1-11, 2011.
[36] V. Mbarika, C. Okoli, T. Byrd and P. Datta, "The neglected continent of IS research: A research agenda for Sub-Saharan Africa," *Journal of the Association for Information Systems*, vol. 6 no. 5, pp. 130-170, 2005.
[37] N. McBride and B. Stahl, "Analysing a national information strategy: a critical approach," *International Journal of Intercultural Information Management*, vol. 2 no.3, pp. 232-262, 2010.
[38] M. Mureithi, "Evolution of Telecommunications Policy Reforms in East Africa: Setting New Policy Strategies to Anchor Benefits of Policy Reforms," *The Southern African Journal of Information and Communication*, pp.1-17, 2002.
[39] W. Olatokun, "Gender and national ICT policy in Africa: issues, strategies, and policy options," *Information Development*, vol. 24 no. 1, pp. 53-65, 2008.
[40] S. Singh, "The South African 'Information Society', 1994–2008: Problems with Policy, Legislation, Rhetoric and Implementation," *Journal of Southern African Studies*, vol. 36 no. 1, pp. 209-227, 2010.
[41] A. Strauss and J. Corbin, "The Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory," *London: Sage*, 1998.
[42] H. Twaakyondo, "Key issues in Information Communication Policy review process: The case of Tanzania," *International Journal of Computing and ICT Research*, vol. 5 no. 2, pp. 46-58, 2011.
[43] L. van Audenhove, "The African Information Society: rhetoric and practice*," Communicatio: South African Journal for Communication Theory and Research*, vol. 24 no.1, pp. 76-84, 1998.
[44] L. van Audenhove, "Towards an integrated information society policy in South Africa: an overview of political rhetoric and policy initiatives 1994–2000," *Communicatio: South African Journal for Communication Theory and Research,* vol. 29 no. 1-2, pp. 129-147, 2003.
[45] J. Webster and R.Watson, "Analyzing the past to prepare for the future: Writing a literature review," *MIS Quarterly*, vol. 26 no.2, pp. xxi-xxii, 2002.
[46] E. Wilson and K. Wong, "African Information Revolution: a balance sheet," *Telecommunications Policy*, vol. 23, pp. 155-177, 2003.
[47] J. Wolfswinkel, E. Furtmueller and C. Wilderom, "Using grounded theory as a method for rigorously reviewing literature," *European Journal of Information Systems*, pp.1-11, 2011.
[48] A. Yadana and M. Amppiah, "The Role of Information Communications Technology (ICT) in National Development: The challenges for our society," *Mathematics Connection,* vol. 3, pp. 35-43, 2003.
[49] M. Yusuf, "Information and communication technology and education: Analysing the Nigerian national policy for information technology," *International Educational Journal*, vol. 6 no. 3, pp. 316-321, 2005.
[50] P. Browne, "Study of effectiveness of informatics policy instruments in Africa," IDRC, Canada, 1996.
[51] P. Zirima, "Engendering ICT policies: Practices from Mozambique, Zambia and Zimbabwe," *Agenda*, vol. 21 no. 71, pp. 70-79, 2007.
[52] R. Checchi, J. Hsieh and D. Straub, "Public IT Policies in less developed countries: A critical assessment of the literature and a reference framework," *Journal of Global Information Technology Management*, vol. 6 no. 4, pp. 45-64, 2003.
[53] Human Development Report, "*Sustainable and Equity; A better future for all*", UNDP, New York, 2011.

THE ROLE OF INTELLIGENT TRANSPORTATION SYSTEMS IN DEVELOPING COUNTRIES AND IMPORTANCE OF STANDARDIZATION

Muzaffar Djalalov

Scientific Engineering and Marketing Researches Center – SUE "UNICON.UZ", Uzbekistan
djalalov@unicon.uz

ABSTRACT

The traffic accidents and congestions are getting a serious problem all over the world. The problem is growing fastest in developing countries where urbanization and the use of motorized vehicles is increasing most rapidly. One alternative solution is a concept of Intelligent Transportation Systems (ITS). It provides the ability to gather, organize, analyze, use, and share information about transportation systems. In this paper such issues as concept of sustainable communities with transportation and ITS is discussed and scheme of ITS deployment in developing countries and its benefits are presented. Moreover, the analysis of ITS current situation in some developed and developing countries and the role of standardization are presented.

Keywords—Intelligent Transportation Systems (ITS), Sustainable communities, Standardization

1. INTRODUCTION

Maintaining air quality was once viewed as a luxury of developed countries, which could more easily bear the cost of technology to keep emissions under control. However, the impact of poor air quality, especially on health and productivity, is now recognized as a large cost to all national economies, including developing and transitional economies.

However, traffic congestion is a serious problem in all parts of the world. The problem is growing fastest in developing countries where urbanization and the use of motorized vehicles is increasing most rapidly. Congestion causes delays and uncertainty, wastes fuel, results in greater air pollution, and produces a larger number of crashes [1-3].

Many newly developing countries are growing rapidly. One Such rapid growth in vehicles without a comparable growth in transportation infrastructure leads to increasing traffic congestion. One solution to the growing congestion problem is to build more transportation infrastructure. It is cost prohibitive, however, to build new infrastructure in built-up areas. So developing countries around the world are looking for other alternatives to deal with this problem. One such alternative is a set of practices called Intelligent Transportation Systems (ITS). ITS is commonly understood to denote systems that combine recent advances in information and communication technologies to better manage the transport system. ITS comprises a wide range of tools for managing transport networks as well as for providing services to travelers. One of the basic features of ITS is the collection of data and conversion of the data into information that can be used to fulfill a user need. Through ITS, transportation authorities, operators, and individual travelers are able to make more coordinated and intelligent decisions based on timely information.

Besides sustainable development and sustainable communities, "sustainable transportation" forms a third conceptually important link between ITS and sustainable communities. And defining sustainable transportation, in turn, begins with the understanding that transportation is a complex system that integrates people, modes, and land uses. Despite widespread understanding of such system attributes, most conceptual approaches to transportation still reflect separate modes (i.e., driving as separate from bicycling, transit, or walking).

Figure 1 provides a systems framework for linking the concept of sustainable communities with transportation and ITS [1, 15].

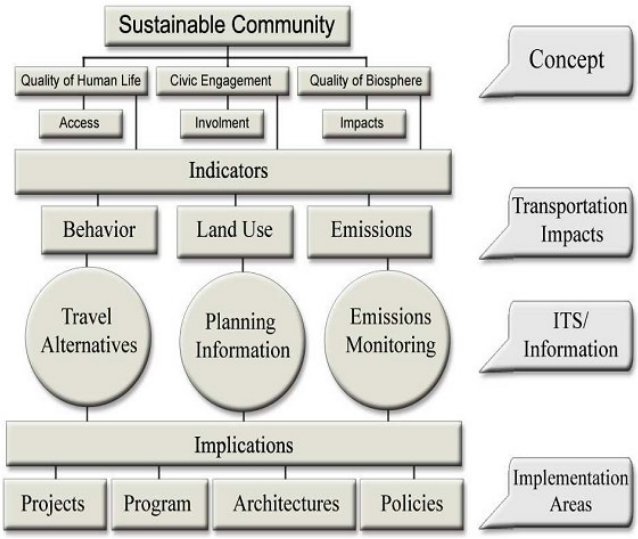


Figure 1. Sustainable communities and ITS

2. ITS AROUND THE WORLD

The world regions which led the introduction of ITS – Europe, the U.S., and Japan – use approaches that have many features in common, including:

- An interest in pursuing advanced technology and applying it to social and economic problems;
- A desire to expand the capabilities of the transportation system in a well-integrated manner;
- A strong desire to expand existing markets and open new ones;
- A belief that the best results are produced through the cooperative efforts of industry, government, and academia.

However, each world region also has its own individual approach to introducing ITS – its own "ITS Culture" [4-6].

2.1. Europe

Europe has a long tradition of applying technology, including computers and communications, to broad social issues including safety and mobility. As Europe continues to move toward a common continental economy, ITS is playing an important role in lowering barriers for the movement of people and freight throughout Europe. ITS is regarded both as a transportation tool and as part of Europe's Information Society [7, 12].

Europe has been able to introduce transport control technology to advance social goals. This includes using technology to limit the speed at which trucks can travel and "intelligent speed adaptation" which advises vehicles of safe speed limits and has the capability to limit driving speed. Europe is taking a very aggressive approach to traffic safety, with the objective to halve the number of traffic fatalities by 2010 and aim for "zero traffic fatalities" by 2020. This is part of a public-private sector initiative called "eSafety" that is being led by the EU's Information Society Directorate. eSafety's objective is to improve road safety by using Intelligent Vehicle Safety Systems, and it has established a timetable for the Europe-wide adopt ion of in-vehicle systems like antilock brakes, electronic stability control, automatic crash notification, etc.

Europe has been very successful in establishing partnerships to conduct tests and demonstrations among European national and city governments, vehicle manufacturers and suppliers, and universities. Europe has also succeeded in establishing a robust industry for ITS infrastructure systems and end-user products, with customers worldwide.

Europe is attempting to reform the way it charges for the use of the road infrastructure, although this process is still regarded as very controversial politically. An EC directive called Eurovignette prescribed moving toward a road-charging system, starting with heavy trucks, based on vehicle weight, distance traveled, and other criteria. This has been experimentally introduced in Germany and Switzerland, and there are plans to introduce it in the UK in 2006. Although huge interest has been expressed for these schemes, there are continuing arguments about their consistency and fairness.

Like other parts of the world, Europe has had little success in introducing telematics (wireless delivery of information and services to vehicles). Several attempts to develop telematics services by vehicle manufacturers and wireless carriers have been unsuccessful. General Motors' OnStar service is attempting to enter the European market, but is far from being prof i table. A few companies in the UK (Trafficmaster and ITIS) and France (MediaMobile) are delivering rudimentary real-time traffic information.

2.2. USA

Safety is also an important issue in the U.S., although it is pursued less aggressively than in Europe. The U.S. Dept. of Transportation sponsored an Intelligent Vehicle Initiative to test and demonstrate in-vehicle technology to enhance driving safety. This program has now concluded, but important portions are continuing, namely the development of intersection collision avoidance systems and integrated vehicle safety systems [8, 9].

The deployment of safety products in the U.S. has been hindered by concerns about product liability lawsuits and, in general, the domestic ITS industry is less robust in the U.S. than in Europe or Japan, especially for ITS in-vehicle and consumer products.

There is a program in the U.S. called "Vehicle-Infrastructure Integration." The objective of this program is to create an integrated, intercommunicating surface transportation system. The system will use wireless communications, primarily DSRC (Dedicated Short Range Communication) to link the infrastructure and its managers with vehicles and their drivers. It will gather and share information about the transportation system to help improve the performance of the infrastructure, vehicles, and drivers.

Another program that is becoming wide-spread in the U.S. is called "511". The digits 511 have been reserved as a nationwide telephone number for obtaining traveler information.

Electronic toll collection (ETC) is also becoming widespread in the U.S. as a means to reduce delays at toll barriers and lower the cost of collecting tolls.

2.3. Japan

Japan has been very successful in translating its strengths in electronics technology into successful ITS. The most prominent ITS programs in Japan are the widespread adoption of car navigation systems and the nationwide deployment of the Vehicle Information and Communication System (VICS), which provides real-time traffic information to vehicles. Japan's complex and congested road system has made these technologies particularly attractive to the driving public. In addition, Japanese consumers have traditionally been early adopters of new technology-based products and services [13].

Japan's emphasis on ETC deployment has mainly been to reduce congestion at toll barriers, with less emphasis on improving the efficiency and reliability of collection.

Dedicated Short-Range Communications (DSRC), which is used for ETC, is also being deployed for use with VICS. The intention is to use this communications infrastructure as a basis for multiple other ITS applications.

2.4. Developing countries

In East Asia: Traffic information services have become common using multiple broadcasting and communications media [1, 3, 10, 15].

In Eastern Europe: Road management systems have been introduced to identify road surface conditions, reflecting an emphasis on improving infrastructure maintenance. In addition, the trading of "empty cargo space" has become common, to improve the efficiency of freight logistics.

In Latin America: Border-crossing systems have been introduced as a result of the regional emphasis on promoting cross-border trade to increase the economic strength of the region.

All three regions have introduced basic systems to manage road traffic. These include traffic signal systems, traffic surveillance systems using CCTV, and traveler information systems based on variable message signs (VMS). As expected, systems that provide a high rate of return on investment have the greatest likelihood of being introduced. These include electronic toll collection and fare payment systems, commercial vehicle tracking systems, and bus management systems.

In Central Asia, as well as also in Uzbekistan, car number plate recognition system has successfully been used by customs services to automatically register a vehicle that enters/leaves the country.

A very popular ITS application in developed countries is in-vehicle navigation. An in-vehicle navigation system calculates and delivers driving directions to a destination stated by the driver. In-vehicle navigation systems include a map database, location sensors, a computer, and a user interface (e.g., a touch screen). The user interface lets the driver specify a destination and lets the system deliver directions. Navigation systems can generate efficient routes and help drivers keep from getting lost . In the future, navigation systems will receive real-time traffic information and adapt routes dynamically based on current conditions. As the cost of navigation systems continues to come down, it is expected that these systems will start to appear in developing countries as well.

World Bank Reports provide an overview of different categories of ITS applications' deployment in three major developing regions including East Asia, Latin America and Eastern Europe (Table 1). Traffic information traffic management, commercial vehicle operations, and public transportation management are most widely used applications in the selected regions. The applications can be regarded as a manipulated version of commonly known applications in the industrialized world such as Advanced Traffic Information System (ATIS), Advanced Traffic Management System (ATMS) and Commercial Vehicle Operation (AVO).

Table 1. ITS Application in the three selected regions of the world.

<table>
<tr><th>Scope of application</th><th>Application</th><th>East Asia</th><th>Eastern Europe</th><th>Latin America</th></tr>
<tr><td rowspan="5">Common Applications</td><td>Travel Information</td><td colspan="3">Traffic information systems using roadside veriable message signs</td></tr>
<tr><td>Traffic Management</td><td colspan="3">Traffic signal systems,
Traffic surveillance systems using CCTV</td></tr>
<tr><td>Commercial Vehicle Operation</td><td colspan="3">Commercial vehicle tracking systems</td></tr>
<tr><td rowspan="2">Public Transport Management</td><td></td><td>Bus operation management systems</td><td></td></tr>
<tr><td>- IC card systems,
-Electronic ticketing</td><td></td><td>- IC card systems,
- Electronic ticketing</td></tr>
<tr><td colspan="2">Region-Specific Applications</td><td>Traffic information services utilizing multiple media</td><td>Road management systems "Empty cargo space" trading systems</td><td>Border crossing systems</td></tr>
</table>

3. ITS FOR SUSTAINABLE COMMUNITIES IN DEVELOPING COUNTRIES

The link between ITS and sustainable communities stems from the ability of ITS to create a transportation system rich in information, or what might be called an information-intensive transportation system. An information-intensive transportation system promotes sustainable communities in two ways. First, it substitutes information for new lanes, roads, and highways as a way of increasing the capacity of the transportation system. In doing so, ITS substitutes "information for stuff" (14); in this case, the "stuff" being replaced with information is the material resources necessary to build roads. In addition to using fewer resources, substituting better information for new roads may also conserve open space, decrease the noise and community disruption related to new highways, and reduce the damage to biodiversity. ITS thus supports a fundamental tenet of sustainability: that the earth's resource base has limits, that some of those limits are being approached, and therefore sustainable development depends on accommodating economic growth while consuming fewer resources.

Developing countries are often at a disadvantage, relative to developed countries, in constructing the basic infrastructure that provides the foundation for building their economies and societies. This is largely due to the limited financial, technical, and engineering resources that developing countries have available. However, developing countries also have some advantages relative to developed countries, particularly when the infrastructure to be constructed has high IT content.

Developing countries, more often than developed countries, are able to install electronic infrastructure at the same time that physical infrastructure is being constructed. This is far less expensive than retrofitting existing physical infrastructure. Developing countries are also not generally burdened with an outdated in-place IT infrastructure that has to be updated. Developing countries benefit from the continuing rapid decrease in the cost of IT. Building a new IT infrastructure from scratch is often less expensive than updating an existing system. Developing countries can make immediate use of other systems like cellular telephones and the Internet that are spreading rapidly in parallel. Finally, developing countries can take advantage of IT and ITS products and applications which have already been tested and deployed in developed countries and which are now mature, stable, well understood, and starting to become less expensive to acquire and operate [12, 15].

As a result, developing countries can frequently leapfrog directly to an ITS-enabled transportation infrastructure far more rapidly and far less expensively than developed countries (figure 2).

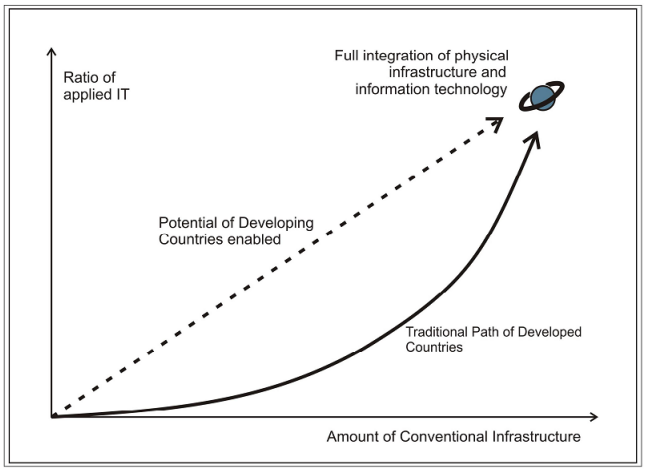


Figure 2. Leapfrogging with ITS

Deploying ITS in developing countries has many direct and indirect benefits for travelers, shippers, transportation system operators, and the country as a whole. However, introducing ITS is a complex undertaking, and decision makers have to understand the prerequisites for deploying ITS before their decisions are made. These prerequisites are both institutional and technological. Planners and developers in developing countries must have both the traditional and the expanded skills and knowledge for the country to succeed in leapfrogging to an ITS-enabled transportation infrastructure,. Having both sets of skills and knowledge makes it possible for planners and developers to understand the interact ion between traditional and ITS-enabled transportation and to develop them in parallel, methodically and systematically.

The concept of affordable ITS means encouraging decision makers in developing countries to focus first on the ITS applications that (1) can be deployed immediately or in the near future and that (2) can provide the greatest return on investment, in terms of lives and money saved and improved services. Affordable ITS applications will generally have the following characteristics:

- Deployment of an application can proceed in parallel and in cooperation with the development of other road infrastructure and public transportation systems.
- Deployment can make good use of the spread of the Internet and cellular phones.
- Applications are flexible enough to cope with rapid urban development and growth.
- The cost of deployment is moderate, functions are basic and simple, and maintenance is easy.
- Systems can incorporate human work where appropriate.
- Developers can make use of the ITS experiences, architectures, and applications of industrialized countries.

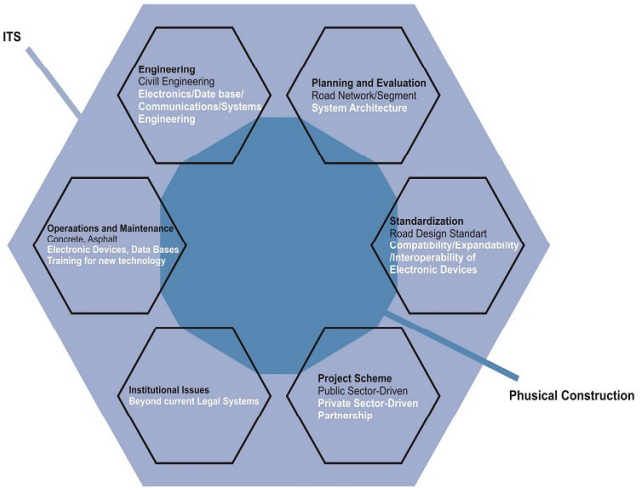


Figure 3. Scheme of ITS deployment in developing countries

In both developed and developing countries, innovative strategies for the deployment of transportation infrastructure have often taken the form of public-private partnerships. One common model for public-private cooperation in infrastructure development and operation is called BOT (for Build, Operate, Transfer). In this model, private companies invest in the construction of infrastructure and, with public sector support, own and operate the infrastructure and collect tolls or usage fees. Once the investment has been recouped, the facility is transferred to the public sector for continuing operation.

In the global economy, freight moves across oceans and through countries and across borders in trucks and on trains, often changing carriers en route to its final destination. It is important to move freight promptly and efficiently. It is also important to keep track of the location and content of containers and vehicles as they travel and to safeguard sensitive or hazardous cargo.

ITS includes many approaches to enhance the mobility of people and freight in al l transportation modes. Traveler Information helps travelers avoid congestion and can help improve traffic conditions. Traffic Management, e.g. the more effective timing of traffic signals, can help increase traffic efficiency. Demand Management, e.g. road and access pricing can help relieve heavily congested urban areas. Commercial Vehicle Management helps to increase security and efficiency not only for carriers but also for related public agencies. There are many more examples as well.

For the case of Dar es Salaam city in Tanzania, Dar es Salaam Urban Transport Authority (DUTA) provides direction and coherence across all sectors so they operate harmoniously. The Institutional Development in Dar es Salaam Transport Authority has proposed to establish the Dar es Salaam Urban Transport Authority (DUTA) as an accountable and transparent authority responsible for the transport system development of Dar es Salaam city. The management of urban transport as a function is almost nonexistent in Dar es Salaam specifically due to the lack of guidance attention for responsible agencies and the number of responsible entities involved, with municipalities and also national authorities cross - cutting into urban transport affairs. This has resulted in a fragmented planning process and lack of coordination, vertically and horizontally between levels of government and departmental disciplines. As one of the strategic policies to meeting these challenges DUTA has been proposed and its conceptual design has to be presented.

4. STANDARDIZATION ISSUE

The requirement for future standards in the ITS field is to be able to provide multiple services, over multiple different platforms, that will work in different countries (as vehicles can easily cross borders), while maintaining a simple-to-use interface that requires minimum intervention from the driver.

However, systems in use in different parts of the world remain incompatible and fragmented, particularly for 5 GHz systems for which USA, Europe and Japan have different implementations using different frequencies.

In many developing countries as well as in large developed countries (like the U.S. and Australia), the transportation systems in various parts of the country often developed separately. As a result, the tools used to manage these transportation systems are often incompatible. This makes it harder and more expensive for regions to cooperate effectively with one another (e.g. , in a time of national emergency). Even within a region, it is often hard for different public agencies to cooperate and coordinate with each other due to differences in procedures and equipment (e.g., radios on different frequencies or using different protocols). The deployment of ITS standards may solve such gaps [11].

Europe takes a very active and aggressive role in regional and international ITS standards activities to advance its ITS technology in the world market. The European Committee for Standardization) has a technical committee (TC278 – Road Traffic and Transport Telematics) focused on ITS standards issues. TC278 works in close cooperation with the ITS Technical Committee (TC204) of ISO. Many ITS standards items are developed in parallel by ISO/TC204 and CEN/TC278.

Japan pursues international ITS standardization with a view to encouraging international competition and safeguarding Japan's competitive position.

Because developing countries, for the most part, have not yet invested heavily in ITS, there is a good chance that they will avoid the standards battles that currently reign in this area. Where compatible standards are in use, the unit price - to consumers as well as to manufacturers - will tend to fall. Thus CALM represents an important opportunity for developing countries to plan their ITS strategies around international standards. Although it may not be feasible for many developing countries to develop their own indigenous car manufacturing sectors, there is scope for them to break into the market for ITS, which tends to be dominated by small and medium-sized enterprises, and by services rather than devices. In addition, there is a genuine possibility for technological leapfrogging in this field because developing countries generally do not have legacy transport management systems.

Standards for electronic toll collection (ETC) and dedicated short-range communications (DSRC, used to communicate between vehicles and ETC systems) provide interesting examples of success and failure in the creation and use of standards in the developed world. Electronic toll collection, in general, has been one of the most widely adopted and economically successful applications of ITS in both developed and developing countries. One reason for its attractiveness and success, of course, is that it helps to generate revenue by eliminating delays at toll barriers and by generally reducing congestion on toll roads and bridges. It can also help reduce labor costs (more important in developed countries) and make collection more reliable (important in all countries).

The lesson, for developing countries who aim for nationally standardized ETC, is to focus first on the institutional and administrative aspects of ETC, and the systems needed for a collective operation. Once these can be harmonized, selecting a common technology for ETC communications is relatively easy.

5. CONCLUSIONS AND RECOMMENDATIONS

For ITS to be successfully introduced in a developing country, a number of institutional prerequisites must be met.

Some of these are common to any large public project. Some are specific to ITS:

- An ITS promotion organization is extremely helpful, like ITS America, ITS Japan, and ERTICO/ITS Europe. Many developing countries in Europe, Asia, and Latin America have ITS promotion organizations. These organizations can help form public-private partnerships and introduce and promote the concept of ITS to the public.
- Capital for investment must be secured.
- ITS needs to be coordinated with existing laws and regulations; and, in some cases, new laws, regulations, and institutions may need to be created.
- New procurement rules are required to purchase software and electronic devices which are different from the rules for procuring infrastructure development.
- Provision must be made for training human resources to develop and administer ITS.
- The viewpoints of consumers and other users need to be understood and incorporated into ITS deployment.

Standards are also being developed for many other aspects of ITS to help provide consistency, enlarge markets, promote competition, and enhance interoperability. Developing countries should look primarily to international standards programs as sources of ITS standards to adopt. Many developing countries are active participants in ISO/TC204, where most international ITS standards are developed. Some countries can send experts to participate in the drafting of standards documents. Other countries participate by reviewing and voting on draft standards. Some countries act simply as observers. A domestic standards oversight group to coordinate standards participation and adoption is very helpful.

A step-by-step architecture approach is recommendable for developing countries. This step-by-step approach can take either of two directions. The first is to adapt the architecture of another region rather than creating an original architecture from scratch. An architecture that suits a region or country can be created by picking the necessary services and modules from an existing comprehensive architecture. Alternatively, a region or country can start by developing a simple architecture, reviewing it from time to time, and allowing it to develop step by step. The base architecture should be chosen with the aim of future interoperability with surrounding regions/countries.

Given the broad scope of the CALM project, which interfaces with virtually all commercially-available wireless interfaces, it is likely that it will have much wider ramifications for ITU's work notably touching upon ITU-T Study Groups 13 (NGN) and 16 (Multimedia), Study Group 12 (Performance and QoS) as well as ETSI TC ITS, car2car communication consortium.

There are a number of avenues through which private industry and public agencies in developing countries can help to ensure that appropriate standards are available and in place in their countries, both for ITS and for many other areas of interest. These include:

- Adopting national and/or international standards developed by others
- Participating in international standards development
- Par t icipat ing in relevant industry-oriented SDOs and consortia

All of these approaches (but especially the first one) will benefit from the existence of a national organization in a developing country that coordinates the adoption and use of standards. Usually, the logical candidate is the organization that serves as the country's national member body at ITU or ISO. In addition, an ITS-oriented industry association or ITS promotion organization in a developing country can serve as a valuable advisor to the national standards organization in the area of ITS standards.

So the answer to the question posed in the title is that ITS development in developing countries should proceed gradually until a national or regional architecture is in place and the bulk of the resources until then should be devoted to non-ITS infrastructure improvements.

REFERENCES

[1] Asian Development Bank Urbanization and Sustainability in Asia: Good Practice Approaches in Urban Region Development Manila, Philippines., 2006;

[2] U.S. Department of Transportation Systems Engineering for Intelligent Transportation Systems: An Introduction for Transportation Professionals. Washington, D. C., 2007;

[3] Khanal, J, "How Rapidly should Developing Countries Implement Intelligent Transportation Systems (ITS) to Solve the Growing Urban Traffic Congestion Problem?" Civil & Environmental Engineering journal, 2012, 2:4;

[4] U.S. Department of Transportation (1998) Road Less Traveled: Intelligent Transportation Systems for Sustainable Communities. Federal Highway Administration report;

[5] United Kingdom Parliamentary Offce of Science and Technology, "Intelligent Transport Systems," Postnote Number 322, January 2009,

[6] Robert D. Atkinson and Daniel D. Castro, "Digital Quality of Life: Understanding the Personal and Social Benefts of the Information Technology Revolution," Information Technology and Innovation Foundation, October 1, 2008, 106,

[7] Daniel Bursaux, "Transportation Policy in France," presentation at 15th ITS World Congress New York City, November 17, 2008.

[8] Tyler Duvall, Acting Undersecretary of Policy at U.S. Department of Transportation under George W. Bush, presentation at 15th ITS World Congress, November 17, 2008.

[9] Intelligent Transportation Society of America, "VII White Paper Series: Primer on Vehicle-Infrastructure Integration," 2005.

[10] Terry Warin, "Overcoming Barriers to ITS Implementation in the Asia Pacifc Region," presentation at 15th ITS World Congress, New York City, November 18, 2008.

[11] U.S. Department of Transportation, Research and Innovative Technology Administration, "Intelligent Transportation Systems (ITS) Strategic Plan: Background and Processes," January 8, 2010,
[12] Frances Penwill-Cook, "Intelligent Transportation Systems: Driving into The Future," Road Traffc Technology, September 26, 2008,
[13] Japan Highway Industry Development Organization, "ITS Handbook Japan 2007-2008."
[14] ITU-T Technology Watch Report #1, Intelligent Transport Systems and CALM, October 2007
[15] Ddaniel R. Jordanand Tthomas A. Horan, Intelligent transportation systems and sustainable communities, Transportation research record, Paper no. 971098
[16] ITS Technical notes for developing countries. http://www.worldbank.org/transport/roads/its%20docs/ITS%20Note%%201.pdf.

ABSTRACTS

Session 1: Infrastructures and platforms to support communities[1]

S1.1 Sustaining life during the early stages of disaster relief with a Frugal Information System: Learning from the Great East Japan Earthquake.*

Mihoko Sakurai, Jiro Kokuryo (Keio University, Japan); Richard Watson (University of Georgia, USA); Chon Abraham (College of William and Mary, USA)

Important lessons for responding to a large-scale disaster can be gleaned from the March 11, 2011 Great East Japan earthquake and tsunami. The failure of the electrical power system and the resultant loss of information communication and processing capability severely constrained the recovery work of many municipalities. It was difficult for supporting organizations to collect and share information. A frugal Information System (IS) designed around the four U-constructs is suggested as a solution to handle the early stages of disaster relief. This paper focuses on the most frequently available device, the cellular phone, as the foundation for a frugal IS for disaster relief. Familiar and available tools place minimal stress on an already stressed system.

S1.2 A Model for Creating and Sustaining Information Services Platform Communities: Lessons learnt from Open Source Software.

Sulayman K Sowe, Koji Zettsu, Yohei Murakami (National Institute of Information and Communications Technology, NICT, Japan)

Many research institutions are building cloud-based information services platforms (ISPs) that enable their researchers, scientists, and the general public use information assets, share knowledge and experience, and create sustainable communities. However, there is no guarantee that when you build an ISP this will happen. Part of the problem is because ISP providers lack the model to help them facilitate the building of sustainable communities. In this paper, we present a model for creating and sustaining communities on the ISP being developed by the National Institute of Information and Communications Technology (NICT) of Japan. Inspired by the way Open Source software communities operate, we describe the model concept, its settings, and the tools ISP communities may need to support their contribution towards the development of products and services. Our experience in the design and implementation of the model provides useful insights into emerging ICT trends and the means for ISP providers to identify, at an early stage, the requirements for creating successful products and services ecosystem.

S1.3 Security technologies for the protection of critical infrastructures - ethical risks and solutions offered by standardization.

Simone Wurster (Technische Universität Berlin, Germany)

The added value of standards is shown in numerous research articles. Several recent studies also highlight the need for security standards. Security products and services may bear ethical and privacy-related risks which can impede acceptance of new security solutions. Specific privacy standards may help to overcome such problems, but privacy issues of security technologies are not covered by standardization research so far. This paper deals with the topic from mainly German and European perspectives. Based on a survey in the German security research program, it gives an overview of security technologies, the specific risks they bear and their importance. Three technology-related categories were identified: surveillance solutions for detection from distance, solutions for obtrusive detection and data processing. Relevant risks were described and discussed. Solutions based on standardization were shown. The paper finishes by giving recommendations for new privacy standards.

[1] Papers marked with an "*" were nominated for the three best paper awards.

Session 2: Future communication services to sustain communities

S2.1 Invited Paper: Visible Light Communication Using Sustainable Led Lights.

Shinichiro Haruyama (Keio University, Japan)

LED lights are becoming widely used for homes and offices for their luminous efficacy improvement. Visible light communication (VLC) is a new way of wireless communication using visible light. Typical transmitters used for visible light communication are visible light LEDs and receivers are photodiodes and image sensors. We present new applications which will be made possible by visible light communication technology. Location-based services are considered to be especially suitable for visible light communication applications.

S2.2 Selecting the Best Communication Service in Future Network Architectures.

Rahamatullah Khondoker (University of Kaiserslautern & Fraunhofer Institute of Secure Information Technology, Germany); Paul Mueller (University of Kaiserslautern, Germany); Kpatcha Bayarou (Fraunhofer Institute for Secure Information Technology, Germany)

As the number of future network architectural approaches increases, the possibility of offering many similar services with different qualities of service is increasing. Therefore, it will be required to select a suitable, or the best, service from the set of alternative services. This paper proposes a matching process and an adapted analytic hierarchy process to accomplish this task. The matching process is used to determine if a service is suitable. When more than one suitable service is available, the adapted analytic hierarchy process is used to select the best service.

S2.3 Using the RFID Technology to Create a Low-Cost Communication Channel for Data Exchange.

Ivan Farris, Antonio Iera, Silverio Carlo Spinella (University Mediterranea of Reggio Calabria, Italy)

This paper proposes a methodology to use the RFID technology (more specifically the RFID tags) as a novel "communication channel", to support data exchanges in high pervasive environments, analogously to more traditional short-range communication technologies (WiFi, ZigBee, Bluetooth). To this aim, the further research issue of creating so called RANs (RFID-Area Networks - in analogy with LANs, Local Area Networks, PANs, Personal Area Networks, etc.) is addressed. These are made up of groups of RFID readers into which the functionality for exchanging data over the introduced "RFID virtual channel" within the generic RAN, in either a broadcast or a unicast modality, is embedded. From initial studies on its functional behavior, it emerges that the proposed method may actually allow to exploit a further (currently largely wasted although available "at no cost") channel in future scenarios populated by tagged everyday-life objects.

S2.4 Non-Directed Indoor Optical Wireless Network with a Grid of Direct Fiber Coupled Ceiling Transceivers for Wireless EPON Connectivity.

Dimitar Kolev, Kazuhiko Wakamori (Waseda University, Japan); Takahiro Kubo, Takashi Yamada, Naoto Yoshimoto (NTT Corporation, Japan)

In this paper we propose an optical wireless system for indoor communication with a grid of ceiling transceivers, based on direct fiber coupling technology. The proposed network is fully compatible with EPON standard that uses point to multipoint broadcasting in the downstream and can guarantee a high speed two-way connection for multiple mobile devices. We present the transmission analysis for the both downlink and uplink and discuss the eye safety issues regarding our proposal. Furthermore, deeper analysis of the system synchronization is conducted and the distribution of the delay in the overlapping zones is presented.

Session 3: Supporting remote communities

S3.1 Implementation Roadmap for Downscaling Drought Forecasts in Mbeere Using ITIKI.*****

Muthoni Masinde, Antoine Bigomokero Bagula (University of Cape Town, South Africa); Nzioka Muthama (University of Nairobi, Kenya)

Mbeere is in Eastern Kenya and it has an average of 550 mm annual rainfall and therefore classified under Arid and Semi-Arid Lands. It has fragile ecosystems, unfavorable climate, poor infrastructure and historical marginalization; the perennial natural disasters here are droughts. Of importance to this paper is the fact that despite its vast area of 2,093 km2, there is no single weather station serving the area. The main source of livelihood is rain-fed marginal farming and livestock keeping by small-scale and peasant farmers who rely mostly on the indigenous knowledge of seasons in making cropping decisions. ITIKI; acronym for Information Technology and Indigenous Knowledge with Intelligence is a bridge that integrates indigenous drought forecasting approach into the scientific drought forecasting approach. ITIKI, a framework initiated by the authors of this paper was adopted and adapted from the word itiki which is the name used among the Mbeere people to refer to an indigenous bridge used for decades to go across rivers. ITIKI makes use of mobile phones, wireless sensor networks and artificial intelligence to downscale weather/drought forecasts to individual farmers. ITIKI implementation project in Mbeere commenced in August 2012; this paper describes the implementation roadmap for this project.

S3.2 A Sustainable Integrated-Services Community Learning Center.

Prasit Prapinmongkolkarn, Supavadee Aramvith, Chaodit Aswakul, Anegpon Kuama, Sucharit Koontanakulvong (Chulalongkorn University, Thailand); Ekachai Phakdurong (THAICOM PLC, Thailand)

This paper proposes the concept of integrating the community learning center with e-health facility and natural disaster warning system. The model for sustainability and ubiquity of ICT facilities in community has been achieved through three years of experiences in implementation of universal service obligation (USO) schemes in Thailand. From the beginning, the community learning centers have been designed with the principle of sustainability, flexibility, easy-to-use, cost saving and local participation concepts. With the country's lesson learned in the recent great flood last year and to prepare our country for future natural disasters, it is natural that the community learning center is proposed to extend its conventional services with real-time information and data service system for flood warning. This new service of the center can expectedly cowork with the conventional national television broadcasting, radio, mobile phone, satellite and amateur radio services. It is our belief that such integrated-services community learning center concept, the first of its sort, will enhance the education of people by bridging the digital divide in USO, to improve health care and wellness of people by telehealth service, as well as to make our country ready for unforeseen natural disaster crisis in the future.

Session 4: Resource discovery and management

S4.1 System design and numerical analysis of adaptive resource discovery in wireless application networks.*****

Wei Liu, Takayuki Nishio, Ryoichi Shinkuma (Kyoto University, Japan)

In this paper, we propose an adaptive resource discovery method in heterogeneous wireless application networks. The adaptive method uses either centralized mode or flooding mode to discover available resources according to different network status. The proposed adaptive method is used to reduce energy consumption in resource discovery process. We establish theoretical energy model for both modes. A heuristic algorithm is designed to implement the proposed adaptive method. It is also proved to be energy efficient through extensive evaluations.

S4.2 Design and Implementation of virtualized ICT resource management system for carrier network services toward Cloud computing era.*****

Yoshihiro Nakajima, Hitoshi Masutani, Wenyu Shen, Osamu Kamatani, Masaki Fukui (NTT Network Innovation Laboratories, Japan); Hiroyuki Tanaka, Katsuhiro Shimano (NTT, Japan); Ryutaro Kawamura (Cyber Solutions Laboratories, NTT corporation, Japan)

This paper describes the design and implementation of a virtualized information and communications technology (ICT)resource management system called "Management Engine" (ME) for carrier network services to realize flexible service operation and dynamic resource accommodation between multiple services in the cloud computing era. To facilitate network services using virtualized ICT resources in a carrier network, a virtualized ICT information model is designed that expresses the relationship and mapping between physical resources and virtual resources for failure handling and analysis required in network carrier operations and management. A disaster recovery scenario to guarantee high-priority voice communication service in case of a largescale natural disaster is used to examine MEs capability and functionality for providing next generation mobile network service over an OpenFlow network and virtualized servers. As a result, it is found that ME performs both integrated ICT resource management and inter-service dynamic resource accommodation. Further research areas and standardization issues ascertained from prototype experiment results are presented.

S4.3 Harmonized Q-Learning For Radio Resource Management In LTE Based Networks.

Dhananjay Kumar (Anna University, India); Kanagaraj Nachimuthu Nallasamy (Alcatel-Lucent India Limited, India); Sri Lakshmi (Anna University MIT Campus, India)

The efficient management of radio resource is highly imperative so as to meet the vast application requirements in future high speed wireless networks such as Long Term Evolution-Advanced (LTE-A). The current research on applying machine learning algorithms either focuses on packet scheduling in infrastructure network or in cognitive radio in ad-hoc environment. Our study on spectrum usage indicates that there is a lot of room for optimization of spectrum in a multi-operator scenario of LTE systems which covers large customer over a vast geographical area. In this paper, we introduce the concept of Harmonized QLearning (HQL) for the radio resource management in LTE based networks that efficiently manage its resource pool dynamically. The multi-operator system is modeled on the game theory based Q-Learning. Our system level simulation of the proposed algorithm shows higher throughput while meeting the real-time resource requirement of each player.

Session 5: Supporting future applications

S5.1 Invited Paper: Hybridcast: a new media experience by integration of broadcasting and broadband

Hisayuki Ohmata; Masaru Takechi; Shigeaki Mitsuya; Kazuhiro Otsuki; Akitsugu Baba; Kinji Matsumura; Keigo Majima; Shunji Sunasaki (NHK, Japan)

Broadcasting has a role for the public service. Providing the same information to a large number of people at the same time has benefitted modern society in many ways, including presenting the forefront of lifestyle trends, offering dependable media during disasters and costeffective transmissions. On the other hand, services over the Internet satisfy the individual's needs as seen in customization for each, interactive communication and user-generated media. NHK is developing "Hybridcast", a service platform integrating broadcasting with the Internet. This platform can enhance broadcasting programs and provide other various services by the best mix of features of both media. This paper describes the system and examples of service on Hybridcast for the general public including minority viewers. The next-generation media for a sustainable society will emerge from Hybridcast which is expected to be launched in 2013.

S5.2 Standard-based Publish-Subscribe Service Enabler for Social Applications and Augmented Reality Services.*****

Oscar Rodríguez Rocha (Politecnico di Torino, Italy); Boris Moltchanov (Telecom Italia, Italy)

A Publish/Subscribe mechanism based on the Open Mobile Alliance's (OMA) Next Generation Services Interface (NGSI) open standard, allows interfacing the information available from many publishers with heterogeneous customers. Pervasive devices (including mobile smartphones) publish a huge amount of real world information, which afterwards is accessed through web browsers and applications. The adoption of an open standard interface between information publishers and consumers allows to reduce the gap in the technologies used on both sides, therefore, include new actors into the services, increase the service offers and increment the world-wide and cross-domain usage of services based on the Publish/Subscribe paradigm. Major European Industrial Entities supported by the EU Research Program are deriving a cross-domain Future Internet open standard technology to be adopted and used in any application domain by any customer for any needs. The reference open standard chosen is OMA's NGSI. The open standard based technological binding created in the FI-WARE EU funded project and provided with an open reference implementation performed by Telecom Italia is demonstrated through examples of Augmented-Reality and social-impacting services that improve the quality of life for people (including those decease affected).

S5.3 QoXphere: A New QoS Framework for Future Networks.*****

Eva Ibarrola, Eduardo Saiz, Luis Zabala, Leire Cristobo (University of the Basque Country, Spain); Jin Xiao (University of Waterloo, Canada)

The telecommunications sector has experienced significant changes over the past few years. The advent and rise of new applications and services, together with a competitive market, has led to a complex scenario in which quality of service (QoS) plays a major role. Under this condition, novel QoS regulation and standardization initiatives are required. During the last few years new terms and concepts, such as Quality of Experience (QoE) or QoS Perceived (QoP), have been included in the updated and new QoS-related standards as to better integrate the user's point of view, as opposed to only network performance parameters. The influence of the user's satisfaction on the Quality of Business (QoBiz) has also been given increased attention in the regulation and standardization bodies recently. The result is a loose collection of metrics and models that are not standardized and do not integrate all aspects of quality. Such integration is necessary to assure the successful development of this sector. This paper presents a new and integrated QoS model (QoXphere) that is spherical, adaptive and multi-layered.

S5.4 Telebiometric Information Security and Safety Management.*****

Phillip H Griffin (Booz Allen Hamilton, USA)

Organizations that rely on human-oriented technologies such as telebiometrics should protect and manage the safety and security of their physical and information assets. Data that documents the safe and secure operation of telebiometric system devices should be collected and captured in an information security and safety event journal. Event journal data provides an audit trail that should be protected using digital signatures, encryption and other safeguards. A system heartbeat record should document and monitor the safety, performance, and availability of telebiometric system devices and alert system administrators to security and safety events and changes. Heartbeat data should provide metrics that inform the continuous improvement of a telebiometric information security and safety management program. A signcryption cryptographic message wrapper should protect event journal, biometric reference template, and other telebiometric information to promote user security and respect for user privacy rights.

Session 6: Standardisation Issues

S6.1 Invited Paper: Open Standards: a Shrinking Public Space in the Future Network Economy?

William Melody (LIRNE.NET, Aalborg University Copenhagen, Denmark)

The capacity to sustain most communities depends on their access to technical, economic and political resources. The interaction among technologies, markets and government policies that direct technology and market activity generates the resources. Most countries, and many international agencies have adopted information society policies to promote broadband infrastructure and NGN development as a platform for ICT applications. The goals are economic development and universal access, i.e., more inclusive communities. This paper examines how ICT sector market and policy trends are influencing the environment for developing innovative NGN applications and their essential supporting standards.
The interaction between ever more expansive and generous patent and copyright (IPR) awards in the ICT sector with the winner-take-all network characteristics of most ICT and content markets is fostering oligopoly markets based on proprietary standards. This trend toward increasing policy-permitted standards and market exclusivity as the foundation for asset values and industry growth steadily narrows the scope for NGN applications based on open standards. It reduces opportunities for participation in the development of future knowledge communities. A major initiative is proposed to build the evidentiary and analytical support for policy reforms that will reverse this trend.

S6.2 Innovation Management of Electrical Vehicle Charging Infrastructure Standards in the Sino-European Context.*****

Martina Gerst, Xudong Gao (Tsinghua University, P.R. China)

Energy challenges, changing consumer attitudes and evolving government mobility policies impact today's automotive industry. Mobility in sustainable communities of the 21st Century may to a considerable degree be based on New Electrical Vehicles (NEV) as an important part of electric mobility (e-mobility) concepts. One of the central factors to gain market acceptance is the interoperability of the different NEV sub-systems, particularly the standardization of the charging infrastructure. E-mobility is embedded in a rapidly changing, competitive and complex global environment, highly influenced by competing regional innovation policies. Therefore, this paper highlights some of the tensions in standardization management by Multi National Automotive Enterprises (MNAE) of a charging infrastructure in a Sino-European context.

Session 7: Energy Issues

S7.1 An Analytical Evaluation of Energy Consumption in Cooperative Cognitive Radio Networks.

Mahdi Pirmoradian, Olayinka Adigun, Christos Politis (Kingston University of London, United Kingdom)

This paper studies the total energy consumption of a cooperative cognitive radio network in coexistence with a stochastic multi-channel licensed network. Energy consumption of each cognition phase at the secondary user is mathematically analyzed and obtained. The numerical results presented show the various interactions between the secondary network size, availability of appropriate spectrum holes and the total energy consumption for different states of the cognitive radio network under discussion.

S7.2 Solar-Powered Cell Phone Access Point for Cell Phone Users in Emerging Regions.

Takuya Kato, Yoshihiro Kawahara (The University of Tokyo, Japan)

The availability of electrical power is a critical issue for building sustainable communication networks in emerging regions. Low-cost energy delivery encourages people to use communication services. This paper presents simulation results on the effect of the distribution of the surplus power generated for an access point (AP) to cell phone users. We show that 9.3% of the user population can use the excess power generated for the AP. Furthermore, we propose an energy-proportional server cluster to ensure computational resources for information services, such as for charging cell phones. The existing server hardware often wastes power in the idle state and is not energy proportional, and thus we designed the cluster to reduce this energy waste by matching the number of working servers to the number of incoming requests. Our prototype system with low-power and off-theshelf devices cuts energy consumption in the frequently observed idle state by 50% compared with an existing server cluster with equivalent performance.

S7.3 Proposal of a Sub- λ Switching Network and its Time-Slot Assignment Algorithm for Network with Asynchronous Time-Slot Phase.

Keisuke Okamoto (Kyoto University, Japan); Atsushi Hiramatsu (NTT, Japan)

We propose a sub-λ switching network which has fine granularity and low cost/power consumption. In this network, each wavelength is divided in time domain to achieve fine granularity, but buffers are eliminated from the core network to reduce switching cost and power consumption. Buffers are located only at the entrance of the network in order to groom ingress traffic, and all nodes in this network are operated synchronously under a certain time-control mechanism. The problem in this network is the rather long guard-time which is required to absorb the clock synchronization error and time-slot-phase difference, which is caused by the various fiber lengths between nodes (asynchronous phase network). To solve this problem, we propose a novel time-slot assignment algorithm using multi-time-slot bonding technique and delay-shift packing technique with global-time-based delay shift. By using the proposed method, we could improve the utilization of link capacity by 45% compared with the conventional method.

Poster Session

P.1 A Proposal of a New Packet Scheduling Algorithm and Its Evaluation.

Tetsushi Matsuda (Mitsubishi Electric Corp, Japan)

ITU and other SDOs have launched oneM2M initiative recently and the standardization of M2M is now accelerating. The current access and core networks built for today's network services will be used as a common network infrastructure for M2M network with some modifications. When the current access and core networks are used for both the current network services and M2M services, communications equipments at the network edge need to handle a large number of communication flows which are a mix of large volume data communication such as web access and M2M data communication at the same time. To satisfy the QoS requirements of many applications including M2M applications, communications equipments at the network edge will need to support both minimum guaranteed rate service and low delay forwarding service for small sized packets. In this paper, we propose a packet scheduling algorithm which can provide minimum guaranteed rate service and which can reduce the scheduling delay of small packets. It can be used in access network communications equipment such as edge router and OLT. We also evaluate the proposed algorithm by simulation.

P.2 Digital Space Transmission of An Interference Fringe-Type Computer-Generated Hologram Using IrSimple.

Masataka Tozuka, Koki Sato, Makoto Ohki (Shonan Institute of Technology, Japan); Kunihiko Takano (Tokyo Metropolitan College of Industrial Technology, Japan)

In this paper, we present a method to perform a digital space transmission of an interference fringe-type computer generated hologram using IrSimple. IrSimple was defined by IrDA technical standard, and it was developed for the purpose of sending image data at high speed using an infrared-rays. We performed infrared digital transmission using IrSimple, and we minimized the size of the transmitted file by using a suitable compression method. The size of the compressed file was very small compared with that of the bitmap file. The transmission preserved the quality of the representation while requiring a short transmission time.

P.3 Integrated Telecommunication Technology for the Next Generation Networks.

Victor Tikhonov (A S Popov Odesa National Academy of Telecommunications, Ukraine); Petro Vorobiyenko (ONAT, Ukraine)

The paper focuses researches on next generation network (NGN) convergence. A set of comprehensive data-transfer axioms premise holistic approach to benefit diverse packetto-circuit switching techniques. A novel dynamic flow switching (DFS) method introduced to facilitate digital telecommunication channels along with appropriate multipurpose network meta-protocol (MNP). The tenets of integrated telecommunication technology (ITT) platform developed for the transport stratum in ITU-T model of NGN. Two-dimensional quality of service (QoS) palette and related cost-to-quality ratio (CQR) function proposed for multimedia traffic control. An elastic ITT-address system originated for ITT-platform to meet the challenge of Internet-scope expansion. The paper intends to contribute future network engineering.

P.4 Research on ICT service energy impact assessment method: How much energy to manufacture a chip.

Sebastien Schinella, Stephane Le Masson, Tomoko Tanaka (France Télécom-Orange, France); Didier Marquet (Orange Labs, France); Xavier Chavanne, Jean-Pierre Frangi (Université Paris Diderot, France)

Telecommunications are expected to reduce the energetic impact of human society through dematerialization. But in addition to their utilization consumption, the ICT equipments are responsible of energy consumption for their fabrication. To assess as precisely as possible their consumption and the gain compared to physical services, we use a new modular and open method intended to reduce errors and highlight possible improvements. The lithography phase of chips manufactured from extra-pure silicon wafers is responsible for about 70% of the consumption of chips fabrication. The main elements of this phase are the tools, which make the operations, and the air circuit, which cleans the air and control its levels of temperature and humidity 24h/24. This element is deeply analyzed in this paper, as a key module of the method which can be reused for other studies. Many parameters take part in the air circuit consumption, particularly the climate of the place where the factory is built, the class of the clean room and the amount of particles emitted. We try to put them ahead to understand this consumption and to know how to improve the energy efficiency.

P.5 Robust Audio Watermarking Based on Dynamic DWT with Error Correction.

Hemam Ayed Alshammas (The University of Jordan, Jordan)

Audio watermarking was introduced as a solution for the arising challenges facing audio ownership verification. These challenges are a result of easiness and high speed of copying and distribution digital audio. This paper presents enhancements in the performance of an audio ownership verification system that has been reported previously. The proposed system is based on the Discrete Wavelet Transform (DWT). A new approach for audio signal framing, dynamic DWT leveling, error correction code and new embedding methods are suggested to improve the watermark bit rate, minimum audio-cover period, quality of the watermarked audio and watermark robustness against audio attacks. The evaluation of the suggested system showed the following improvements: the watermark bit rate increased 23.4 times, 92% reduction in the minimum required audio-cover period, 54% increase in the Signalto-Noise Ratio (SNR) of watermarked audio, and it demonstrated better robustness against watermarking benchmark attacks.

P.6 Self-Verified DNS Reverse Resolution.

Zheng Wang, Rui Wang (China Organizational Name Administration Center, P.R. China)

Domain Name System (DNS) reverse resolution is commonly relied on by anti-spam techniques to verify the email origins and by measurements or applications to uncover the host information. But the current practice is not able to clarify the IP addresses with no reverse resolution response and the source verification process is not optimized in terms of network bandwidth and response latency. This paper proposes an explicit scheme to bind A/AAAA resource records (RRs) with their matching PTR RRs by introducing APTR/AAAAPTR RR types. The DNS cache server can automatically switch from forward resolution to reverse resolution when handling the APTR/AAAAPTR RR types. This scheme enables the negative verification if no reverse records are returned for APTR/AAAAPTR records. Furthermore, the analytical and numerical results show that the number of queries and response delay are significantly cut by the proposed scheme.

P.7 A Periodic Combined-Content Distribution Mechanism in Peer-Assisted Content Delivery Networks.

Naoya Maki, Ryoichi Shinkuma (Kyoto University, Japan); Tatsuya Mori (NTT, Japan); Noriaki Kamiyama, Ryoichi Kawahara (NTT Service Integration Laboratories, Japan)

The concept of peer-assisted content delivery networks (CDNs) lets other nearby altruistic clients forward requested content files instead of the source servers, which works to localize overall traffic. Our prior work proposed a traffic engineering scheme to localize traffic in peer-assisted CDNs. To induce altruistic clients to download content files that are most likely to contribute to localizing network traffic, this scheme combines the content files and allows them to obtain the combined content file while keeping the price unchanged from the single-content price. Although we have discussed how much traffic only a set of combined content files can theoretically reduce, we can expect further traffic localization by distributing combined content files to multiple altruistic clients. This paper proposes a periodic combined-content distribution mechanism based on our scheme. This mechanism determines when combined content files should be provided by considering the network cache state. Computer simulations confirmed that our new mechanism could double performance.

P.8 Medication Error Protection System with a Body Area Communication Tag.

Yoshitoshi Murata, Nobuyoshi Sato, Tsuyoshi Takayama (Iwate Prefectural University, Japan); Shuji Ikuta (NTT ComTechnology Corporation, Japan)

Errors in administering medication are serious problems. Most errors are due to some confusion between the patient and the medication. The bar-code has been used to deal with this problem. However, since a nurse has to put a reader over a bar-code tag, there is still room for improvement when, say, a nurse is affected by stress due to too heavy a workload. As a safer alternative to bar-codes and RFIDs, we propose Touch-tag, a body area communication tag. The concept is that the patient wears a tag and the nurse has a Touch-tag reader that reads the ID on the tag by a nurse just by touching the patient. Since a nurse usually touches a patient during administration of medication, the medical error protection system using the Touch tag does not involve any additional work. We describe the medical error protection system with the Touch-tag and experiments to confirm whether the Touch tag works well or not.

P.9 Intra-City Digital Divide Measurements Through Clustering.

Tugra Sahiner, Gunes Karabulut Kurt (Istanbul Technical University, Turkey); Aysegul Ozbakır (Yildiz Technical University, Turkey)

With latest development of telecommunication technologies and end user's increased bandwidth and mobility demand reachability of information and communication technologies (ICT) became more critical. In this paper, we approach to end user behaviors and the reachability to ICT by tackling digital divide concept along with clustering analysis. To the best of our knowledge, this research is a unique case study that attempts to analyze digital divide at intra-city level by neighborhoods. While governments and institutions, such as ITU, are in question of whether the global divide is widening or narrowing, there are no studies, neither in the literature nor in practice to understand the gap between ICT users in a city. With this goal, Istanbul habitants were asked to fill a questionnaire, in order to be classified in terms of their technology reachability and reasons of using ICT. Then, clustering analysis was performed to questionnaire results. Respondents have been clustered into sub groups from digital divide perspective. The required steps and suitable clustering techniques during this process are discussed with determination of questions at the end which are commonly answered by respondents with different ICT knowledge, which may lead us determine precise reasons of digital gap later on.

P.10 ICT Innovation In South Africa: Lessons Learnt From Mxit.

Michael Kahn (University of Stellenbosch, South Africa)

In the last decade South African innovators have produced two game changing innovations: Thawte and Mxit, the former for Internet security and the latter for mobile messaging. It is of interest to understand what it is in the local system of innovation that has enabled such innovations to emerge. Mxit is one of the first instant messaging systems to be freely available for almost all mobile phones. From a simple text service it has now evolved into a multimedia platform that offers gaming, education, community support services, and that is poised to enter the money transfer and payment market. The paper locates this development in the context of the telemetry sectoral system of innovation of South Africa, and the innovation ecology of the city of Stellenbosch that is designated as 'Silicon Vineyard.' Mxit has particular importance as a means of bridging the various divides that continue to characterize post-Apartheid South Africa.

P.11 Review of challenges in national ICT policy process for African countries.

Frank Makoza, Wallace Chigona (University of Cape Town, South Africa)

National Information and Communication Technology (ICT) policies are vital in supporting socioeconomic development agendas. However, the formulation and implementation of national ICT policies is often beset with myriad of challenges which lender the policies ineffective in most developing countries. While there is substantial body of knowledge on the challenges for national ICT policies, no study has not yet dealt with the challenges holistically. This paper reports on the results of review of literature on the challenges for national ICT policy process in African countries using Grounded Theory method. The review categorised the challenges related to agenda setting, policy formulation, legal frameworks, implementation and evaluation. To mitigate some of the challenges, it is suggested that stakeholders' participation should be encouraged; monitoring and evaluation with mechanisms for learning should be integrated in all the phases of the policy process.

P.12 The role of intelligent transportation systems in developing countries and importance of standardization.

Muzaffar Djalalov (Scientific Engineering and Marketing Researches Center, Uzbekistan)

The traffic accidents and congestions are getting a serious problem all over the world. The problem is growing fastest in developing countries where urbanization and the use of motorized vehicles is increasing most rapidly. One alternative solution is a concept of Intelligent Transportation Systems (ITS). It provides the ability to gather, organize, analyze, use, and share information about transportation systems. In this paper such issues as concept of sustainable communities with transportation and ITS is discussed and scheme of ITS deployment in developing countries and its benefits are presented. Moreover, the analysis of ITS current situation in some developed and developing countries and the role of standardization are presented.

INDEX OF AUTHORS

Index of Authors

Printed in Switzerland
Geneva, 2013