UBICOMM 2010

The Fourth International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies

October 25-30, 2010 - Florence, Italy

Editors

Jaime Lloret Mauri

Sergey Balandin

Cosmin Dini

UBICOMM 2010

Foreword

The Fourth International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies (UBICOMM 2010), held from October 25 to October 30, 2010 in Florence, Italy, was a multi-track event covering a large spectrum of topics related to developments that operate in the intersection of mobile and ubiquitous technologies on the one hand, and educational settings in open, distance and corporate learning on the other, including learning theories, applications, and systems.

The rapid advances in ubiquitous technologies make fruition of more than 35 years of research in distributed computing systems, and more than two decades of mobile computing. The ubiquity vision is becoming a reality. Hardware and software components evolved to deliver functionality under failure-prone environments with limited resources. The advent of web services and the progress on wearable devices, ambient components, user-generated content, mobile communications, and new business models generated new applications and services. The conference made a bridge between issues with software and hardware challenges through mobile communications.

The goal of UBICOMM 2010 was to bring together researchers from the academia and practitioners from the industry in order to address fundamentals of ubiquitous systems and the new applications related to them. The conference provided a forum where researchers were able to present recent research results and new research problems and directions related to them.

Advances in web services technologies along with their integration into mobility, online and new business models provide a technical infrastructure that enables the progress of mobile services and applications. These include dynamic and on-demand service, context-aware services, and mobile web services. While driving new business models and new online services, particular techniques must be developed for web service composition, web service-driven system design methodology, creation of web services, and on-demand web services.

As mobile and ubiquitous computing becomes a reality, more formal and informal learning will take pace out of the confines of the traditional classroom. Two trends converge to make this possible; increasingly powerful cell phones and PDAs, and improved access to wireless broadband. At the same time, due to the increasing complexity, modern learners will need tools that operate in an intuitive manner and are flexibly integrated in the surrounding learning environment.

Educational services will become more customized and personalized, and more frequently subjected to changes. Learning and teaching are now becoming less tied to physical locations, co-located members of a group, and co-presence in time. Learning and teaching increasingly take place in fluid combinations of virtual and "real" contexts, and fluid combinations of presence in time, space and participation in community. To the learner full access and abundance in communicative opportunities and information retrieval represents new challenges and affordances. Consequently, the educational challenges are

numerous in the intersection of technology development, curriculum development, content development and educational infrastructure.

We take here the opportunity to warmly thank all the members of the UBICOMM 2010 technical program committee as well as the numerous reviewers. The creation of such a broad and high quality conference program would not have been possible without their involvement. We also kindly thank all the authors that dedicated much of their time and efforts to contribute to UBICOMM 2010. We truly believe that, thanks to all these efforts, the final conference program consisted of top quality contributions.

This event could also not have been a reality without the support of many individuals, organizations and sponsors. We also gratefully thank the members of the UBICOMM 2010 organizing committee for their help in handling the logistics and for their work that is making this professional meeting a success. We gratefully appreciate to the technical program committee co-chairs that contributed to identify the appropriate groups to submit contributions.

We hope Florence provided a pleasant environment during the conference and everyone saved some time for exploring this historic city.

UBICOMM 2010 Chairs:

Jaime Lloret Mauri, Polytechnic University of Valencia, Spain
Petre Dini, IARIA / Concordia University, Canada
Sathiamoorthy Manoharan, University of Auckland, New Zealand
Zary Segal, UMBC, USA
Sergey Balandin, Nokia, Finland
Cao Le Thanh Man, Hitachi Ltd., Japan
Tetsuji Takada, Advanced Industrial Science and Technology, Japan
Carlo Mastroianni, CNR, Italy
Junya Nakata, Hokuriku Research Center/ Japan Advanced Institute of Science and Technology, Japan
Shaya Potter, IBM Watson Research Lab, USA

UBICOMM 2010

Committee

UBICOMM Advisory Chairs
Jaime Lloret Mauri, Polytechnic University of Valencia, Spain
Petre Dini, IARIA / Concordia University, Canada
Sathiamoorthy Manoharan, University of Auckland, New Zealand
Zary Segal, UMBC, USA

UBICOMM 2010 Industry Liaison Chairs
Sergey Balandin, Nokia, Finland
Cao Le Thanh Man, Hitachi Ltd., Japan
Tetsuji Takada, Advanced Industrial Science and Technology, Japan

UBICOMM 2010 Research/Industry Chairs
Carlo Mastroianni, CNR, Italy
Junya Nakata, Hokuriku Research Center/ Japan Advanced Institute of Science and Technology, Japan
Shaya Potter, IBM Watson Research Lab, USA

UBICOMM 2010 Technical Program Committee
Afrand Agah, West Chester University of Pennsylvania, USA
Chang-Jun Ahn, Hiroshima City University, Japan
Mehran Asadi, West Chester University of Pennsylvania, USA
Sergey Balandin, Nokia, Finland
Matthias Baumgarten, University of Ulster-Belfast, Northern Ireland, UK
Shlomo Berkovski, CSIRO, Australia
Gennaro Boggia, Politecnico di Bari, Italy
Matthias Böhmer, Münster University of Applied Sciences, Germany
Jihen Bokri, ENSI, Tunisia
Mahmoud Boufaida, Mentouri University of Constantine, Algeria
Song Boyeon, Korea University, South Korea
Anisoara Calinescu, University of Oxford, UK
Jose Manuel Cantera Fonseca, Telefonica Investigacion y Desarrollo, Spain
Juan Vicente Capella Hernández, Universidad Politécnica de Valencia, Spain
Kevin (Bongsug) Chae, Kansas State University - Manhattan, USA
Ruay-Shiung Chang, National Dong Hwa University, Taiwan
Michael Collins, University College Dublin, Ireland
Luca De Nardis, University of Rome La Sapienza, Italy
Michael Decker, University of Karlsruhe (TH), Germany
Steven A. Demurjian, The University of Connecticut - Storrs, USA
Kamil Dimililer, Near East University, Turkey
Jianguo Ding, Norwegian University of Science and Technology (NTNU) - Trondheim, Norway
Santiago Eibe, Universidad Politecnica de Madrid, Spain
Josu Etxaniz Marañón, Euskal Herriko Unibertsitatea/Universidad del País Vasco, Spain
Andras Farago, The University of Texas at Dallas - Richardson, USA
George Fiotakis, Univeristy of Patras, Greece

Kary Främling, Helsinki University of Technology, Finland
Korbinian Frank, German Aerospace Center - Institute of Communications and Navigation, Germany
Flavius Frasincar, Erasmus University Rotterdam, The Netherlands
Crescenzio Gallo, University of Foggia, Italy
Miguel Garcia Pineda, Polytechnic University of Valencia, Spain
Dimitris Gavrilis, IMIS/Athena Research Centre, Greece
Laurent George, Université Pairs XII - Val de Marne, France
Kyller Gorgonio, Signove Tecnologia S.A., Brazil
Dominic Greenwood, Whitestein Technologies AG, Switzerland
Frederic Guidec, Université de Bretagne Sud / Université Européenne de Bretagne, France
Arthur Herzog, Technische Universität Darmstadt, Germany
Javier Alexander Hurtado, University of Cauca, Colombia
Noha Ibrahim, Grenoble Informatics Laboratory, France
Ivan Jelinek, Czech Technical University in Prague, Czech Republic
Lang Jia, UCD-Dublin, Ireland
Zhao Junhui, NEC Laboratories China, China
Faouzi Kamoun, University of Dubai, UAE
Dimitris Karagiannis, University of Vienna, Austria
Iman Keivanloo, Concordia University - Montreal, Canada
Reinhard Klemm, Avaya Labs Research-Basking Ridge, USA
Dattatraya Vishnu Kodavade, Shivaji University-Kolhapur, India
Ruwini Kodikara, RMIT University, Australia
Ioannis Krontiris, University of Mannheim, Germany
Natalie Kryvinski, University of Vienna, Austria
Hien Nam Le, Norwegian University of Science and Technology - Trondheim, Norway
Frédéric Le Mouël, INRIA/INSA Lyon, France
Nicolas Le Sommer, Université Européenne de Bretagne, France
Philippe Le Parc, Université de Bretagne Occidentale - Brest, France
Juong-Sik Lee, Rensselaer Polytechnic Institute, USA
Jian Liang, Cork Institute of Technology, Ireland
David Lizcano Casas, Universidad Politécnica de Madrid, Spain
Jaime Lloret Mauri, Polytechnic University of Valencia, Spain
Elsa María Macías López, University of Las Palmas de Gran Canaria, Spain
Sathiamoorthy Manoharan, University of Auckland, New Zealand
Teddy Mantoro, KICT/IIUM, Malaysia
Sergio Martín Gutiérrez, UNED-Spanish University for Distance Education, Spain
Oscar Martinez Bonastre, Miguel Hernández University, Spain
Carlo Mastroianni, CNR, Italy
Natarajan Meghanathan, Jackson State University, USA
Elisabeth Métais, CNAM/CEDRIC, France
Moeiz Miraoui, Université du Québec/École de Technologie Supérieure - Montréal, Canada
Neeraj Mittal, The University of Texas at Dallas - Richardson, USA
Hamid Mukhtar, National University of Science and Technology - Islamabad, Pakistan
Maurice Mulvenna, University of Ulster, UK
Junya Nakata, Hokuriku Research Center/ Japan Advanced Institute of Science and Technology, Japan
Rui Neves Madeira, New University of Lisbon, Portugal
Cao Le Thanh Man, Hitachi Ltd., Japan
Quang Nhat Nuyen, Hanoi University of Technology, Vietnam

Akihiko Ohsuga, The University of Electro-Communications (UEC) - Tokyo, Japan
Netzahualcoyotl Ornelas Garcia, University Paris 13, France
Carlos Enrique Palau Salvador, University Polytechnic of Valencia, Spain
Pejman Panahi, Urmia Azad University, Iran
Andre Peters, University of Rostock, Germany
Jari Porras, Lappeenranta University of Technology, Finland
Daniel Porta, German Research Center for Artificial Intelligence (DFKI) - Saarbrücken, Germany
Shaya Potter, IBM Watson Research Lab, USA
Michele Ruta, Politecnico di Bari, Italy
Ichiro Satoh, National Institute of Informatics - Tokyo, Japan
Zary Segal, UMBC, USA
Álvaro Suárez Sarmiento, University of Las Palmas de Gran Canaria, Spain
Tetsuji Takada, Advanced Industrial Science and Technology, Japan
Yoshiaki Taniguchi, Osaka University, Japan
Saïd Tazi, LAAS-CNRS, Université de Toulouse / Université Toulouse1, France
Stephanie Teufel, University of Fribourg, Switzerland
Giannis Tsakonas, University of Patras, Greece
Chih-Cheng Tseng, National Ilan University, Taiwan
Maarten Weyn, Artesis University College of Antwerp, Belgium
Matthias Wieland, Universität Stuttgart, Deutschland
Fatos Xhafa, University of London, UK
Yu Xiao, Aalto University - Espoo, Finland
Xiaochun Xu, University of Florida, USA
Chai Kiat Yeo, Nanyang Technological University, Singapore

Copyright Information

For your reference, this is the text governing the copyright release for material published by IARIA.

The copyright release is a transfer of publication rights, which allows IARIA and its partners to drive the dissemination of the published material. This allows IARIA to give articles increased visibility via distribution, inclusion in libraries, and arrangements for submission to indexes.

I, the undersigned, declare that the article is original, and that I represent the authors of this article in the copyright release matters. If this work has been done as work-for-hire, I have obtained all necessary clearances to execute a copyright release. I hereby irrevocably transfer exclusive copyright for this material to IARIA. I give IARIA permission or reproduce the work in any media format such as, but not limited to, print, digital, or electronic. I give IARIA permission to distribute the materials without restriction to any institutions or individuals. I give IARIA permission to submit the work for inclusion in article repositories as IARIA sees fit.

I, the undersigned, declare that to the best of my knowledge, the article is does not contain libelous or otherwise unlawful contents or invading the right of privacy or infringing on a proprietary right.

Following the copyright release, any circulated version of the article must bear the copyright notice and any header and footer information that IARIA applies to the published article.

IARIA grants royalty-free permission to the authors to disseminate the work, under the above provisions, for any academic, commercial, or industrial use. IARIA grants royalty-free permission to any individuals or institutions to make the article available electronically, online, or in print.

IARIA acknowledges that rights to any algorithm, process, procedure, apparatus, or articles of manufacture remain with the authors and their employers.

I, the undersigned, understand that IARIA will not be liable, in contract, tort (including, without limitation, negligence), pre-contract or other representations (other than fraudulent misrepresentations) or otherwise in connection with the publication of my work.

Exception to the above is made for work-for-hire performed while employed by the government. In that case, copyright to the material remains with the said government. The rightful owners (authors and government entity) grant unlimited and unrestricted permission to IARIA, IARIA's contractors, and IARIA's partners to further distribute the work.

Table of Contents

An Application for Protecting Personal Information on Social Networking Websites

Mehmet Erkan Yüksel
Computer Engineering Department
Faculty of Engineering
Istanbul University
Istanbul, Turkey
e-mail: eyuksel@istanbul.edu.tr

Asım Sinan Yüksel
Computer Science Department
School of Informatics and Computing
Indiana University
Bloomington/Indiana, USA
e-mail: asyuksel@umail.iu.edu

***Abstract*— We redundantly share our personal information and applications with people on the Internet. Depending on this, social networking websites have also become indispensable parts of our lives and allow the users to share just about everything: photos, videos, favorite music, and games. Sharing large amounts of information causes privacy problems for the users in these websites. In order to prevent these problems, we can provide trusted and built-in applications that help to protect our privacy by limiting the friends who get access to our personal information and applications. Thus, the security and privacy problem has prompted us to provide a solution that offers the users of these social networking websites an opportunity to protect their information. In this paper, an application that can be used in social networking websites, its design, algorithm and database structure are mentioned. Our application offers a trusted architecture to the social network users. It finds social circles and helps the users to group their friends easily and meaningfully for protecting their privacy and security. This system provides grouping of users through an automated system into different social circles by analyzing the user's social situation and depending on what common information or application they would like to share that should not be accessed by other users.**

***Keywords*— social networking websites; clustering; sharing information; protecting privacy; graph database**

I. INTRODUCTION

Meeting new friends and socializing are parts of our lives. With great advances being made at this age of information technology, socialization has greatly increased with people being able to meet and communicate friends from different regions of the world by using social networking websites. These websites enable friends to easily communicate online, and provide many Internet features and functionalities for social network users such as publishable personal profiles, repositories for sharing information and applications, and the abilities to provide social connectivity between the users. Although social networking websites in the Internet offer an opportunity to meet and communicate many friends; it creates a privacy and security problem because people of all ages, interests and backgrounds have free access to social networking websites, and you may not want to share some of your personal information with some of your friends or network users who you do not know. In these websites, there are a number of cases where the users have been able to identify and locate other users through the personal information that was posted. Inappropriate information might be published that leads computer hackers, sexual predators and other malicious users to alter the person's profile and information or to access their computer. Users can find damaging information about a person's past and they can learn what he/she is doing on the Internet. Therefore, the users must allow as many or as few friends to view their personal web pages by choosing some kind of restrictions. They must determine the accessing permissions using tools on the website.

Building personal web pages and using social networking technologies, services and applications can be a very creative, useful, effective and beneficial outlet for users to share and express their thoughts and opinions, to learn how to manipulate and use large amounts of information, and to learn skills needed to build web pages and applications. Most popular examples including Facebook, Twitter, MySpace, and Hi5 are public social networking websites offering free accounts to the users to share personal information such as "About Me", "My Friends", sexual orientation, emails, message boards, religion, politics, user groups, favorite tunes, movies/videos, interests, preferences, education achieved, networking organizations, photographs, applications and other information about themselves. However, social networking websites have potential effects on people's life, and there are very serious privacy issues when these websites are not used appropriately.

Personal information like your profile that is posted on a social network can be accessed by all your friends that you share the network with. Unauthorized people may also get access to some of your personal information that you do not want to share. We must know what is appropriate to put on the web pages, and be clear about what is not safe to post on the web: full name, address, specific places we go, phone numbers, ethnicity, and anything else that would help someone identify or locate us. Once something is posted on the web, it is no longer private [1, 2].

 ISBN: 978-1-61208-100-7

Social networking websites increase popularity of the Internet usage, for the purpose of socializing and networking with users across the world, and they are becoming a growing issue of concern for researchers. Therefore, protecting privacy, sharing information and applications in social networking websites are really important issues. These websites provide some features for protecting privacy, and controlling what information can be accessed. However, most people are unaware or do not know how to use these features. Even if users were to perform these tasks of categorization, on what basis would they categorize their friends in a meaningful way to set privacy and security policies? Our study proposes an application to help the users to make better decisions about their privacy settings.

The remainder of this paper is organized as follows: In Section 2, we provide a literature review highlighting the works already carried out in this area, explain what we want to achieve, and reveal what was/is missing. We present the details of our application, application design platform, algorithm design, clustering approach and graph based database design in Section 3. In conclusion, we discuss the future directions, limitations, contributions of our study.

II. RELATED WORK

Social network theorists have discussed the relevance of relations of different depth and strength in a person's social network. In a recent study [3], the privacy relevance of these arguments has recently been studied and researchers concentrated on the role and importance of social connections as we call social circles. In a study by [4], researchers reveal the relation between personal information, privacy and a user's social network. They state that a social network provides a visual map of the relevant social connections between the nodes of participation which can be used to measure the degree of connectivity. This work is one of the studies we inspired and supports our idea of protecting personal information by creating social circles with their crucial explanation "Safety must be first and foremost because we want to share information about ourselves to be known only by a trusted circle of close friends, and not by anonymous strangers or distant friends who does not know us better."

In [5], researchers studied the information disclosure in social networks, and they found that by looking at certain characteristics, such as knowing which groups people belong to or their favorite applications, it was possible to predict their political affiliation.

In [6], Canadian Privacy Commissioner published a must-read report about personal information protection on Facebook. This report clearly supports our idea of improving and simplifying the privacy, but it does not go beyond further than being a criticism. We believe that our study will inspire Facebook developers to implement more user friendly, more successful privacy management features.

All of the recent researches show the importance of protecting information in social networks. Lack of the privacy in social networks causes some members to unregister so as to protect their privacy. Our study differs from recent studies. Instead of proving the existence of privacy problems and presenting attacks, we proposed a solution and its implementation for current problems that social network users encounter.

III. APPLICATION

Our application provides an implementation of a web based solution to protect personal information. It helps the users to automatically categorize a large number of friends into meaningful lists. The main assumption we make to build the social circles is that users would mostly present similar information to all friends in a social group, and therefore social circles provide a meaningful and trusted categorization of friends for setting privacy policies [7, 8].

Our application interface design has two aims. The first aim is to discover whether social circles exist on a social networking website. The second aim is to discover whether these social circles would help the users in social networking applications in setting effective privacy and security policies. In our system, we have developed a trusted application which is shown in Fig. 1 to identify the social circles in social networking websites. The users can add this application to their personal web pages on any social networking website (e.g., Facebook, Twitter or MySpace). The users have been asked randomly generated questions about their willingness to share a piece of their information with a social network friend of theirs. These questions are based on the fields of social networking website database tables that are available for application developers. Each question is formed in a way which does not reveal the real aim of the study, and does not disturb the users. This is to prevent the bias such as evaluating the concept of trusted social circles in the context of privacy and security. The answers to the questions are saved in our secure, anonymized graph database. This data collection method provides us with quantitative results that we can statistically analyze. When all questions are answered, the application runs the clustering algorithm and finds the visual graph of users.

Figure 1. Main Page of Our Application

We have developed a web application which finds social circles of the users in their social networks. Users can add this application to their personal pages which are stored on social networking websites such as Facebook, Twitter, and MySpace. Users are able to delete this application after they have completed their studies. Our application is built on a trusted structure and suitable for protecting privacy. It provides the following features:

1) Creating Visual Graph of Social Circles

As shown in Fig. 2, our application produces such a graph that helps users to see each social circle and to make better decisions about their applications and privacy settings.

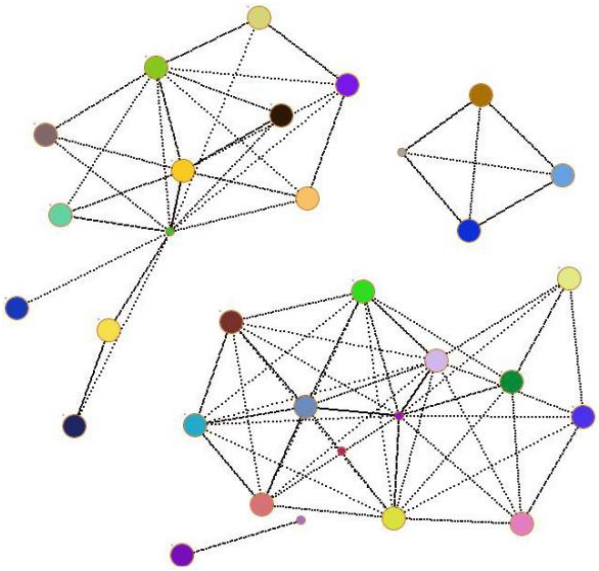


Figure 2. Visualizing Users' Social Graphs.
(Friend pictures are anonymized by using circles)

2) Suggested-Settings For Proteting Privacy

Our application suggests the set of friend lists that users should create, and the friend lists into which they should put each of their current friends based on the identified social groups. It is also designed to attach importance to user's privacy and security on social networking websites.

3) Graph Database For Effective Data Representation

We used a simple, robust, massive scalability, and convenient object-oriented graph database structure that provides an intuitive graph-oriented model for data representation and collection. This database is an embedded, disk-based, fully transactional, more effective and flexible system that stores data structures in graphs rather than in tables. Instead of static and rigid tables, rows and columns, our application works with a flexible graph network consisting of nodes, relationships and properties.

A. Application Design Platform

Applications for social networking websites can be created by using a variety of software technologies, including HTML, XML, OpenSocial Templates, JavaScript, CSS, Flash, Python, Java, Perl, PHP, .NET, or Ruby on Rails. This section gives us several approaches for developing our application, depending on our project requirements.

Most social network application designs in the Internet have similar structures: social network data, application data, and an appropriate template are used to provide a rendered view to the user. In many social networking websites, these components can come from several places. Client-side applications can use scripting language such as JavaScript and VBScript to render data into a template. Social networking websites can store both social network and application data, and server-side applications can take advantage of databases and server-side frameworks to produce rendered output.

Our application runs inside of a social networking website such as Facebook and MySpace but relies on an external server or host for processing and rendering data. It can provide advanced functionality but may run into scaling problems when the users increase so much. Its platform consists of some hardware and software components. These components are given below:

- A markup language derived from HTML
- General-purpose scripting languages PHP
- JavaScript scripting language
- MySQL database language for interacting with social networking website database.
- Object-based web API for handling communication between a social network site and our application.
- A set of client libraries (ASP.NET, C++, C#, PHP, Python) for different programming languages.

For our application, we used Linux Fedora Operating System Version 12.0, an open source JavaScript library [9] to draw the edges and nodes, Social Network API to gather necessary information to draw edges and nodes. PHP language is chosen as a server side technology to query database, run the clustering algorithms, and display the results on the social networking website. Our application can be embedded within a social networking website itself, or access a website's social data from anywhere on the Internet.

B. Algorithm Design

Our algorithm consists of two phases. In the first phase, we create the nodes for the users. In the second phase, we create and draw the connections between the nodes to determine the relationship and privacy between users who are registered on a social networking website.

The algorithm collects information such as friends' ids. This structure successfully detects social circles if the users choose to share the similar combination of personal information with friends in the same social circle, and if they choose different combinations with friends in other social circles. By using more data collected from our application, we have been finding out the effectiveness of our algorithm.

1) Creating Nodes

In this phase, we create all nodes of the graphs that we are going to draw. The algorithm for creating the nodes is shown in Fig. 3. In our node creation algorithm, we first go through all friends of the user and create nodes for each

 ISBN: 978-1-61208-100-7

friend. Then, for each friend, we go through all mutual friends and create nodes for each mutual friend. By saying mutual friends, we mean the common friends of the user with a user's friend.

```
for ( i = 0 ; i < Total_Friends ; i++)
{
Create_Node ( friend_ids [i] );
    for ( j = 0 ; j < Total_Mutual_Friends[i] ; j++)
    {
    Create_Node ( mutual_friends_id[i][j] );
    }
}
```

Figure 3. Node Creation Algorithm

2) Creating Edges

In this phase, we create the connections between friends of user and between mutual friends of user. By using the nodes that we created in the first phase of the algorithm, we add the edges according to the following algorithm shown in Fig. 4. In edge creation algorithm, we go through all friends of the user and find out if the friends are friends with each other. If they are friends, we add an edge between those friends. At the same time, we go through the mutual friends of the user and find out if they are friends with each other. If the mutual friends are friends with each other, we again add an edge between those mutual friends.

```
for ( i = 0 ; i < Total_Friends ; i++)
{
friend1 = friends_id[i];
    for ( j=0 ; j < Total_Friends[i] ; j++)
    {
    friend2 = friends_id[j];
        if ( friends.arefriends (friend1,friend2))
        {
        AddEdge(friend1,friend2);
        }
    }
    for ( j = 0 ; j < Total_Mutual_Friends[i] ; j++)
    {
    mutual_friend1 = mutual_friends_id[i][j];
    for ( k = 0 ; k < Total_Mutual_Friends[i] ; k++)
    {
    mutual_friend2 = mutual_friends_id[i][k];
        if(friends.arefriends(mutual_friend1,mutual_friend2)
        {
        AddEdge(mutual_friend1,mutual_friend2);
        }
    }
    }
}
```

Figure 4. Edge Creation Algorithm

Fig. 5 shows the output of node and edge creation algorithm. The colorful circles are the nodes that represent the social network user's friends, and the black lines are the edges that represent the friendship relation.

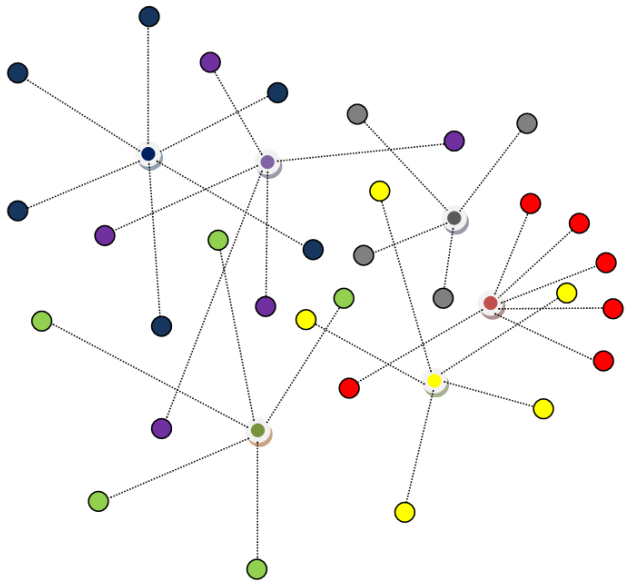


Figure 5. Output of Node and Edge Creation Algorithm

C. Clustering Method To Determine Social Circles

Methods for clustering have been deeply studied. But our aim is not to study them. Clustering is just one of the steps to achieve our privacy goal. Our main aim is to use the right clustering algorithm for social networks and develop an application to provide privacy by adapting this clustering algorithm to our application.

The clustering in social networks requires grouping users into classes based on their attributes, properties of personal relationship, web page links, spreads of messages and other applications. It is the process of organizing users into groups whose members are similar in some way. Our algorithm is different from other clustering algorithms, and it can dynamically group users in a social network into different classes based on their properties and effectively identify relations among classes. It collects some data which are similar between social network users and are dissimilar to the users belonging to other groups. It creates active cells like network grids and builds visual graph of social groups. The similar structure applied in the algorithm [8, 10] for finding (α, β) clusters has been used in our algorithm. Friends sharing common personal information are the adjacent nodes to α-fraction. The α-fraction represents the cluster that has a large density. On the other hand those friends not sharing common personal information are the adjacent nodes to β-fraction. The β-fraction represents the cluster that has a low density. It is therefore possible to use the social graph of network users as an input to our algorithm. One might ask that what if a friend belongs to more than one group? For example, a user can have a friend from high-school or university that is currently his/her work mate. The overlapping sets or being

 ISBN: 978-1-61208-100-7

in more than one group does not cause a problem from the privacy perspective. Our application groups friends according the common information that a user wants to share with his/her friends. For example, if we just want to share our photos and status with our college friends, then we will be showing them a profile where they will only able to see our photos and status. If there are some other friends that we just want to share our photos and status, they will also be in this group. Therefore, it is perfectly normal and possible that a person can be in one or more social circles. The user will show some information in one group and different information in another group. In other words, we limit who sees what. Fig. 6 shows our pseudo-algorithm of our clustering process for the users in a social network.

1. Write all answers of the users to DB
2. Select User's answers from DB and create result_array
3. FOREACH (result_array as value)
    3.1. Get selected friends' friendids for each question
    3.2. Create [question no, friendids] array
4. FOR i=0 to size (result_array)
    4.1. FOR k=i+1to size (result_array)
    4.2. Create the clustering_array[][]
5. Sort (clustering)
6. Create unique values for clustering_array[][]
7. Find how many times a friend is chosen in 10 questions
    7.1. Eliminate the friend: IF NOT a friend chosen >= 3 times
8. Find Min (set of information that the user wants to share)
9. Eliminate the sets: IF sets do not contain mutual friends
10. Display (Groups or circles)
11. Suggest (Privacy Settings)

Figure 6. Clustering Algorithm

D. Database Design

Building the visual graph of a social network user is an expensive task. Instead of creating the graph while executing the social network API calls, we decided to store the necessary information in our own database. The main reason to use our own database is because having too many API calls causes time outs. Another important reason is the difference between our database design and the social networking websites. Current social networking websites use relational databases to store social network data. For better performance, more effective querying, to extend our work and develop a knowledge based approach, we used graph database.

1) Graph Database Design

In graph based databases, information is stored as nodes, edges and properties. Since social networking data has similar properties, graph database is the powerful way of representing social relationships between people. In our application, we used Neo4j, an open source graph database. According to developers' of Neo4j [11], it is an embedded, disk-based, fully transactional Java persistence engine that stores data structured in graphs rather than in tables. More importantly, it includes the database features such as ACID transactions, durable persistence, concurrency control, transaction recovery, and other features of enterprise-strength databases. The following figures show our transition from relational database to Neo4j graph database. As it is seen in Fig. 7 and Fig. 8, it is very easy to see the connection between two people. However, in a relational database, it is hard to see who is friend with whom. In addition to this, whenever we introduce a relationship such as mutual-friends relationship, we need to add one more table to represent this relationship. As a result, the number of table joins increase and the performance decreases.

ID	Name	Sex	Age	Relationship Status
0001	Asim	Male	27	Single
0002	Erkan	Male	30	Single

User Table

ID1	ID2
0001	0002
0002	0001

Friendship Table

Figure 7. Our Social World Modeled in Relational Database

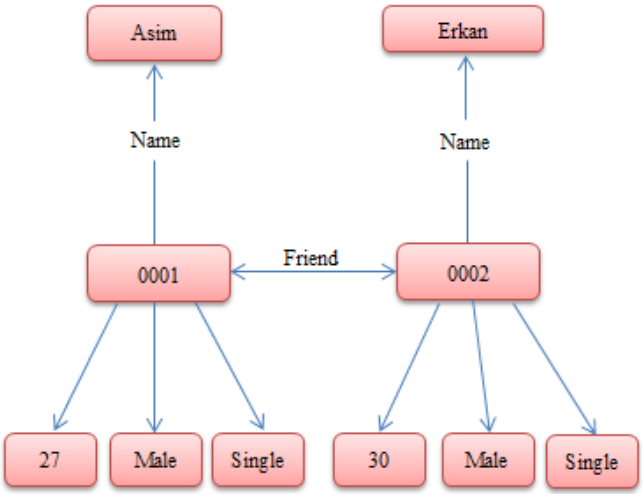


Figure 8: Our Social World Modeled in Neo4j Graph Database

2) Graph Database vs. Relational Database

Relational databases are around for many decades. They are the database choice of most traditional data-intensive storage and retrieval applications. SQL language is used to retreive the data. Relational databases are not efficient, if data contain many relations and require many joining of tables which are expensive operations. Thus, graph base database has better performance than relational database when representing relations. Recent studies by [12, 13] provide a detailed performance evalution of MySQL database and Neo4j. According to their results, graph

 ISBN: 978-1-61208-100-7

database is more flexible, easy to program, and perfoms better.

IV. CONCLUSION

Most of the related work present attacks for social networks and they do not provide a useful solution to protect privacy. We believe that our study is the first study that contains an implemented application for social network privacy. This system is a secure web application for social networking websites such as Facebook, Twitter, MySpace, and it includes an implementation of our original idea. Currently, it is running on Facebook. As a future work, we are planning to develop a general API that can be applicable for any social networking website such as Twitter and MySpace.

Recently, we have been inviting social network users to our study and collecting data. Furthermore, we are helping the users to get acquainted with our application. After collecting enough data, we will evaluate the effectiveness of our approach.

Our study uses a combination of clustering approaches. Firstly, the users are grouped according to their friendship relations (i.e., by using friendship and mutual friendship queries). Secondly, we group them based on the information that a user wants to share with his/her friends. The second one is the heart of grouping, since it will provide the privacy. Privacy is provided by showing different profiles to different combination of groups. For example; if a user wants to share his/her relationship status, photos, date of birth with his/her Friend-A and Friend-B, then Friend-A and Friend-B only see this information. Therefore, we are able to limit who sees what.

Although we successfully create the social graph of a user, we have limitations which affect the performance of our application. Our social graph visualization algorithm works for a subset of friends and mutual friends. We limited the number of friends and number of mutual friends that will participate in our study. The reason behind the limitation is because of large amount of social network API calls. There are millions of social network developers who are querying social network servers, and these queries cause a delay in response time. Drawing the social graph of a user and displaying it takes more time. Sometimes, the queries are even dropped because of the delay, and the graph is not drawn.

In this study, we proposed an application to identify the social circles of the users by using graph database system. In order to see the effectiveness of our algorithm we have been testing our application. As future work, we also want to develop a knowledge base system to provide intelligent decisions about sharing of personal information with people.

REFERENCES

[1] E. Hooper, “Intelligent strategies and techniques for effective cyber security, infrastructure protection and privacy” The 5th International Conference for Internet Technology and Secured Transactions (ICITST-2010), London, UK, 2009, pp. 1-7.

[2] A. Beach, M. Gartrell, and R. Han, “Solutions to Security and Privacy Issues in Mobile Social Networking”, The 12th IEEE International Conference on Computational Science and Engineering (CSE '09). Vancouver, Canada, 2009, pp. 1036 – 1042.

[3] B. Zhou, J. Pei, “Preserving Privacy in Social Networks Against Neighborhood Attacks”, The 24th International Conference on Data Engineering (ICDE’08), Cancún, México, 2008, pp. 506-515

[4] R. Gross, A. Acquisti, “Information Revelation and Privacy in Online Social Networks (The Facebook Case)”, ACM Workshop on Privacy in the Electronic Society, Alexandria, VA, USA, 2005, pp. 71-80

[5] J. Lindamood, R. Heatherly, M. Kantarcioglu, and B. Thuraisingham, “Inferring private information using social network data.”, The 18th International Conference on World Wide Web, Madrid, Spain, ACM 978-1-60558-487-4, 2009, pp. 1145-1146.

[6] E. Denham, “Report of Findings into the Complaint Filed by the Canadian Internet Policy and Public Interest Clinic (CIPPIC) against Facebook Inc. Under the Personal Information Protection and Electronic Documents Act”, http://www.priv.gc.ca/cf-dc/2009/2009_008_0716_e.cfm, Last accessed: July 18, 2010.

[7] E. Zheleva and L. Getoor, “To Join or Not to Join: The Illusion of Privacy in Social Networks with Mixed Public and Private User Profiles”, The 18th International World Wide Web Conference, Madrid, Spain, ACM 978-1-60558-487-4, 2009, pp. 531-540.

[8] F. Adu-Oppong, C. K. Gardiner, A. Kapadia, and P. P. Tsang, “Social Circles: Tackling Privacy in Social Networks (Poster Abstract)”, The 4th Symposium on Usable Privacy and Security (SOUPS), Pittsburgh, PA, USA, 2008.

[9] K. Scholz, “Using Force Directed Graphs in Your App.” http://www.kylescholz.com/blog/2006/06/using_force_directed_graphs.html, Last accessed: July 18, 2010.

[10] N. Mishra, R. Schreiber, I. Stanton, and R. Tarjan, "Clustering Social Networks", The 5th International Conference on Algorithms and Models for the Web-Graph, San Diego, CA, USA, 2007, pp. 56-67.

[11] Neo4j, http://neo4j.org/, Last accessed: July 18, 2010.

[12] M. A. Rodrigez, “MySQL vs. Neo4j on a Large-Scale Graph Traversal.”,http://markorodriguez.com/Blarko/Entries/2010/3/29_MySQL_vs._Neo4j_on_a_Large-Scale_Graph_Traversal.html, Last accessed: July 18, 2010.

[13] C. Vicknair, M. Macias, Z. Zhao, X. Nan, Y. Chen, and D. Wilkins, “A Comparison of a Graph Database and a Relational Database”, The 48th ACM Southeast Conference, Oxford, Mississipi, USA, 2010.

 ISBN: 978-1-61208-100-7

Improvement of Channel Decoding using Block Cypher

Natasa Zivic
Institute of Data Communications Systems
University of Siegen
Siegen, Germany
natasa.zivic@uni-siegen.de

Ayyaz Mahmood
Institute of Data Communications Systems
University of Siegen
Siegen, Germany
ayyaz.mahmood@uni-siegen.de

***Abstract*—This paper introduces two methods for the improvement of performance of channel coding using cryptography, based on concatenation of codes. Cryptography as an outer code is combined with channel coding as an inner code. The first method improves decoding of cryptographic functions. The second one uses the first method for improvement of information decoding using a block cipher Advanced Encryption Standard. Computer simulation results are included.**

Keywords- *Advanced Encryption Standard, Cryptography, Concatenated Codes, Soft Input Decryption, Encryption, Maximum A Posteriori Probability (MAP)*

I. INTRODUCTION

This paper researches the interoperability between channel coding and cryptography in order to reduce BER of the channel decoding. Therefore, soft output or so called *L*-values of SISO (Soft Input Soft Output) channel decoding are used for correction of the input of inverse cryptographic mechanisms. The channel code can be considered as an inner code and the output of the cryptographic mechanism as an outer code (Fig. 1).

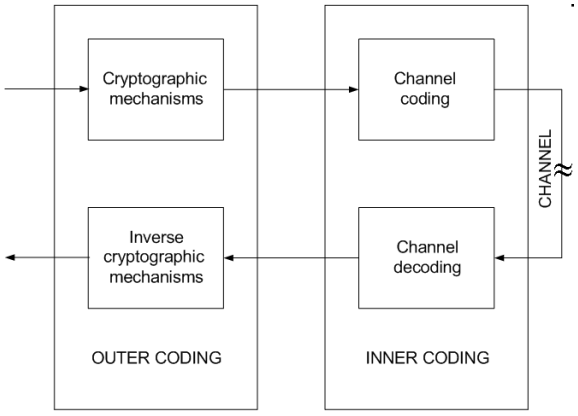


Figure 1. Cryptography and channel coding as concatenated codes.

Cryptographic mechanisms are used for the recognition of modifications by errors or manipulation. Soft output of the channel decoder enables cryptographic mechanisms to perform error corrections by Soft Input Decryption [1].

The following problems are investigated and solutions for them are proposed:

1. Improving cryptography using *L*-values of channel decoding (solution: Soft Input Decryption) – explained in Chapter II of the paper
2. Improving channel decoding using *L*-values and avalanche effect of error spreading [2] by wrong input of a decryptor (solution: Improving channel decoding using block cipher) – explained in Chapter III of the paper. AES is used as a block cipher because it is one of the most widely accepted block ciphers [3].

II. SOFT INPUT DECRYPTION

Soft Input Decryption (SID) improves decrypting mechanisms using soft output of the channel decoder [1]. A decryptor is used for verification of cryptographic check values.

Algorithm of SID is as follows:

The security mechanism is successfully completed on the receiving side if the verification results is positive. In case of negative verification, the decryptor analyzes soft output of the channel decoder, changes the bits with the lowest $|L|$-values, performs the verification process and checks the result of the verification again.

If the first verification after starting Soft Input Decryption is not successful, the bit with the lowest $|L|$-value flipped, assuming that the wrong bits are probably those with the lowest $|L|$-values. If the verification is again negative, the bit with the second lowest $|L|$-value is changed. The next try will flip the bits with the lowest and second lowest $|L|$-value, then the bit with the third lowest $|L|$-value, etc. The process is limited by the number of bits with the lowest $|L|$-values, which should be tested. The strategy follows a representation of an increasing binary counter, whereby the lowest bit corresponds to the bit with the lowest $|L|$-value, etc.

If the attempts for correction of cryptographic check values fail, the number of errors is too large as a result of a very noisy channel or an attack, so that resources are not sufficient to try enough combinations of flipping bits of low $|L|$-values.

III. CHANNEL CODING USING BLOCK CYPHER

This chapter explains the method of improving channel decoding using *L*-values and avalanche effect of error proposed error correction improvement scheme.

 ISBN: 978-1-61208-100-7

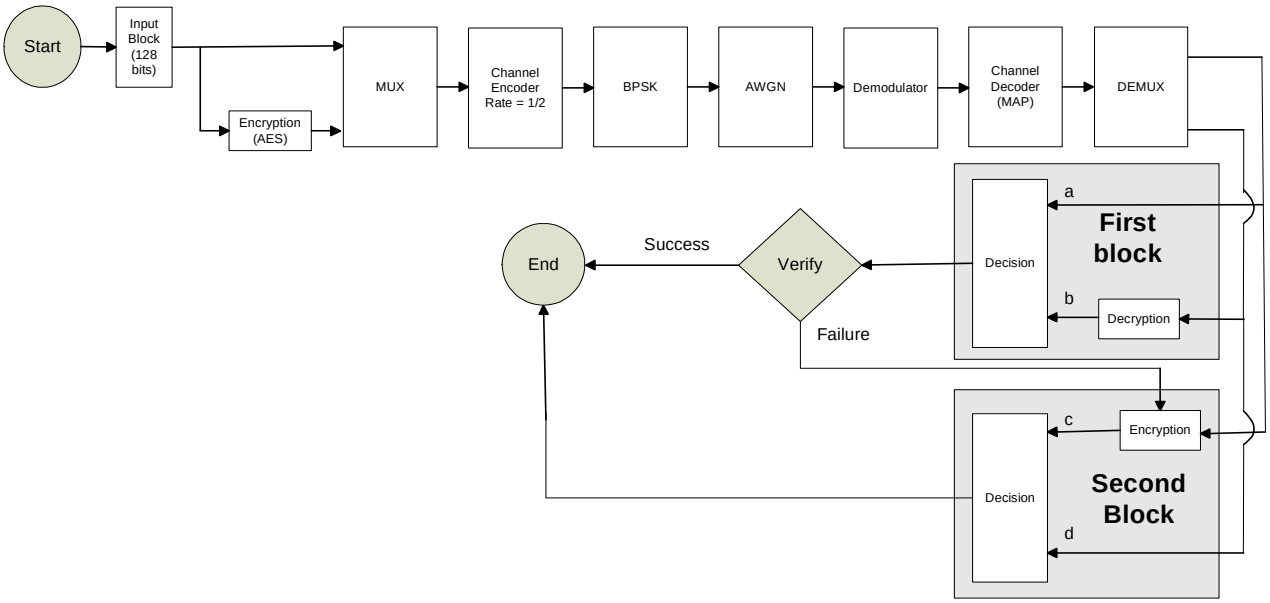


Figure 2. Error correction system using block cypher.

Fig. 2 shows the proposed error correction improvement scheme. The two blocks, which are named as first block and second block, are actually responsible for error correction improvement. The second block comes into operation in the case that first block is not able to do error correction.

Scheme in Fig. 2 includes decision blocks, which are used for making decision between two inputs applied to them. The decision block selects only one input which has the minimum number of errors.

The AES is a symmetric key (uses the same key for encryption and decryption) block encryption algorithm. The AES block size is 128 bits and that is the reason for using a block of 128 input bits in the simulations. Fig. 2 shows that in the case of the first block, decryption is used because the input data was encrypted. If the first block is not able to make error correction, the second block comes into operation. In the second block, the outputs 'c' and 'd' will also be compared to perform error correction in the case that first block is not able to do it.

A. Error Correction considering the First Block

The two outputs 'a' and 'b' applied to decision block have the following possibilities:

1) Both have errors
2) The output 'a' applied to decision block is error free whereas the output 'b' applied to decision block has about 50% errors (avalanche effect)
3) The output 'b' is error free whereas the output 'a' has errors.

The first block will be able to improve error correction considering all of the above possibilities. The output 'b' in lower branch exhibits avalanche effect because of the use of AES. It means that if MAP decoder is not able to correct all errors, then the output 'b' will have about 50% errors. The output 'a' will have significantly smaller number of errors as compared to the output 'b'. Therefore decision block will always compare output 'a' and output 'b' to check if this difference is above a certain value. This value depends upon the signal to noise ratio and is named threshold. The decision block calculates a value, which is called BER_compare for each iteration. It is calculated as a difference between BER of the output 'a' and BER of the output 'b'. If BER_compare for each iteration is higher than the threshold, then the output 'b' has about 50% errors. In this case SID is used for achieving error free output 'b' (if SID is successful). If BER_compare is lower than the threshold, then the output 'b' is error free; the decision block will select the output 'b'.

B. Soft Input Decryption using AES Block Cypher

Soft Input Decryption using AES is able to correct all errors (if it is successful) occurring after decryption at output 'b' by taking the output 'a' as a reference. It uses soft output of the channel decoder. As the magnitude of *L*-value gives the reliability of the decision, it can be used to correct all erroneous bits at output 'b'. Soft Input Decryption using block cipher AES uses the lowest sixteen *L*-values, which means that SID will have 65536 attempts for error correction. In each attempt a bit or a combination of bits are flipped (0 to 1 or 1 to 0) and then decryption is performed. For each attempt BER_compare is calculated and compared with the threshold until it becomes less than the threshold.

When BER_compare is less than the threshold, all of the errors at output 'b' are corrected. The decision block will then select the output 'b' because it is error free. It can also happen that within 65536 attempts, SID is not successful. In that case second block comes into operation.

 ISBN: 978-1-61208-100-7

C. Error Correction considering the Second Block

If Soft Input Decryption is not able to correct errors in the first block, the second block attempts to achieve it. In the case of the second block, upper branch is encrypted after MAP decoder, so the avalanche effect will be present at output 'c'. The decision block will therefore treat output 'c' exactly like output 'b' and output 'd' exactly like output 'a'. The error correction can be done in the same way as it was performed for the first block. Instead of SID, the second block performs Soft Input Encryption. If BER_compare is higher than threshold, the lowest sixteen $|L|$-values will be flipped at the input of the encryption block until all errors are corrected (if Soft Input Encryption is successful).

IV. SIMULATION RESULTS

It is explained that the improvement in error correction can be achieved using Soft Input Decryption and Soft Input Encryption which depends upon the threshold. The simulated curve for threshold vs. E_b/N_0 is shown in Fig. 3. The curve shows that threshold decreases with the increase of E_b/N_0. The reason is that with the increase of E_b/N_0, the channel introduces fewer errors.

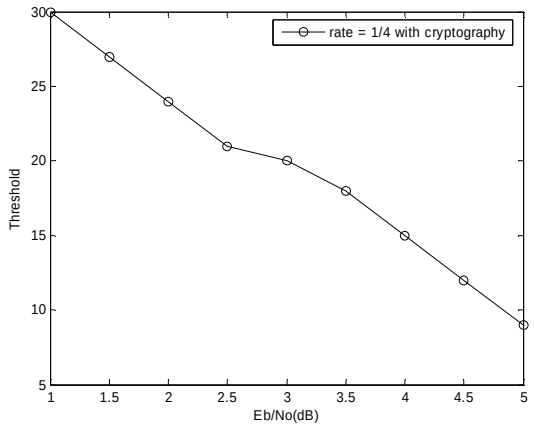


Figure 3. Threshold versus E_b/N_0 for error correction system using cryptography.

The proposed system shown in Fig. 2 has an overall code rate of 1/4, if it compared to a standard error correction system without cryptography having a convolutional encoder of rate 1/4. The convolutional encoder of rate 1/2 is a non-systematic (2,1,3) convolutional encoder and a convolutional encoder of rate 1/4 is a non-systematic (4, 1, 3) convolutional encoder [4]. These two encoders were selected because they have the same coding gain and the similar structure, which enables fair comparison of decoding results [4].

BPSK modulation, AWGN channel and MAP [5] convolutional decoder are used in simulations. For purposes of Soft Input Decryption / Encryption, maximum 16 lowest L-values are used (2^{16} correction trials).

Fig. 4 shows that the error correction system using cryptography achieves considerable coding gain over 1/4 convolutional decoder: 1.3 dB for BER of 10^{-6} and 1.85 dB for BER of 10^{-7}.

For E_b/N_0 higher than 2.4 dB, there is a coding gain of the error correction system presented in Fig. 3, which increases with increase of E_b/N_0 in comparison to 1/4 convolutional decoder. For E_b/N_0 lower than 2.4 dB, presented error correction system gives worse decoding results than the comparable 1/4 convolutional decoder.

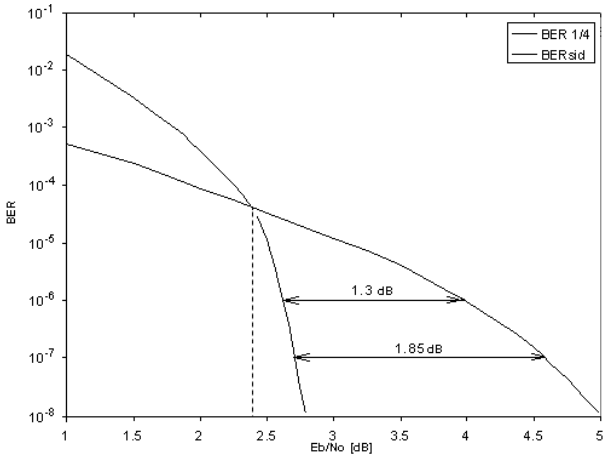


Figure 4. BER versus E_b/N_0 for error correction system with and without cryptography.

V. CONCLUSION

The paper introduces two methods of interoperability of channel coding and cryptography: improving decryption using Soft Input Decryption and improving decoding results using block cipher and principles of Soft Input Decryption. The characteristic of cryptographic check values to give about 50 % of wrong bits at output of decryptor if one or more bits at input of decryptor are wrong, is used for bit error correction of decoded information. Bit error correction scheme with two blocks (for Soft Input Decryption and Soft Input Encryption) is presented and simulated.

Simulation results show that, if 16 lowest L-values are used for Soft Input Decryption / Encryption, a remarkable coding gain in comparison to the standard 1/4 convolutional decoder can be achieved for E_b/N_0 higher than 2.4 dB: for BER of 10^{-6} coding gain is equal 1.3 dB and for BER of 10^{-7} coding gain achieves 1.85 dB. For E_b/N_0 lower than 2.4 dB, 1/4 convolutional decoder is stronger in error correction than the presented error correction scheme.

REFERENCES

[1] N. Zivic, C. Ruland, "Soft Input Decryption", 6th Source and Channel Code Conference, VDE/IEEE, Munich, vol. April 2006.

 ISBN: 978-1-61208-100-7

[2] S. Fernandez-Gomez , J. J. Rodriquez-Andina, E. Mandado, “Concurrent error detection in block ciphers”, in Proc. IEEE Int. Test Conf., Atlantic City, NJ, 2000, pp. 979-984

[3] National Institute of Standards and Technology, Advanced Encryption Standard (AES), Federal Information Processing Standard FIPS PUB 197, November 26, 2001

[4] S. Lin, D.J. Costello, “Error Control Coding”, Pearson Prentice Hall, USA, 2004

 ISBN: 978-1-61208-100-7

Distortion Free Steganographic System Based on Genetic Algorithm

Heshem A. El Zouka
Department of Computer Engineering, College of Engineering and Technology
Arab Academy for Science & Technology and Maritime Transport,
Alexandria, Egypt

helzouka@aast.edu

Abstract — **Steganography is the art of secret communication. Its purpose is to hide the very presence of communication as opposed to cryptography whose goal is to make communication unintelligible to those who do not possess the right keys. The research work in this paper shows that most of the steganographic systems proposed during the past decades usually substitute the insignificant parts of the image with the secret message. However, these systems don't pay attention to these parts, and the original image is distorted by some small amount of noise due the data embedding itself. This noise could reveal the existence of secret message and hence change the statistical profile of the cover image significantly. A simple attack such as Laplace filtering can exploit this fact and make the system detectable by the eavesdropper. In attempt to minimize the error introduced due to hiding foreign message into the cover image, a genetic algorithm approach will be employed efficiently in this paper in a way that examine the embedded bits. This has the advantage that the image remains nearly unchanged.**

Keywords - Security; Watermarking; Steganography; Information Hiding; Genetic Algorithm.

I. INTRODUCTION

Steganography is the art of secret communication. Its purpose is to hide the very presence of communication as opposed to cryptography whose goal is to make communication unintelligible to those who do not possess the right keys. By embedding a secret message into a digital image, a stego-image is obtained. It is important that the stego-image does not contain any easily detectable artifacts due to message embedding that could be detected by eavesdropper. There are many different steganographic methods that have been overviewed and analyzed by many researchers over the last few years. However, one common drawback of all current data embedding methods, is the fact that the original image is distorted by some small amount of noise due the data embedding itself. This noise could reveal the existence of secret message and hence, weakness the security value of the applied steganography method.

In this paper, we investigated most of the steganography methods which are based on substitution systems and have been proposed in the last few years. After analyzing the drawbacks in the analyzed steganography systems, we propose a steganographic system, which is based on genetic algorithm technology. The proposed system maintains such noise in a way that makes the transmitted signal undetectable. In addition, we built the software programs that provide the simulation results including the histograms of both the cover images and the stego images, and the variance between them. Also, the simulation results for all equivalent substitution techniques, which covered in this paper, are used here for comparisons purpose.

II. RELATED WORK

There are many different steganographic methods that have been proposed during the past decades. Most of the simple techniques can be broken by careful analysis of statistical properties of the channel's noise [1]. In substitution steganography technique [2], one can notice that: this method substitute insignificant parts of the image, e.g., the noise component of the cover with the secret message. These parts have specific statistical properties and the embedding process usually does not pay attention to them, and consequently change the statistical profile of the cover significantly. A simple attack such as "Laplace filtering" [3] can exploit this fact and detect the use of the steganographic system. In addition, these systems are extremely sensitive to cover modification and the attacker, who is not able to extract or prove the existence of a secure message can add a random noise to the transmitted cover in attempt to destroy the secret message.

Meanwhile, most of these stenography systems have a vital drawback, which is that the system doesn't discard image blocks where the desired relation of DCT coefficients for example [4] cannot be enforced without severely damaging the image data contained in this specific block. TCP/IP packet headers [5], [6] can also be reviewed easily. For example, firewall filters are set to test the validity of the source and destination IP addresses. Those filters can also be configured to catch packets that have information in supposed unused or reserved space. Based on the analysis of spread spectrum techniques [7], it can be observed that phase coding provide robustness against resembling of the carrier

 ISBN: 978-1-61208-100-7

signal, but at the same time it has a low data transmission rate. These techniques have a problem with the absolute phase of all following segment that followed the first modified one, since all of them will have a change that could be noticeable to the attacker. Moreover, at the receiver end; the embedding process is reversed and image restoration technique such as adaptive Wiener filter [8] is needed to estimate the original image.

III. SECURITY IMPLICATIONS AND SOLUTIONS

The proposed technique in this paper distorts the image insignificantly by making small modifications over a large number of pixels. We spread the secret message over a large area of cover image to produce a small modification on the carrier media. The new approach combines both Genetic approach and steganography to exchange secret messages in a way that it's impossible to discover without the knowledge of the cover image and the genetic algorithm that have been used.

Firstly, the image is divided into blocks and the parity bits from each block b1, b2,..,bi are computed and encoded with the corresponding bits in the text file which contains the secret message. The process is repeated for the whole selected blocks. If the computed parity bit pi and the secret bit mi are equal, then the encoded bit is zero and if the 2 input bits are different, then the output is one. Finally, the encoded bits are lined up to reconstruct the encoded file. Now, the file is ready to be encrypted and sent in any insecure channel to the receiver who had both, the secret key and a copy of the cover image which has been used. Therefore, the receiver of the encoded message will decipher the message using his secret key and the shared cover image.

3.1 ENHANCED EMBEDDING SOLUTION

After the encoding process had taken place, the output file was encrypted and sent directly to the other party as a cipher file. However, instead of sending an encrypted stream of bits, an alternative scheme can be adopted by injecting the stream of bits back to the cover image with a probability of 50% of changing the LSB of embedded pixels in target blocks using a fitness function. Our goal here is to embed one bit of the secret message m_i into one block of the cover image C, where C is composed of all the blocks {b1, b2,….,bi}, and since length of the message L(m) is less than the number of the target blocks N(b) the rest of the image can left unchanged. Moreover it's possible to select only some blocks bi in a rather random manner according to a secret key and leave the other unchanged. Therefore, the idea depends on spreading the secret message over the cover image using both a pseudo random number generator (PRNG) and a fitness function [9] that specifies one bit from each block of pixels randomly as follows:

$$P(I) = \sum_{j \in i} LSB(b_j) \bmod 2 \quad (1)$$

If the parity bit of one cover block bi doesn't match with the secret bit mi, the proposed genetic model will flip the LSB of one pixel in the block in attempt to make p(Ii) equals to mi and according to the employed fitness function which minimize the noise introduced to stego image as a result of embedding foreign bits of the secret message. Studying the properties of pixels surrounding the target pixel in a certain block by invoking a statistical fitness function that examine the number of 1's or 0's inside the chosen block will conceal the very existence of hidden information inside the stego image

3.2 EMPLOYED GENETIC ALGORITHM

Before communication starts, both sender and receiver have to agree on the location of the target blocks bi using a shared key value as the seed to a known PRNG algorithm. Each seed number can generate a set of random numbers, each of which allocates different locations of the blocks within the cover image. These blocks will be used as subjected pixels, from which the parity bits P(bi) are computed. The stego image file is then sent directly to the other parity with the encoded parity bits as mentioned before. The embedding process is preceded by extracting the parity bit through a reverse process to reconstruct the transmitted-hidden secret message.

However, the genetic algorithm is used to minimize the error due to hiding the foreign message carrier into cover images. The method is based on statistical analysis of images and their values are varying according to the applied key (seed) number. This is done by studying the neighboring pixels surrounding the chosen bit and changing its value to match the adjacent one in a way that prevent any statistical tracing. The Lina cover image in Figure 1 provides special features and will be used in this research work as a test image.

Figure 1. Lina Cover Image

The Genetic algorithm proceeds by dividing the image into blocks of 8*8 pixels and chooses the blocks in sequence according to a given seed number. The intensity of each

 ISBN: 978-1-61208-100-7

pixel x[i][j] within the chosen blocks is predicted according to the value of pixels in a specific neighborhood. Hence, the difference between the intensity of each pixel and its adjacent pixels is calculated as follows:

$$out[i][j] = in[i][j] - \frac{1}{64}\sum_{i:1-8}\ \sum_{j=1-8} x[i][j] \qquad (2)$$

where x[i][j] represents the pixel coordinates in the selected block region bi. For each tested pixel in the block, the average weighted sum of the surrounding pixels is computed and compared with the target pixel.

IV. IMPLEMENTATION AND DESIGN

In this section we will show how the Genetic algorithm is employed to search for suitable pixels within the blocks which by changing their least significant bits (LSBs) will introduce a lesser embedded noise to the stega image. If the system fails with one pixel within the target block another offspring is generated. The matching process will start over again till the program is succeeded in finding the flipped LSB which introduce lesser cost compared with the suggested noise threshold value. The genetic algorithm (GA) will finish the task either when it successfully finds the appropriate matched threshold value, or the closer one using different seed numbers. Finally, the program will print out the largest match between the generated offspring and the threshold value in a way that fulfil the objective of steganography which is concealment of existence. The whole proposed Genetic algorithm is illustrat6ed in the following statements:

4.1 MODEL DESCRIPTION

Chosen the target pixels in the selected blocks is a classic case of a combinatorial optimization problem. The intensity of each pixel is associated with a cost Cij. The cost of pixels is varying according to difference between the target pixels and the weighted sum of the surrounding pixels within one block. The objective is to obtain a solution with a minimum cost in terms of intensity difference and for each computed offspring; the elapsed time is computed in the provided searching algorithm. If n is the number of pixels in each tested block, then, the GA which has a behavior of cyclic approach will search for the fittest and optimal solution obtained by n!. It will be a huge number as the size of the block is increased, so the complexity of searching for the appropriate pixel within the block will not be suitable at all. However, the full optimality is not an a objective behind this paper. The employed genetic algorithm can be classified into 2 main parts based on the algorithm proposed recently by Ray et al. [10], which is called the modified order crossover (MOC) which in turn based on the order crossover (OCX) for solving Travelling Sales Person (TSP) problem [11]. The authors of these two algorithms used a new operation called the Nearest Fragment (NF) Heuristic and thy referred to their GA as (FRAG_GA). In addition the authors compared their results with other GA's that use different crossover methods, such as SWAP_GATSP [12] and OX_SIM [13]. We decided to choose an adaptive Ray et al. method on selecting the most suitable pixels in the target blocks which are subjecting to change in attempt to change the computed parity bit of each block separately. The adaptive algorithm is based on two main parts; swapped inverted crossover mechanism and the Fitness function as illustrated in the following sections.

4.1.1 SWAPPED INVERTED CROSSOVER

The main idea behind swapped inverted crossover (SIC) genetic algorithm is to find a better pixel on which its LSB is subject to change and introduce a minimum distortion to the image. The most suitable pixel will be chosen according to the crossover criteria which are based on the computed cost: difference between the intensity of the tested pixel and the weight sum of its neighborhood. Hence, the most suitable pixels from the target blocks are chosen with minimum introduced artifacts. The process is repeated with different seed numbers to all blocks populations in the tested image in attempt to calculate minimum value of all costs (Cij). The process of applying one, two, or both cutting points on the formulated population is illustrated in the following steps:

1. One point SIC - This is can be done by selecting on crossover point randomly to cut on. Suppose the two parents and a cut point from parent 1 is 4:
   a. Parent1 (1 2 3 4 5 6 7 8 9)
   b. Parent2 (5 1 4 6 8 9 2 7 3)

   where these numbers represent the cost of each pixel found in the target block for a given seed number.

2. The head of Paent1 will be flipped to be 4 3 2 1. By removing these point from Parent 2; the remaining pixels on parent 2 will be 5 6 8 9 7 to produce O1 and O2 offspring:

   O1 (4 3 2 1 5 6 8 9 7)
   O2 (5 6 8 9 7 4 3 2 1)

3. Doing the same procedure with parent 2, as cutting on point 6 and flip the head to be 6 4 1 5. Then removing these points from parent one. Similarly, by alternating output of parent 2 the offspring O3 and O4 will be generated:

   O3 (6 4 1 5 8 9 2 7 3)
   O4 (8 9 2 7 3 6 4 1 5)

4. The processes will continue until examine all generated outputs O5,O6, O7, and O8.

5. Two points SIC – where each parent will be cut to three pieces (head, middle, and tail) using two cut points (P1 and P2). For example 4 and 6 on parent1 & 6 and 9 on parent 2:
   Parent1 (1 2 3 4 5 6 7 8 9)
   Parent2 (5 1 4 6 8 9 2 7 3)

6. By flipping the parent 1 head to be 3 2 1 and flipping parent 1 tail also to be 9 8 7 to produce :
   O9 (3 2 1 4 5 6 7 8 9)
   O10 (7 8 9 4 5 6 3 2 1)

7. The same thing happens with the second parent producing O11 and O12.

4.1.2 THE FITNESS FUNCTION

After generating the above 12 offspring (O1 to O12) , the fitness function runs sequentially on all populations of the chosen blocks and the suitable pixels which are subjected to change with minimum cost can be chosen based on the following algorithm:

```
---------------------------------------------------------------------
i = 1
S0 = S
max = 0
while i <  n-1
      if Ci,j+1 > max
            Begin
                  max = Ci,j+1
                  x = i
            End
S1 ← swap pixelx with pixel1
S2 ← swap pixelx with pixel Lx; where L is the left pixel
.....
S8 ← swap pixelx with pixeln
S  ← max ( S0, S1,.....,S8)
Return S
End
---------------------------------------------------------------------
```

Clearly, the matching process will run, comparing the extracted parity bits with the embedded secret message bits. Therefore, this operation will be applied to the selected parents that will be copied several times. Each copy is then mutated using a different seed number with suitable neighbor pixels. The process proceeds till we find the minimum of these switching criteria and exchange it to match the original embedded bit. This operation is undertaken on one of the selected seed after the mutation operation is performed to produce the most suitable pixel which is subject to change and hence no statistical artifacts are introduced to the image. Using the same cover image, the recipient will treat each block separately by running the same algorithm reversely until the secret message is extracted.

A further improvement is possible if the sander and recipient generate a number of cover images to be verified with the embedded messages. By doing this we can minimize the searching time of the fitness function by about half, and increases the probability of finding the closest matches with the suggested threshold noise value. Therefore, the two communicating parties can agree on which group of images they are going to use. This method guaranteeing that a minimum changes will be detected within the image that hold the information. Although the idea is attractive in that it allows a smaller message to be hidden in the cover image without changing the image significantly, the difficulty is that the complexity of time and space will be increased exponentially if the number of the pixels in the subjected block is doubled. However, our aim is not to increase the block size, but to increase only the length of the embedded message, which will be achieved by increasing the number of blocks within the image and hence reducing the blocks sizes. On the other hand, a compression technique could be employed and allow us to compress the transmitted message even further more.

V. SIMULATION RESULTS

The simulation results showed that the text message could be embedded with a non noticeable degradation of the image. Studying the histogram of Laplace filter in Figure 2 for the provided lena image, we notice that on average, the amount by which the image is modified is smaller than some known substitution embedding systems that we investigated in this research work.

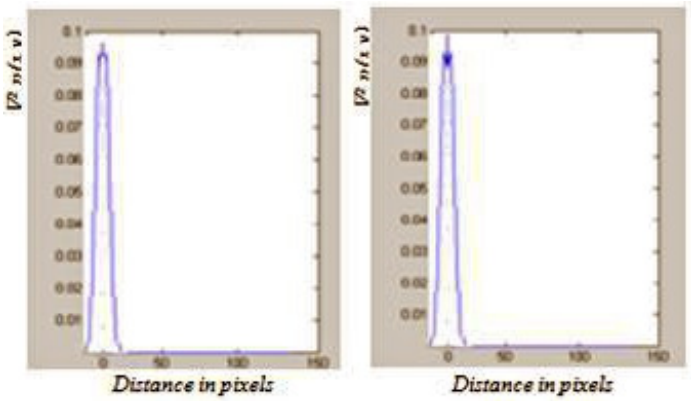


Figure 2. Laplace Filter : (a) the Cover Image (b) The Stego image

For example, comparing the distortion introduced by PGMStealth [1991] and the new technique, we can clearly see that the new technique provides visibly fewer and less peaks than PGMStealth filtered histogram which has a wider band and many peaks clustered around zero as shown in Figure 3. Also comparing our results with the experimental results obtained by other Genetic watermarking algorithms such as Shieh et al. [14], we see that our proposed method

greatly minimizes the effect of the noise caused by the embedding process itself, since it has the ability to keep and refine the results within the selected regions; identifies the one of the most suitable pixels corresponding to the marks placed on the image and allow choice of the correct settings to the threshold healthy pixels**,** and thus making the watermarked image perceptually similar to the original one.

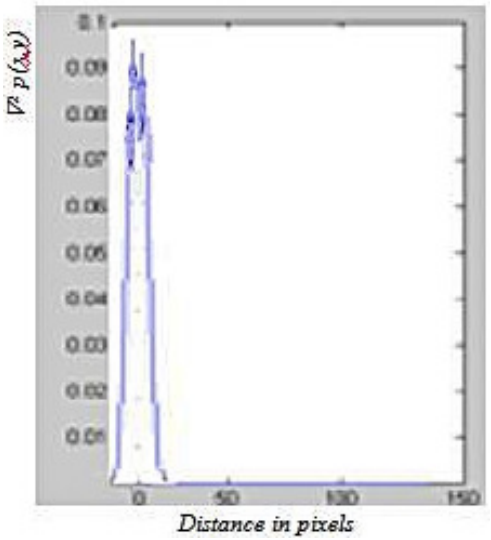


Figure 3. Laplace Filter of PGMStealth Stego Image

5.1 CONSIDERATION OF VALIDITY

Consider a message that is intended to be transmitted using an embedding process based on genetic algorithm. Without losing generality we may consider this to be a sting of n bits. We also require an appropriate fitness function that embed one bit in one block of the cover image such that the number of blocks exceed the length of the message itself (L). Using a seed of length l bits and a random number generator to identifies the target blocks in the cover image where l << number of blocks. This idea is attractive in that it allows short messages to be embedded within a cover image without introducing any noticeable noise in the stego image. The communicating stego image will also be meaningless to the attacker unless he knows both the entire random number generator, fitness function, and the original cover image. In this approach we use the seed number to generate a sequence of numbers representing the target blocks in the cover image where the parity bits are computed. The parity bits are compared to the stream bit sequences of the secret message. The difficulty with this approach is that the results obtained from the experiments we ran that the longest match between the message and the obtained parity bits was nearly within the ratio of one half. Therefore by changing one least significant bit in the pixels contained the target block will flip the value of the computed parity bit and hence match the subjected message bit. Using a genetic algorithm equipped with a fitness function allowed us to select the appropriate pixel within the block which introduces a minimum cost compared to suggested threshold noise value. Thus enhancing the security value of the whole system by improving the and the statistical properties of the stego image significantly. Moreover if we use a block size of 8x8 pixels, there is a one in 64 chance of finding one LSB that complementing its value will match the fitness function. However, to improve this ratio with larger block size, the chance embedding large messages will be reduced accordingly. In addition the searching time could be increased by nearly 100 times compared to the smaller block size as the experimental results showed.

5.2 ATTACKS ON THE GENETIC BASED ALGORITHM

The effort the attacker needs to break the proposed system will not rely only on discovering which fitness function has been used or which number generator method has been applied to select the target blocks where the embedding process had take place, but also on which cover image the two communicating parties are sharing. In addition, the sender will send the seed number to the recipient either encrypted or hidden into another unimportant cover image, which if discovered and applied to the transmitted stego image, the extracted information will be useless. Only the recipient who has the security knowledge of both the stenographic system and the true cover image, which is shared in advance, will be able to extract the true message information using a reverse embedding process.

VI. CONCLUSION AND FUTURE WORK

In this paper we proposed a new steganography algorithm that uses a Genetic algorithm and a random number generator to produce minimum distortion free images from the original ones. The results of the proposed approach showed that the encoding process distorted the image insignificantly by making small modifications over large numbers of pixels. The algorithm divides the image into small blocks that are analyzed for parity check values equivalent to the embedded bit. The technique was designed with the intent of maximizing the quality of the stego image by the aid of fitness function that introduced extremely small modification to the cover images. Initial investigations showed that this modification was difficult to detect visually, and there is no tell-tale artifact could be picked up during the investigation process. In order to compare the provided approach with other established methods, many stego-analysis techniques were investigated and applied to the stego image. Testing our proposed steganographic algorithm's using the adaptive fitness function, we found that the stego image does not show any artifacts and thus, it gives no indication that the image contains any hidden

 ISBN: 978-1-61208-100-7

information. Comparing the Laplace filter histogram for the provided cover image with the one which contains the embedded message, we noticed that on average, the amount by which the image is modified is smaller than other known substitution steganographic systems that we investigated. Looking at the pixel repetition histogram of the stego image and comparing it with the histogram of the original image, it can be observed that there are only very small differences between them, and there are a few fine lines distributed over some parts of the histogram. For the future work we recommend that the cryptography methods should be taken more seriously into account in order to design a more successful steganographic system and in an attempt to provide a secure function to the steganography process. In addition to make the communication even more secure, we recommend that the secret message should be compressed or encoded before the encryption process takes place. This is important because in this way we will minimize the amount of information that is sent, and hence minimizing the chance of degrading the image.

VII. REFERENCES

1. E. Franz. Steganography preserving statistical properties, proceeding of the 5th internationally Workshop on information Hiding, Noordwijkerhout, The Netherlands, October 2002, LNCS 2578, pp. 278-294, Springer 2003.

2. Fabien A. Petitcolas, R. J. Anderson, and M. G. Kuhn, Information Hiding- A Survey, Proceedings of the IEEE, vol. 87, no.7, pp. 1062-1078. Jul. 1999.

3. A. Heurtas and G. Medioni, Detection of Intensity Changes with Subpixel Accuracy Using Laplacian- Gaussian Masks, IEEE Transactions on pattern analysis and machine intelligence, vol. pami-8, no. 5, pp. 651-664, September 1986.

4. R. Anderson, R. Needham, and A. Shamir, The Steganographic File System, in Proceedings of the Second International Workshop on Information Hiding, vol. 1525 of Lecture Notes in Computer Science, Springer, pp. 73-82 , 1998

5. D. Piscitello and A. Chapin, Open Systems Networking: TCP/IP and OSI, Addison-Wesley, Reading, Massachusetts, **pp**. 582-596, 1993.

6. Joseph J. K. Ó Ruanaidh and Gabriella Csurka, A Bayesian Approach to Spread Spectrum Watermark Detection and Secure copyright protection for Digital Libraries. IEEE Conference on Computer Vision and Pattern Recognition (CVPR'99), Vol. 1, pp. 207-212, Fort Collins, Colorado, USA, 23-25 June 1999.

7. A. Westfield and A. Pfitzmann, “Attacks on steganographic Systems”. 3rd International Workshop on. Information Hiding, Dresden, Germany, pp. 61-75, September 1999.

8. M. Peyravian, A. Roginsky, and N. Zunic, Hash-Based Encryption System. Computers & Security Vol. 18, No.4, pp. 345-350 , 2003.

9. The USC–SIPI. Image database, Signal & image processing Institute, Electrical Engineering Department, University of Southern California. <URL: htpp://sipi.usc.edu/database/>, May 10, 2010.

10. M. Kwan, How gifshuffle works, Technical report, Helsinki University of Technology, Full-text, June 2004. < URL: http://www.darkside.com.au/gifshuffle/description.html>

11. P. Borovska., “Solving the Travelling Salesman Problem in Parallel by Genetic Algorithm and Multicomputer Cluster”, International Conference on computer Systems and Technologies, pp. 421-430, 2006.

12. S. Ray, S. Bandyopadhyay and S. Pal, “New operators of genetic algorithms for traveling salesman problem” Cambridge : icpr, vol. 2, pp. 497-500, 2004, Vol. 2.

13. E. Lawie, “Combinatorial Optimization; Networks and Moatroids”., Holt, Rinehart, and Winston, New York, pp. Full-text, 1976.

14. Shieh, C., et al., Genetic watermarking based on transform domain techniques. Pattern Recognition, vol. 37, no. 3, pp. 555-565, 2004.

 ISBN: 978-1-61208-100-7

Optimal Activation of Intrusion Detection Agents for Wireless Sensor Networks

Yulia Ponomarchuk and Dae-Wha Seo
Department of Electrical Engineering and Computer Science
Kyungpook National University
Daegu, Republic of Korea
rus_flash@hotmail.com, dwseo@ee.knu.ac.kr

***Abstract*—Recent technological advancements and low price of deployment and maintenance of wireless sensor networks (WSNs) allow their use in numerous applications in industry, research, and commerce, in order to gather environmental data in an unattended manner. Since WSNs usually function in open environments, they may become a target of attacks or malicious activities aiming to gain access to data, manipulate aggregation result, or disrupt the network service. Therefore, intrusion detection becomes crucial for WSNs as a second line of defense. In order to detect "smart" attacks of colluding devices, active monitoring of behaviors of neighboring nodes was proposed, but it is too energy expensive for resource-constrained WSN nodes, making an adaptive technique for activation of intrusion detection system (IDS) agents extremely important. This paper proposed a model for optimal activation of IDS agents for WSNs on the basis of the Ising model.**

Keywords - wireless sensor networks; anomaly detection; intrusion detection system; Ising model

I. INTRODUCTION

WSNs have become one of the most interesting areas of research owing to the recent advancements of technology, low price and easiness of deployment and maintenance, and application flexibility. A common WSN includes a large number of static sensor nodes and one or several base stations (BSs). Sensor nodes are very simple and cheap devices with constrained resources of memory and power, poor processing and communication capabilities. They monitor environment parameters (e.g., temperature, pressure, humidity) and transmit the sensed data in a hop-by-hop manner towards the BS. BSs or sink nodes are usually more powerful and secure, capable of maintaining WSN topology, collecting data from nodes, storing, preprocessing, and sending them to a user or another network, such as Internet.

Commonly, WSNs function in an unattended manner in open environments with easy access. WSN devices communicate via open radio channel and they are prone to occasional network failures, such as HW/SW faults, loss of connectivity, natural disasters [1-4]. Moreover, WSN nodes are vulnerable to a large variety of attacks that may target physical integrity of devices, as well as routing protocols and data, transmitted within the network. Traditional security schemes can not be applied to WSNs directly because of severe resource constraints of nodes and absence of central authority, therefore, new simple, lightweight and efficient algorithms are needed [1, 2, 4]. Moreover, since no security scheme can guarantee that an attacker will not succeed eventually, an IDS is required as a second line of defense. It may detect nodes' anomalous behavior and activate response measures to avoid an assault or minimize its effect on WSN performance. An IDS has to include global agents, responsible for constant monitoring of behaviors of neighboring devices ([5-13]) and node cooperation, because detection of complex attacks of colluding nodes may not be reliable by means of traffic analysis alone. However, this method incurs significant increase of power consumption at a monitoring node, which makes the problem of IDS's self-organization and adaptation extremely important.

In this paper, a model for distributed and adaptive activation of global IDS (GIDS) agents is proposed. It is based on statistical mechanical approach and the Ising ferromagnetic spin model [14-15]. It incurs insignificant computation overheads and small communication costs, since the decision on activation of IDS agents is done locally and there is no need to send all relevant data to the BS. Combined with a reliable traffic analysis technique for local intrusion detection, it is capable of detecting "smart" attacks of colluding devices.

The paper is organized as follows. Section 2 provides description of the problem and brief background information. Section 3 presents the Ising model for activation of GIDS agents. In Section 4, an algorithm for activation of IDS agents is proposed. Section 5, finally, concludes the paper.

II. BACKGROUND

The problem of design of a distributed, lightweight, and efficient IDS for WSNs has been drawing attention of researchers in recent years. Traditional IDSs are classified into network-based (NIDS) and host-based (HIDS) [16]. While HIDS analyzes incoming and outgoing traffic from individual hosts, NIDS is placed at strategic points of the network to analyze the traffic from all devices.

Another approach to IDS classification is based on detection technique: signature or misuse detection, anomaly detection, or specification-based detection [16-17]. *Misuse detection-based* IDSs rely on a priori knowledge of attacks. Therefore, they detect the majority of known intrusions and have rather low false positive rate (number of false alarms). However, new types of assaults can be missed and signature database may require large memory resources. *Anomaly-based* IDSs detect intrusions by comparison of newly acquired traffic profiles to previously created normal profiles.

 ISBN: 978-1-61208-100-7

They are capable of detection of new attacks, but have higher false positive rate, since random network failures are confused with intended assaults. *Specification-based* IDSs use a set of rules or constraints, specific for running protocols and applications. They are considered to be the most suitable for WSNs, since they are able to detect new types of intrusions, have low false positive rate, and require less memory to store specification database.

Significant number of the previously proposed IDSs relies on analysis of incoming and outgoing traffic of a node and monitoring neighbors' behaviors (watchdogs technique) [5-13]. While the former is not energy consuming and can be performed constantly by a local IDS (LIDS) agent, the latter is expensive in terms of energy and memory resources and it is done by a GIDS agent [5, 7-8, 12]. LIDS module detects intrusions against traffic flow in the nearest vicinity, but it is not capable of reliable detection of complex assaults initiated by collaborating malicious devices, which makes GIDS desirable to operate at least on a portion of nodes. Moreover, since any WSN node may be compromised, the network must defend itself from false accusations and GIDS modules may take responsibility for nodes' cooperation and protect a WSN more efficiently. Recent papers on IDS design take this into account and propose algorithms to optimize GIDS deployment and activation [6-8, 12, 18-19]. However, the suggested schemes still require too many nodes to perform overhearing in normal conditions and lack of adaptability.

Techateerawat and Jennings [18] proposed an adaptive activation of IDS agents. When a WSN is not suffering from an attack, the IDS agents are activated according to core, boundary, or distributed defense strategy. As soon as an intrusion is detected, alarm messages are broadcast to activate IDS agents on nodes in the vicinity of an intruder. This results in isolation of the malicious device and limitation of its effect on network performance. This paper proposes an approach to GIDS agents activation, based on the Ising's ferromagnetic spin model [14-15] of statistical mechanics. The Ising model has been used to study critical phenomena in various systems in diverse disciplines, e.g., finance, biology and sociophysics [20]. It is used to describe the collective behavior of an ensemble of interacting components of a complex system, represented by a lattice [15], where each component has a magnetic dipole (spin), e.g., ± 1. It enables modeling local and global influences on constituting components. In [20], the authors proposed to use the Ising model to provide self-organization of a sensor network in detection of pervasive faults. However, the proposed scheme is centralized: the BS aggregates and stores all relevant information and decides, which nodes to activate. This incurs extra communication costs, results in significant time delay, and may lead to problems with scheduling and node synchronization. In this paper, we suggest to use the Ising model to design a distributed and lightweight scheme for optimal and adaptive GIDS agents activation in WSNs.

III. ISING MODEL FORMULATION FOR AN IDS

The Ising model deals with systems, which can be represented as graphs with vertices as interacting components. A common WSN may be considered as a weighted graph with nodes as vertices and links between nodes as edges. Let $G = (V, E, W)$ denote a weighted graph of a WSN, where $V = \{v_i,\ i = \overline{1,N}\}$ is the set of components (sensor nodes), $E = \{(v_i, v_j) \mid v_i, v_j \in V,\ i, j = \overline{1,N}, i \neq j\}$ is the set of edges or possible links between any two nodes, representing interdependences between a pair of components (there are no self-loops), and $W = \{w_{ij} \mid w_{ij} \geq 0,\ i, j = \overline{1,N},\ i \neq j\}$ is a set of weights assigned to edges $(v_i, v_j),\ i, j = \overline{1,N}$ and representing the strength of interaction between nodes v_i and v_j. Thus, each interaction in G is defined by an edge (communication link) and its weight (link quality or trust value). Though G may be a directed graph, in common WSNs nodes change their roles in time course, as well as routing paths. In general, w_{ij} are time dependent, but they are assumed to be constant within a given time interval. Each node $v_i,\ i = \overline{1,N}$, is assigned a spin σ_i, representing the state of its GIDS agent: $\sigma_i = -1$ - GIDS agent is inactive and the node performs only analysis of incoming traffic from neighbors, $\sigma_i = +1$ - GIDS module is active, monitoring and analyzing communication within the radio range.

Given a weighted graph G, a time-dependent Hamiltonian H^{τ} is constructed; it represents the energy in terms of the Ising model:

$$\mathrm{H}^{\tau} = -\sum_{\langle i,j \rangle} w_{ij}\sigma_i\sigma_j - \mathrm{B}^{\tau}\sum_i \sigma_i , \qquad (1)$$

where $\langle i, j \rangle$ denotes pairs of spins σ_i, σ_j of nearest neighbors v_i, v_j; w_{ij} and B^{τ} represent local interactions and external time-dependent field respectively. Since $w_{ij} \geq 0$, nodes tend to have the same state as their neighbors, unless affected by B^{τ}. The assumptions, provided above, correspond to a multicomponent system, where neighbors with anomalous behavior, make a node more likely to change its state from -1 to +1 under similar external influences, which is analogous to ferromagnetic influences of the magnetization model. It should be noted that while the Ising model deals only with binary states of a spin, there are general models, i.e., the Potts model (where a spin may take integer values $\sigma_i = \overline{1,q}$) and the continuous spin models (the XY model and the Heisenberg model) [15], which are considered for further research.

As it was mentioned earlier, external influence of environment is represented by B^{τ}. According to (1), spins tend to line up in the same direction as the external field. In other words, they want to be positive if $\mathrm{B}^{\tau} > 0$ and negative if $\mathrm{B}^{\tau} < 0$. The value B^{τ} at node v_i in time τ can be written as follows:

 ISBN: 978-1-61208-100-7

$$\mathrm{B}^{\tau}(i) = \sum_{k=1}^{N} B_k\left(\mu_k^{\tau}, \left\{\mu_{k_j}^{\tau}\right\}\right)\delta_k(i), \quad (2)$$

where $B_k\left(\mu_k^{\tau}, \left\{\mu_{k_j}^{\tau}\right\}\right)$ is a function that represents external field in the neighborhood of node v_k and depends on the scalar anomaly measure μ_k^{τ} at node v_k and the set of anomaly measures $\left\{\mu_{k_j}^{\tau}\right\}$ of its nearest neighbors v_{k_j}; $\delta_k(i)$ is the Kronecker delta, i.e., $\delta_k(i)=1$ if $k=i$ and $\delta_k(i)=0$ if $k \neq i$. The functional form of $B_k\left(\mu_k^{\tau}, \left\{\mu_{k_j}^{\tau}\right\}\right)$ is taken identical for all nodes v_k, following [20]:

$$B\left(\mu_k^{\tau}, \left\{\mu_{k_j}^{\tau}\right\}\right) = B_0\left(\mu_k^{\tau} + \sum_{k_j} \mu_{k_j}^{\tau} \cdot \exp\left(-\alpha\left|k-k_j\right|\right)\right), \quad (3)$$

where $\left|k-k_j\right|$ represents the distance from node v_k to its neighbor v_{k_j}, α is a weight coefficient for the distance measure, B_0 is a parameter of the function. The value of the anomaly measure μ_k^{τ} is a result of LIDS agent's monitoring of incoming traffic from other devices at node v_k and the set $\left\{\mu_{k_j}^{\tau}\right\}$ represents alerts from its neighbors v_{k_j}.

Given the spin states of nodes and anomaly measures at a given time instant, the problem of self-organization of IDS agents in a WSN is reduced to the estimation of probabilities of the possible subsequent states of the Ising system. Since each spin can take two values, there are 2^N states in total for a graph with N vertices [15]. In order to compute the probabilities of subsequent states, a statistical mechanical approach is used. It is described in the next section.

IV. Optimal Activation of IDS Agents in WSNs

In this section we apply a statistical mechanical approach [14-15] to activation of an IDS, specifically activation and switching off GIDS agents in WSNs. Unlike a typical problem of statistical mechanics, the goal of the proposed model is to estimate probabilities of future thermodynamic states of the system, provided a particular state at time instant τ, not to compute macroscopic parameters (internal energy, the entropy, the specific heat, etc.). The model addresses two problems: it measures the degree of anomaly of traffic flow, using the enhanced version of traffic analysis method [21], and defines the distribution of nodes with active GIDS agents in a WSN.

In terms of statistical mechanics, a thermodynamic state of a system, represented by graph G, is given by the spin states of graph's vertices. The probability P_I of the system being in state I is defined by the Gibbs distribution [14-15]:

$$P_I(\sigma_1,...,\sigma_N) = \frac{1}{Z} e^{-\beta E_I}, \quad (3)$$

where the energy of the state E_I is defined by Hamiltonian H (1), β is proportional to the inverse temperature, and Z is the partition function of the model, defined as the sum:

$$Z = \sum_{\{\sigma_i\}} e^{-\beta \mathrm{H}}. \quad (4)$$

The partition function may be difficult to compute for systems with an irregular lattice and large number of interacting nodes. However, computations are tractable if the simplifying assumptions are made [15].

- The system follows *Markov dynamics*, i.e., the future state depends only on the present state.
- The system has *quasi-static equilibrium* at all time instants, i.e., the probability of transitions between states, having large energy difference is infinitesimal, the system follows single-spin-flip dynamics.
- The system follows the condition of *detailed balance*, i.e., probabilities P_I and P_J of the system being in states I and J respectively and transition probabilities p_{IJ} and p_{JI} are related as:

$$P_I p_{IJ} = P_J p_{JI} \Leftrightarrow \frac{P_I}{P_J} = \frac{p_{JI}}{p_{IJ}}, \quad (5)$$

where $p_{IJ} = p_{JI} \cdot e^{-\beta(E_J - E_I)}$ from (3) and all transition probabilities should satisfy the constraint:

$$\forall I: \sum_J p_{IJ} = 1. \quad (6)$$

Requirements (5-6) allow to break the transition probability into two parts and apply the Metropolis algorithm, the most efficient and widely used for the Ising model [15]:

$$p_{IJ} = g_{IJ} A_{IJ}, \quad (7)$$

where g_{IJ} is the probability that given an initial state I, a new target state J will be generated (selection probability) and A_{IJ} is the acceptance ratio, showing that if the system starts off from state I and the algorithm generates state J from it, the transition will be accepted. In the Metropolis algorithm, selection probabilities are chosen to be equal, resulting in:

$$\forall I \neq J, g_{IJ} = g_{JI} = \frac{1}{N} \Rightarrow \frac{p_{IJ}}{p_{JI}} = \frac{A_{IJ}}{A_{JI}} = e^{-\beta(E_J - E_I)}. \quad (8)$$

According to [15], the optimal algorithm is chosen when

$$A_{IJ} = \begin{cases} e^{-\beta(E_J - E_I)}, & if \;\; E_J - E_I > 0, \\ 1, & otherwise. \end{cases} \quad (9)$$

In other words, if the algorithm selects a state, which has the energy lower than or equal to the present one, such a transition will be always accepted. If a selected state has higher energy, then it may be accepted with the probability, defined by (9). Since each A_{IJ} is strictly positive, the state transition matrix $[A_{IJ}]$ is irreducible.

The change in energy ΔE_i due to a single-spin-flip (from +1 to -1 or vice versa) at a node v_i is defined by (1) and may be rewritten as:

$$\Delta E_i = 2 \sum_{\langle i,j \rangle} w_{ij} \sigma_i \sigma_j - \mathrm{B}^{\tau} \sigma_i , \qquad (10)$$

where $\langle i, j \rangle$ denotes the set of the nearest neighbors v_j of node v_i . Expression (10) shows that areas with traffic anomalies will have higher energy and ensures that more nodes will be able to detect an intrusion. As soon as an anomaly is eliminated, ΔE_i decreases and nodes tend to stop constant monitoring. The flip probability is computed for v_i using (9):

$$p_i^{flip} = \begin{cases} e^{-\beta \Delta E_i}, & if \quad \Delta E_i > 0, \\ 1, & otherwise. \end{cases} \qquad (11)$$

The value p_i^{flip} shows the likelihood of event that node v_i changes its spin $\sigma_i \to -\sigma_i$ under the influence of its neighbors and external field. Initially, all nodes may have the spin $\sigma_i = -1$ and switch to the active GIDS state ($\sigma_i = +1$) with probability p_0 , defined by the minimal number of active GIDS agents over the whole network.

The algorithm for optimal activation of GIDS agents is summarized in Fig. 1. Each node performs it periodically and is able to switch on/off its GIDS agent in dependence on the information from its close neighbors and traffic intensity. There is no need to transmit anomaly measure values to the BS. The weight coefficients are stored at each node.

Algorithm 1: Self-Organization of IDS agents
while (1) do
Collect traffic data
Compute local anomaly measure μ_i^{τ} at current time instant τ and broadcast it to the one-hop neighbors
Compute the external field B_i^{τ} using (2-3)
Compute the change of energy ΔE_i (10) and p_i^{flip} (11)
Change the spin state with probability p_i^{flip}
end

Figure 1. Algorithm for GIDS agents activation, performed by each node.

V. CONCLUSIONS

The paper proposes a model for adaptive optimal activation of GIDS agents for intrusion detection in WSNs. The model is based on the principles of graph theory and statistical mechanics. Given estimations of traffic anomalies, a small fraction of nodes is activated to monitor neighbors behavior, only when it is necessary. Thus, the scheme reduces power consumption due to overhearing and prolongs network's lifetime. The proposed scheme is distributed and lightweight in terms of computation and communication overheads and may be applied in large WSNs, since BSs are not required to gather and store information about all nodes' behaviors. Further research will be devoted to the performance evaluation via simulations and comparison with other strategies for GIDS agents' deployment and activation.

REFERENCES

[1] V.C. Giruka, M. Singhal, J. Royalty, and S. Varanasi, "Security in wireless sensor networks," Journal on Wireless Communications and Mobile Computing, vol. 8, issue 1, 2006, pp. 1-24.

[2] C.Karlof and D.Wagner, "Secure routing in wireless sensor networks: Attacks and countermeasures," Proc. 1st IEEE Int. Workshop on Sensor Network Protocols and Applications, 2003, pp. 113-127.

[3] D.R. Raymond, and S.F. Midkiff, "Denial-of-service in wireless sensor networks: Attacks and defenses," IEEE Pervasive Computing, vol. 7, issue 1, 2008, pp. 74-81.

[4] Y. Wang, G. Attebury, and B. Ramamurthy, "A survey of security issues in wireless sensor networks," IEEE Communications Surveys & Tutorials, v. 8, issue 2, 2006, pp. 2-23.

[5] C. Besemann, S. Kawamura, and F. Rizzo, "Intrusion detection system in wireless ad-hoc networks: Sybil attack detection and others," TechRepublic, 2004, 14p.

[6] I. Chatzigiannakis and A. Strikos, "A decentralized intrusion detection system for increasing security of wireless sensor networks," Proc. IEEE Conf. on Emerging Technologies and Factory Automation (ETFA), 2007, pp. 1408-1411.

[7] T.H. Hai, F. Khan, and E.N. Huh, "Hybrid intrusion detection system for wireless sensor networks," LNCS 4706, Part II, 2007, pp. 383-396.

[8] T.H. Hai and E.-N. Huh, "Optimal selection and activation of intrusion detection agents for wireless sensor networks," Proc. Future Generation Communication and Networking, 2007, vol.1, pp.350-355.

[9] K. Ioannis, T. Dimitrou, and F.C. Freiling, "Towards intrusion detection in wireless sensor networks," Proc. 13th European Wireless Conf., 2007, 7 p.

[10] G. Li, J. He, and Y. Fu, "Group-based intrusion detection system in wireless sensor networks," Computer Communications, vol. 31, issue 18, 2008, pp. 4324-4332.

[11] I. Onat and A. Miri, "An intrusion detection system for wireless sensor networks," Proc. IEEE Int. Conf. on Wireless and Mobile Computing, Networking and Communications (WiMob'2005), 2005, vol. 3, pp. 253-259.

[12] R. Roman, J. Zhou, and J. Lopez, "Applying intrusion detection systems to wireless sensor networks," Proc. 3rd IEEE Consumer Communications and Networking Conf., 2006, vol. 1, pp. 640-644.

[13] A.P.R. da Silva, M.H.T. Martins, B.P.S. Rocha, A.A.F. Loureiro, L.B. Ruiz, and H.C. Wong, "Decentralized intrusion detection in wireless sensor networks," Proc. 1st ACM Int. Workshop on Quality of Service & Security in Wireless and Mobile Networks (Q2SWinet'05), 2005, pp. 16-23.

[14] K. Huang, Statistical Mechanics, 2nd ed. Wiley, New York, 1987.

[15] M.E.J. Newman and G.T. Barkema, Monte Carlo Methods in Statistical Physics. Oxford University Press, New York, 1999.

[16] Security in Distributed, Grid, Mobile, and Pervasive Computing, ed. by Xiao Y. Auerbach Publications, CRC Press, 2007.

[17] A. Mitrokotsa and A. Karygiannis, "Intrusion Detection Techniques in Sensor Networks" in Wireless Sensor Network Security, J. Lopez, J. Zhou, Eds. Amsterdam: IOS Press, 2008, pp. 251-272.

[18] P. Techateerawat and A. Jennings, "Energy efficiency of intrusion detection systems in wireless sensor networks," Proc. IEEE/WIC/ACM Int. Conf. on Web Intelligence and Intelligent Agent Technology (WI-IATW), 2006, pp. 227-230.

[19] F. Anjum, D. Subhadrabandhu, S. Sarkar, R. Shetty, "On optimal placement of intrusion detection modules in sensor networks," Proc. 1st Int. Conf. on Broadband Networks, 2004, pp. 690-699.

[20] A. Srivastav and A. Ray, "Self-organization of sensor networks for detection of pervasive faults," Signal, Image nad Video Processing, vol. 4, No. 1, 2010, pp. 99-104.

[21] Yu.V. Ponomarchuk and D.-W. Seo, "Intrusion detection based on traffic analysis in wireless sensor networks," Proc. 19th Annual Wireless and Optical Communications Conf., 2010, pp. 229-235.

 ISBN: 978-1-61208-100-7

GeCSen – A Generic and Cross-Platform Sensor Framework for LocON

Mitch De Coster and Steven Mattheussen
Dept. of Applied Engineering
Artesis University College
Antwerp, Belgium
mitch.decoster@student.artesis.be
steven.mattheussen@student.artesis.be

Martin Klepal
Centre for Adaptive Wireless Systems
Cork Institute of Technology
Cork, Ireland
martin.klepal@cit.ie

Maarten Weyn and Glenn Ergeerts
Dept. of Applied Engineering
Artesis University College
Antwerp, Belgium
maarten.weyn@artesis.be
glenn.ergeerts@artesis.be

***Abstract*—In this paper, we present a generic and cross-platform sensor framework for the LocON location and sensor middleware. This framework is developed using C++/Qt. Sensor information is rapidly gaining importance in automating processes and ubiquitous computing, and so it is for projects like FP7 LocON where the main goal is to integrate embedded location systems and embedded wireless communication in order to manage and secure large scale environments. Accessing sensor information mostly requires platform specific code, but for merging this information and sending it over the internet or displaying it on a device, Qt provides cross-platform libraries. Our sensor framework, called GeCSen, also comes in the form of a library that can load platform specific sensor plugins and is able to communicate with the LocON middleware. The framework acts as a cross-platform layer which sends the sensor information to the LocON middleware. This framework enables all kinds of sensors on a wide range of devices to be used by the LocON platform and thereby adds substantial value to the FP7 LocON project.**

***Keywords*-monitoring, localisation, embedded devices, smartphone, sensors, locon, cross-platform**

I. Introduction

Nowadays, sensor data is widely used for many applications, for example, monitoring patients, personnel and equipment in hospitals, monitoring athletes to optimise training methods, etc. To do monitoring we need a selection of sensors such as location sensors, temperature sensors, accelerometers, heartbeat sensors, etc. and some basic logic to parse the sensor data. This data is used to follow someone or something and draw conclusions on which certain actions can be based, for instance, calling an ambulance and automatically give GPS coordinates when someone is having a heart failure. There are different kinds of monitoring but in general we can conclude that there is environmental monitoring to control an area, object monitoring to track a person or object and process monitoring to monitor the proceedings of a process. All situations require different approaches. We are mostly focused on monitoring moving objects.

Currently, monitoring is mainly done by what are called wireless sensor networks [1]; these networks typically consist of a variety of sensor nodes with a small battery, microcontroller and radio transmitter that collects, processes and communicates sensor data. All sensor nodes work together to monitor an environment whereby all collected sensor data reaches a central sink. The sink acts as a link between the sensors and the application. With GeCSen we aim to use computers, embedded systems and smartphones. Many of these devices already have the means of communication and are equipped with various onboard sensors, for instance, most smartphones have a subset of the following: GSM/UMTS, Wi-Fi, GPS, accelerometers, Bluetooth, etc. and can also be extended with various other sensors to suit your needs. Using these kinds of devices comes at the cost of considerably higher power requirements, but also gives enormous processing power in the node itself.

Most applications of sensor usage today are custom developed with industrial usage as target, thus they are specifically built for their needs and only work with specific hardware. Some commercial applications, for example for athletes, are also available, but they are mostly device specific and all use different approaches of which practically none are generic. Notwithstanding, there are already some efforts in making generic sensor frameworks, for example the S60 Sensor Framework of Nokia, the Moblin sensor framework and mSense [2]. Unfortunately they are mostly platform specific, and no more than merely an API for sensors, thus there is an urgent need for a generic way to access sensors used for monitoring.

This paper describes a generic and cross-platform sensor framework, developed in C++/Qt which is easily extendible with plugins to support all kinds of sensors on the majority of platforms and architectures. It thus provides a consistent method of accessing sensor hardware by adding a cross-platform abstraction layer above the APIs described in the previous paragraph, so the plugins are a platform specific wrapper around the APIs which can be loaded in the sensor framework. We also send the sensor data to the middleware of the FP7 LocON project [3], thus we are able to use the LocON data collection and data mining

 ISBN: 978-1-61208-100-7

abilities, together with its localisation fusion engine. Using the LocON platform gives us the opportunity to thrive with the success of the FP7 LocON project, but GeCSen itself also provides substantial value to LocON as it makes it very easy to enable all kinds of sensors on various devices for the LocON platform. We provide the framework itself as a library which can be used statically or dynamically in a host application. The host is able to control GeCSen and use the collected sensor information to, for example display it in a GUI.

The remainder of the paper is organized as follows. In Section 2, we discuss cross-platform aspect of GeCSen and why it is important. The architecture of the framework is described in Section 3. Section 4 goes deeper into the features of GeCSen. In Section 5 we discuss the test cases used to demonstrate the usefulness of the framework, followed by a comparison in Section 6 which shows the advantage of using GeCSen. Future work is discussed in Section 7 and finally, section 8 concludes the paper.

II. Cross-Platform

GeCSen has a broad target group, thus it needs to be able to run on as much platforms as possible. It is not a trivial task to develop a truly cross-platform application, because every platform uses its own libraries and has its own platform definitions. Writing cross-platform applications, basically means having to choose between two major programming languages, C++/Qt and Java. Java is a platform independent language which uses a virtual machine to run Java programs. Qt, on the other hand, is a toolkit written in C++ that uses the same APIs for different platforms. Choosing one over the other essentially means choosing a subset of platforms that you want to support. The main advantage of C++ over Java is a more efficient use of processor cycles and memory and it allows us to program closer to the hardware[4],[5]. The main disadvantage of C++ is that you have to cross-compile for each platform while with Java you do not.

Applications that use GeCSen are mainly focused on mobile and embedded platforms. The market for embedded devices is relatively stable; the two mostly used operating systems are Windows CE and various embedded Linux distributions. The mobile market, on the other hand, is constantly fluctuating; software platforms come and go. As Canalys [6] concludes the most used operating systems currently are Symbian and Blackberry but they also conclude that both Apple iPhone and Google Android are growing very fast, mostly at the cost of Symbian and Windows Mobile. On top of that, every mobile OS has its own set of APIs and most of them are very restricted. This makes developing an application that works on every one of them nearly impossible.

There is also some research going on about cross-platform mobile development which results in a number of very different approaches. Cha, Bernd and Du [7] propose a comprehensive mobile application framework to support interoperability and mobility of mobile application development and operation. Choi, Yang and Jeong [8] suggest an application framework for writing cross-platform mobile applications in Java. PhoneGap [9] is a project that provides a framework in which you can develop programs using simple HTML, CSS and JavaScript. It supports iPhone, Android and Blackberry. MoSync [10] , another project, provides a codebase with which you can program cross-platform in C/C++ for Java ME, Symbian S60, Windows Mobile and Moblin and they are working on Android, iPhone and Maemo. Unfortunately these methods are still too restricted for us to use and don't allow us to program for desktops and laptops, as well as mobile devices.

We have chosen C++/Qt, because although Java is supported by more smartphones than C++/Qt, most smartphone manufacturers have their own restricted APIs. By using C++/Qt a significant part of the smartphone market will still be supported. In addition, Qt is fully open-source; thereby it has a large developer community that thrives to port Qt to every possible platform. It is also gaining more and more attention since Intel and Nokia announced that they are strongly working together to make Qt the default in cross-platform development. GeCSen is tested and works on various x86/x64 and ARM architectures in combination with the following platforms: Windows, Windows Mobile, Windows CE, Android and various Linux distributions, including Ubuntu and a number of OpenEmbedded variants.

III. Architecture

The GeCSen architecture, which is shown in Figure 1, shows a generic way to enable sensors on a device to communicate with the LocON platform. Previously for every different type of device which had to be connected with LocON, or for adding an extra sensor to the LocON client application, it had to be partially, if not completely rewritten. Using this sensor framework the same code can be reused, one only has to compile it for the specific device and develop plugins to use the sensors. As a result, it becomes very easy to add extra sensor devices to the LocON platform. The plugins form a hardware and platform specific layer between the cross-platform framework and the sensors, thus they are responsible for managing the sensors and sending the sensor data to the sensor framework in a uniform way. LocON is working on the standardisation of the data protocol; therefore, the framework has to translate the sensor information, obtained by the plugins, into LocON compatible messages. This is done by using the LocON protocol library [11]. These messages, which can be extended data messages or position messages, are then sent over the network to the LocON platform. This sensor information is very useful to help achieving the main goal of the LocON project, namely securing large scale environments, but also gives endless other possibilities to LocON.

 ISBN: 978-1-61208-100-7

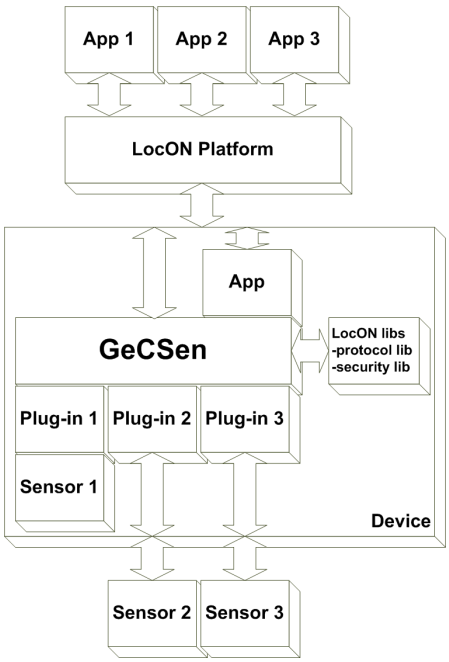


Figure 1. Architecture of the sensor framework

The architecture is in essence a client-server architecture. GeCSen is a client which runs on a device such as a computer, an embedded device or a smartphone. The goal of the GeCSen client is to manage sensor plugins, to collect sensor data from them, to process that data and to send it to the LocON platform. The LocON platform is the server; it is responsible for collecting and processing the sensor data and for seamlessly combining position data. The application on top of the LocON platform is essentially another client that queries the platform for sensor data and position updates so it can be represented in a graphical and useful way or it can be used for further processing.

Figure 2 shows how GeCSen, its plugins and its host applications fit in the OS architecture. It runs on top of the Qt framework and is able to access the system hardware trough the APIs, services and libraries.

IV. Features

A. Plugin System

The plugin system is based upon the Qt Plugins API. The plugins are in essence dynamic libraries that are loaded by the GeCSen framework using a declared interface. This interface is a class containing solely virtual functions that are then used to communicate with the plugins. Every plugin has to have an initialisation and configuration function. The initialisation function, which is the first call to the plugin, is used to initialize the necessary steps, for example, initialise the sensor and start a timer for reading its information. The configuration function is used to configure the plugin and

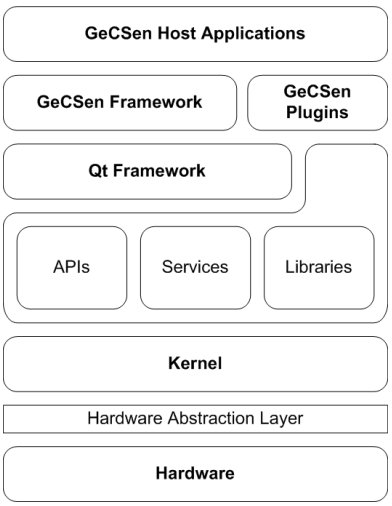


Figure 2. GeCSen on top of OS Architecture

is invoked by GeCSen whenever it receives configuration messages from LocON that are destined for the plugin.

1) Pushing mechanism: Generally plugins are called by the application, accordingly, in our case the framework has to poll the sensors for data. Following the fact that the plugin has to send the sensor information when it is appropriate, for example when a button is pressed, a pushing mechanism had to be developed. When the framework has loaded the plugin and calls the initialisation function, two pointers to functions in GeCSen are given as arguments to the plugin. These two functions can be used to send position messages or extended data messages to the GeCSen framework.

2) Configuration messages: One of the features of LocON is sending configuration messages which can be used for various purposes, GeCSen also supports these configuration messages. By defining and linking them to a sensor in the GeCSen configuration XML, this defined message will be sent to the sensor plugin by using the virtual configuration function in the plugin interface. These configuration messages can be used to set the update rate of a sensor or to manage an actuator for example.

B. Sensor Data Collection

After the plugins are loaded and initialised, they start sending sensor data to the framework by using the function pointers we discussed in the previous paragraph. For extended data messages there are a number of data types available that can be sent which are defined by the LocON protocol and specified for each sensor in the configuration XML. All sensor data is sent as a list of generic pointers, so there is no need for conversion to a certain data type. By using a list we enable the plugin to send multiple values within the same message, for instance multiple MAC addresses of visible Bluetooth devices. Together with these data, the name of the sensor and the plugin is also sent.

 ISBN: 978-1-61208-100-7

Subsequently the data is processed by the sensor handler which checks if the sensor that is sending is allowed to send data and then converts the data into the right format for further processing by the LocON communicator or for usage in the GeCSen host application.

C. LocON Communicator

The LocON communicator uses the LocON protocol and security libraries together with the Qt network libraries to make a connection with the LocON platform and send the right packets accordingly. GeCSen needs to authenticate itself to the LocON platform before it can send any packages. Therefore the communicator sends a handshaking message to the LocON platform. This handshaking message contains a LocON packet with only the subsystem id in it. With this message GeCSen asks to initiate the authentication process. In return LocON sends an open authentication message containing an encrypted blob and a signature blob. Thereafter the paths to the public LocON key and private device key together with the subsystem id are stored in an internal security structure. This structure is used to process the open authentication message received from LocON. When this message is valid an acknowledge message is sent back to LocON to complete the authentication process. After this process GeCSen is able to send signed, and in the future encrypted, messages to LocON. When the connection is lost the communicator will automatically retry to establish a new connection every ten seconds.

Using the data received from the sensor handler a LocON packet is being build. This packet will be filled until the update time interval has passed or when the maximum package size is reached. Thereafter, it is sent to the LocON middleware and the framework starts filling up a new packet.

D. Host Application

Our sensor framework is provided as a library, meaning that it needs a host application to be loaded and initialised. For this purpose GeCSen provides some functions. The most important function is the constructor with parameters to tell GeCSen the location of the configuration file, the plugins directory, the ssl keys for authenticating with LocON and where the log file should be placed. The constructor initialises and starts GeCSen. Some other functions are provided to start and stop GeCSen and to start and stop logging. Qt host applications can also use the signals, of the Qt signal/slot mechanism, we provided for when new extended data, position data or configuration messages arrive. This allows host applications to use sensor data, for example in a GUI on the device itself. Host applications can also be implemented as a console application or even as a service.

E. Configuration

GeCSen can be configured with a XML file similar to the SubsystemDefinitions.xml file used in LocON [**?**]. The GeCSen XML configuration first defines some configuration parameters like server IP and port, whether or not it needs authentication and the update interval in milliseconds. Then a single subsystem is defined which describes a subsystem in which configuration messages and extended data items are defined by a name and data type, exactly the same as in one of the subsystems in the SubsystemDefinitions.xml file of the LocON configuration. The last part of the GeCSen configuration file defines the plugins with their name and sensors that are coupled with extended data items and configuration messages. This allows us to enable or disable plugins and sensors using the configuration file. The configuration file is of great importance to GeCSen and is an agreement between LocON and GeCSen.

V. Test Cases

Various test cases have been developed to test the framework and demonstrate its usefulness. Similar applications already exist in various non-generic, non cross-platform implementations, thereby these test cases prove that we can achieve the same by using our sensor framework. Some test cases we describe here were developed as bachelor theses within our master's thesis, we supported their development and provided the necessary tools and documentation.

A. Remote Athlete Monitoring Application

The first test case is an application for monitoring sporting athletes [13]. GPS data from a HTC Touch Cruise smartphone is combined with the Zephyr HxM [14] external heart rate sensor connected via Bluetooth. Therefore two plugins are written, one to access the GPS sensor and collect the GPS coordinates and another to access the Bluetooth sensor and collect the heart rate values. This data is being presented in GUI applications on the smartphone itself and on top of the LocON platform. This test case should also work on other mobile phones that support Qt and have a GPS and Bluetooth chip, as long as the right plugins are provided.

B. Opportunistic Seamless Localisation Client

Another test case was to build an OSL client using GeCSen [16]. OSL uses the LocON protocol, but without authentication. We have developed a number of OSL compatible plugins such as a Wi-Fi plugin that measures the signal strength (RSSI) of all visible access points, a Bluetooth plugin that detects all Bluetooth devices in the vicinity and a standard GPS plugin. Other possible plugins that we did not provide are a plugin for GSM, an activity plugin that measures the time since your last keystroke and one for step detection with accelerometers. GeCSen, together with a set of OSL plugins is able to act as an OSL client. All these plugins are developed for Windows, Windows Mobile, Linux and Embedded Linux.

 ISBN: 978-1-61208-100-7

C. Internal Social Networking Application

This test case is a social networking application for Windows [17], developed in C#. With this application, friends can locate each other using Wi-Fi based localisation. This application is meant for internal use at our campus to allow students to check if teachers are in their office. To make this possible the application uses GeCSen to send the RSSI values and MAC addresses received from a Wi-Fi plugin to the OSL server using the OSL client implementation described earlier. The OSL server in turn calculates the position of the device sending the Wi-Fi information. For this test case a C# wrapper for GeCSen was developed.

D. Demo Application

During our master's thesis some other simple plugins were developed in order to test GeCSen. This includes a dummy plugin which has two sensors that send dummy values at different intervals, this dummy plugin works on all supported platforms; a CPU plugin that measures CPU usage for Windows, Linux and Embedded Linux and a battery plugin that measures remaining battery life in percent for Windows, Linux and Embedded Linux (ACPI and APM). These plugins are then used together with the other plugins that were previously discussed. For this test case a simple test GUI for the device and an interactive application on top of LocON where you can collect and visualise all the information from the connected devices were developed to demonstrate the advantages of GeCSen in an interactive demo.

VI. Comparison: GeCSen versus OSL client

There are already some systems developed that communicate with and use the LocON platform. For example, the Opportunistic Seamless Localisation System (OSL) client, this is the client that communicates with the OSL server, which uses the LocON protocol. This server combines all sensor data to do opportunistic localisation and in turn sends the location data to the LocON platform [16]. There are a number of OSL clients developed for various platforms including Windows, Windows Mobile and embedded Linux. These clients consist of different implementations to talk to the hardware and to communicate with the OSL server. The OSL client for Windows is developed using C# whereas the version for embedded Linux is using Qt Extended and there is both a C++/Qt as a C# version for Windows Mobile. As you see these are four totally different implementations for achieving the same purpose.

However, the underlying principle is the same for all implementations. When the sensors are being read their data, such as access point information and accelerometer data, is sent to the client controller in a non-generic manner. When new sensors need to be added to the OSL client the client needs to be partly rewritten and compiled. GeCSen is able to do the same job as the OSL client but in a generic manner, adding a new sensor is as easy as developing a small new plugin and adding it to the configuration. The plugin is then loaded by the framework the next time GeCSen starts. This is done without doing any changes to the framework. We believe that GeCSen adds tremendous value for both developers and users of such applications like opportunistic localisation.

VII. Future Work

GeCSen still has to be tested thoroughly with a stress test over a long period of time on various devices as some optimisation of our code might be needed. For example, the device id is currently generated by scanning for network interfaces, chosing the best MAC address and storing it in a file for later usage. In the future when every device has its own certificate, it seems better to use a hash of that certificate as device id, because this is much more difficult to spoof compared to a MAC address. Improvements can also be made to the parser of the configuration XML file. We previously used XML schema validation but it added too much overhead to GeCSen because the Qt library for XML schema validation is too large. This means that our current parser is not completely fail-safe.

The sensor framework should also be tested on more platforms and devices that Qt is ought to work on. For example, the Symbian S60 platform on which we started, but did not succeed within the time that was anticipated. This was mainly because of the fact that we did not have a S60 smartphone to develop on and the standard emulator from the Symbian S60 SDK did not fulfill our needs. Some other mobile platforms on which GeCSen should also work, but are not yet tested include Maemo 5, MeeGo and Symbian 3. The Qt community is currently also working on the Qt Mobility project [18], this is a set of APIs that make it easier for developers to take advantage of mobile features such as using sensors. When this project is completed it could add tremendous value to the future of our project.

Another test case currently is in development as a bachelor thesis, called Smart Environment for Indoor Localisation and Evaluation [15]. The scope of this thesis is to solve the problem where patients who are in need for help still have to wait too long for treatment. Using this system doctors will be automatically warned when patients health conditions are abnormal. GeCSen will be used on a gateway to send sensor and localisation data to the LocON or OSL server for further processing. The GeCSen framework is expected to be included in the following up master thesis next year.

In order to keep GeCSen practically useful, a provisioning system would also be a valuable addition. We intended starting with the development of such a system, but due to time limitations we suspended the project. At last GeCSen should be promoted because of its simplicity and user friendliness. That is why we want to keep it alive by starting up a LocON alliance which recommends GeCSen to be used

 ISBN: 978-1-61208-100-7

as a standard way to connect devices with their sensors to the LocON platform.

VIII. Conclusion

Sensor readings add substantial value to projects like FP7 LocON. Nowadays, there is no generic and cross-platform tool for using sensors. GeCSen is a solution to this ever growing problem that occurs today. Combining various sensors as GSM/UMTS, Wi-Fi, GPS, Bluetooth, RFID and UWB can result in more accurate localisation systems. Other sensors as presence detection sensors can result in more secure environments. Whereas yet other sensors as heart rate sensors and temperature sensors can provide useful applications, for example to monitor sporting athletes or patients. We tried to port GeCSen to as many platforms and architectures we could get our hands on, and we reached even beyond our initial goal. Unlike existing sensor frameworks which are platform specific, GeCSen allows to expose generic sensor information from a broad range of platforms and devices to middleware like LocON, in an easy and uniform way.

References

[1] I. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "A survey on sensor networks," *IEEE Communications magazine*, vol. 40, no. 8, pp. 102–114, 2002.

[2] J. Krösche, A. Jakl, D. Gusenbauer, D. Rothbauer, and B. Ehringer, "Managing Context on a Sensor Enabled Mobile Device-The mSense Approach," in *2009 IEEE International Conference on Wireless and Mobile Computing, Networking and Communications*. IEEE, 2009, pp. 135–140.

[3] S. Couronné, N. Hadaschik, M. Faßbinder, T. von der Grün, M. Weyn, and T. Denis, "LocON–a Platform for an Inter-Working of Embedded Localisation and Communication Systems," *Proceeding of 6th Annnual IEEE SECON*, 2009.

[4] M. Kalle Dalheimer, "A comparison of qt and java for large-scale, industrial-strength gui development," Klarlvdalens Datakonsult AB, Tech. Rep., 2005.

[5] L. Prechelt, "An empirical comparison of c, c++, java, perl, python, rexx, and tcl," University of Karlsruhe, Tech. Rep., 2000.

[6] Canalys. (2010, May) Smartphone market analysis. [Online]. Available: http://www.canalys.com/pr/2009/r2009112.html

[7] S. Cha, J. Bernd, and W. Du, "Toward a Unified Framework for Mobile Applications," in *Proceedings of the 2009 Seventh Annual Communication Networks and Services Research Conference-Volume 00*. IEEE Computer Society Washington, DC, USA, 2009, pp. 209–216.

[8] Y. Choi, J.-S. Yang, and J. Jeong, "Application framework for multi platform mobile application software development," in *11th International Conference on Advanced Communication Technology*. ICACT, 2009, pp. 208–213.

[9] PhoneGap. (2010, May) Cross platform mobile framework. [Online]. Available: http://phonegap.com/

[10] MoSync. (2010, May) the open source standard tool for cross-platform mobile applications. [Online]. Available: http://www.mosync.com/

[11] H. Millner, P. Gulden, M. Weyn, M. Fabinder, and A. Casaca, "Description of the locon protocols. deliverable 4.2 of the fp7 locon project," Symeo, CIT, CEA-LETI, INOV, Artesis, Fraunhofer IIS, Tech. Rep., 2009.

[12] A. Chambron, M. Fabinder, N. Hadaschik, S. Wibowo, P. Gulden, K. Willame, and G. Pestana, "Concept of an inter-operable interface. deliverable 4.3 of the fp7 locon project," ANA, perLocus, Symeo, CEA/LETI, INOV, CIT, Tech. Rep., 2009.

[13] D. Pauwels, K. Fierens, G. Ergeerts, and M. Weyn, "Remote Athlete Monitoring Application for LocON," 2010.

[14] Zephyr. (2010, May) Zephyr hxm — heart rate and accelerometer data analysis. [Online]. Available: http://www.zephyr-technology.com/hxm.html

[15] M. Weyn and M. Klepal, "Adaptive Motion Model for a Smart Phone Based Opportunistic Localization System," *Mobile Entity Localization and Tracking in GPS-less Environnments*, pp. 50–65, 2009.

[16] Y. Budts, G. Ergeerts, and M. Weyn, "Internal Social Networking Application using LocON," 2010.

[17] Q. Labs. (2010, May) Qt mobility project. [Online]. Available: http://labs.trolltech.com/page/Projects/QtMobility

[18] B. Pauwels, D. Lacko, D. Vermeiren, G. Ergeerts, M. Weyn, and R. Steurs, "Smart Environment: Indoor Localization and Evaluation (SEnILE)," 2010.

 ISBN: 978-1-61208-100-7

Securing Off-Card Contract-Policy Matching in Security-By-Contract for Multi-Application Smart Cards

Nicola Dragoni, Eduardo Lostal, Davide Papini
DTU Informatics
Technical University of Denmark
{*ndra,dpap*}*@imm.dtu.dk*
eduardolostal@gmail.com

Javier Fabra
Department of Computer Science and Systems Engineering
University of Zaragoza
jfabra@unizar.es

***Abstract*—The Security-by-Contract (S×C) framework has recently been proposed to support applications' evolution in multi-application smart cards. The key idea is based on the notion of *contract*, a specification of the security behavior of an application that must be compliant with the security policy of a smart card. In this paper we address one of the key features needed to apply the S×C idea to a resource limited device such as a smart card, namely the outsourcing of the contract-policy matching to a Trusted Third Party. The design of the overall system as well as a first implemented prototype are presented.**

***Keywords*- Multi-Application Smart Cards; Security; Contract Matching.**

I. INTRODUCTION

Java card technology has progressed at the point of allowing several Web applications to run on a smart card and to dynamically load and remove applications during the card's active life[1]. With the advent of these new *Web enabled multi-application smart cards* the industry potential is huge. However, concrete deployment of multi-application smart cards have remained extremely rare. One reason is the lack of solutions to an old problem: the control of interactions among applications. Indeed, the business model of the asynchronous download and update of applications by *different parties* requires the control of interactions among possible applications *after* the card has been fielded. In other words, what is missing is a quick way to deploy new applications on the smart card once it is in the field, so that applications are owned and asynchronously controlled by different stakeholders. In particular, owners of different applications (banks, airline companies, etc.) would like to make sure their applications cannot be accessed by new (bad) applications added after theirs, or that their applications will interact only with the ones of some business partners.

To date, current security models and techniques for smart cards (namely, permissions and firewall) do not support any type of applications' evolution. Smart card developers have to prove that all the changes that are possible to apply to the card are security-free, so that their formal proof of compliance with Common Criteria is still valid and they do not need to obtain a new certificate. The result is that there are essentially no multi-application smart cards, though the technology already supports them (Java Card and Global Platform specifications).

The Security-by-Contract (S×C) framework has recently been proposed to address this challenge [1]. The approach was built upon the notion of Model Carrying Code (MCC) [2] and successfully developed for mobile code ([3], [4] to mention only a few). The overall idea is based on the notion of *contract*, that is a specification of the security behavior of an application that must be compliant with the security policy of the hosting platform (i.e., the smart card). This compliance can be checked at load time and in this way avoid the need for costly run-time monitoring.

The effectiveness of S×C has been discussed in [1], [5], where the authors show how the approach can be used to prevent illegal information exchange among several applications on a single smart card, and how to deal with dynamic changes in both contracts and platform policy. However, in those papers the authors assume that the key S×C phase, namely *contract-policy matching*, is done on the card, which is a resource limited device. What they leave open is the issue of outsourcing the contract-matching phase to a Trusted Third Party, in case this phase requires a too expensive computational effort for the card. In this paper we explicitly address this issue, discussing the design and a first prototype of this key functionality of the S×C framework.

The paper is organized as follows. In Section II we introduce the S×C framework and the problem we tackle. Then the discussion of the design and implementation details of the proposed system are depicted in Section III and IV, respectively. Section V concludes the paper summarizing its contribution.

II. SECURITY-BY-CONTRACT (S×C)... IN A NUTSHELL

In the S×C approach, mobile code carries with a claim on its security behavior (an *application's contract*) that could be matched against a mobile *platform's policy* before downloading the code. In this setting, a digital signature does not only certify the origin of the code but also binds together

[1] http://java.sun.com/javacard/specs.html

 ISBN: 978-1-61208-100-7

the code with a contract with the main goal to provide a semantics for digital signatures on mobile code.

At *load time*, the target platform follows a workflow similar to the one depicted in Fig. 1 (see also [6]). First, it checks that the evidence is correct. Such evidence can be a trusted signature as in standard mobile applications [7]. An alternative evidence can be a proof that the code satisfies the contract (and then one can use PCC techniques to check it [8] or specific techniques for smart-cards such as [9]).

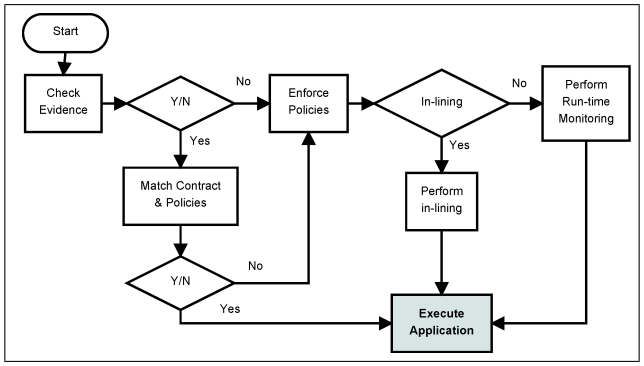


Figure 1. S×C Workflow

Once we have evidence that the contract is trustworthy, the platform checks that the claimed policy is compliant with the policy that our platform wants to enforce. This is a key phase called *contract-policy matching* in the S×C jargon. If it is, then the application can be run without further ado. At run-time, a firewall (such as the one provided by the Java Card Runtime Environment) can just check that only the declared API in the contract can be called. The matching step guarantees that the resulting interactions are correct. This is a significant saving over full in-line reference monitors.

A. Off-Card Contract-Policy Matching

A key issue in the S×C framework concerns who is responsible for executing the contract-policy matching. Due to the computational limitations of a resource limited environment such as a smart card (SC), running a full matching process on the card might be too expensive. In the S×C setting, the choice between "on-card" and "off-card" matching relies on the level of contract/policy abstraction [1], [5]. Indeed, the framework is based on a hierarchy of contracts/policies models for smart cards, so that each level of the hierarchy can be used to specify contracts/policies with different computational efforts and expressivity limitations.

This paper focuses on the situation where contract-policy matching is too expensive to be performed on the card. The idea, depicted in Fig. 2, is that a Trusted Third Party (TTP), for instance the card issuer, provides its computational capabilities to perform the contract-policy compliance check. The TTP could supply a proof of contract-policy compliance to be checked on the smart card. The SC's policy is then updated according to the results received by the TTP: if the compliance check was successful, then the SC's policy is updated with the new contract and the application can be executed. Otherwise, the application is rejected or the policy enforced off-card (for example, by means of a service provided by the TTP in addition to contract-policy matching). In case the TTP includes a proof of compliance in the reply, then a further check is needed to verify the proof, as shown in Fig. 2.

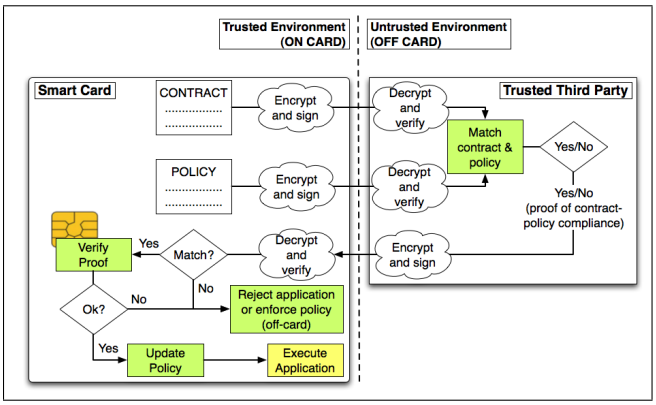


Figure 2. Off-Card S×C Contract-Policy Matching

In this scenario, the communication between SC and TTP must be secured in order to deal with an untrusted environment. Both contract and policy must be encrypted and signed by SC before they are sent to the TTP to ensure authentication, integrity and confidentiality. Analogously, the results of the compliance check should be encrypted and signed by TTP before they are sent back to SC.

III. Securing Off-Card Matching

To secure the system we use *Public Key Infrastructure* (PKI), where keys and identities are handled through certificates (namely, X.509 certificates [10]) that are exchanged between parties during communication. For this reason, the SC must engage an *initialization phase*, where certificates are stored in the SC along with security policies. The security of the system relies on the assumption that the environment in this phase is completely trusted and secure. As above mentioned all messages between SC-TTP will be signed and encrypted. We have decided to use two certificates (i.e. two different key-pairs), one for the signature and one for the encryption, so that in the unlikely event of one being compromised the other is not. The use of two certificates is optional, but it makes the system more secure.

In this Section we first show the design of the *initialization phase* and then pass over the *contract-policy matching* one. Since the system is based on *Java card 2.2.2*, the SC acts as a server which responds only to Application Protocol Data Unit (APDU) commands by

 ISBN: 978-1-61208-100-7

means of APDU-response messages.

Initialization Phase. This phase is divided into three different steps: *Certificate Signing Request* (CSR) building [11], certificates issuing, and finally certificates and policy storage. As shown in Fig. 3 the first step consists in building the CSR for the two certificates to be sent to the *Certification Authority* (CA). The Trusted Reader (TR) queries the SC for its *public key*, then TR builds the CSR and sends it back to SC that signs it. Message #4 `SPrKSCEnc(SCEncCSR)` means that the CSR for encryption is signed (`S`) with private key (`PrK`) of `SC` for encryption (`Enc`). Messages throughout all figures are likewise.

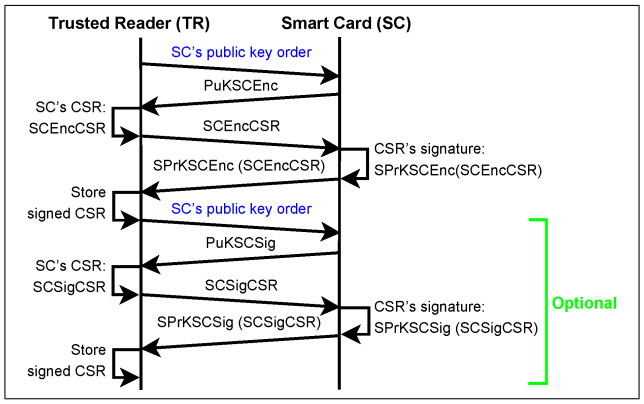


Figure 3. CSRs Building

In the second step (Fig. 4) the TR - Certificates Manager (TRCM) sends to CA the CSRs previously built, CA issues the certificates and then sends them back to the TRCM.

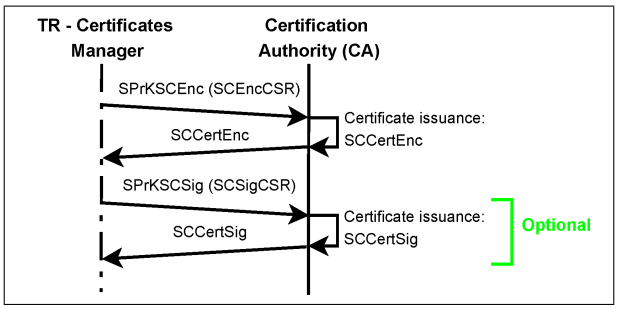


Figure 4. Certificates Issuing

The final step, shown in Fig. 5, completes the *initialization phase* by storing in the SC the two certificates, the security policy and the CA digital certificate (this is needed by the SC to verify certificates of TTP).

After the SC has been initialized it is ready to securely engage in any activity that involves the *contract and policy matching*. Specifically the card will be able to verify the identity of the TTP, authenticate and authorize its requests.

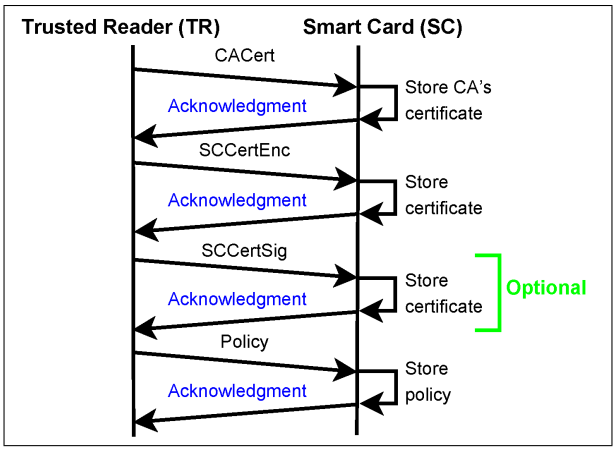


Figure 5. Storage of Keys and Certificates on Smart Card

Contract-Policy Matching Phase. During this phase the contract and the security policy, stored in the card, are sent from SC to some TTP which runs the matching algorithm and then sends the result back to SC. Our goal is to secure communication between TTP and SC in terms of mutual authentication, integrity and confidentiality. The solution we propose is shown in Fig. 6. It is divided into three parts: *certificates exchange*, *contract and policy sending*, *matching result sending*.

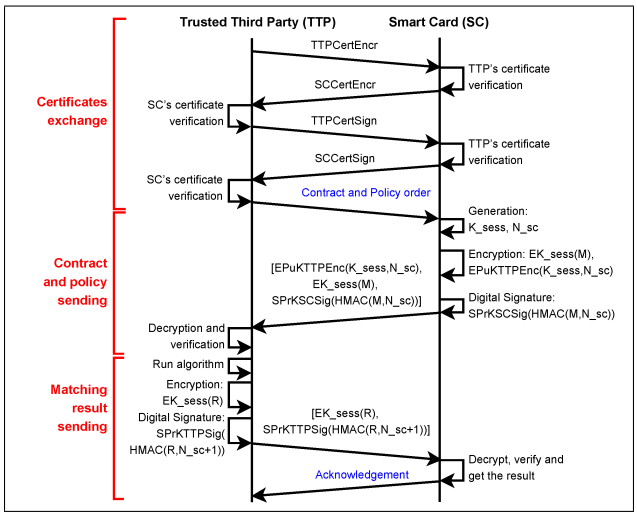


Figure 6. Protocol for Off-Card Contract-Policy Matching

In the first part TTP and SC exchange their own pair of certificates (one for encryption and one for the signature) and then respectively check the validity of those. Particularly, the SC checks them against CA certificate stored during *Initialization phase*. If the certificates are valid then the TTP asks SC for the contract and policy. At this point the SC engages in a sequence of actions aiming to secure the message *M* containing requested information that needs to

 ISBN: 978-1-61208-100-7

be sent back to TTP. Specifically:

1) It generates a *session key* and a *NONCE* (Number used Once) that will be used for this communication.
2) It encrypts the *session key* and the *NONCE* with TTP *Public Key*, and then message *M* with the *session key*.
3) It computes the HMAC (a hash mixed with a salt, i.e. the NONCE (N_{sc})) and it signs it.

Then the message is sent to TTP which verifies the message and extracts the needed information.

In the last part TTP runs the matching algorithm against contract and policy, and builds a secure message containing the algorithm result *R* to be sent to SC. The key used for encryption is still the *session key* generated previously by SC. The signature is done as before except that the HMAC uses as salt the value $N_{sc}+1$. At this point SC decrypts and verifies the result and sends an acknowledgement to TTP.

IV. Prototype Implementation

A first prototype of the proposed framework has been implemented, representing almost a fully-functional implementation. Java version 1.6 has been used to implement the TTP and the TR, and Java Card 2.2.2 was used for the SC. This version was used instead of Java Card 3 due to the lack of mature in version 3 (actually, there are no cards supporting its real implementation). An APDU extended length capability has been implemented in order to allow sending up to 32KB data messages instead of the by-default maximum 255 bytes size.

All message exchange protocols have been implemented and authentication, integrity and confidentiality are ensured by means of X.509 certificates in communications between the TTP and the card. These certificates are managed by means of the CA, which generates self-signed certificates using OpenSSL 0.9.8n.

The implementation of the initialization phase is almost finished. All required data is stored and sent to the installer and also sent back to the card. On the other hand, some work must be done in the contract and policy matching phase. Certificate exchange is working properly, but verification is only carried out in the TTP and not on the card yet. RSA keys are used to achieve PKI encryption, but digital signatures and block ciphering must be developed too.

To test the prototype, two different simulation environments have been used. At first stages, the Java Card platform Workstation Development Environment tool (Java Card WDE) was used. However, saving the status of the card and all the session data is currently being addressed, so the environment has been changed to the C-language Java Card RE (CREF), which eases this feature.

V. Conclusion

In this paper we have addressed the issue of outsourcing the S×C contract-policy matching service to a Trusted Third Party. The design of the overall system as well as a first implemented prototype have been presented. The solution provides confidentiality, integrity and mutual authentication altogether. In particular, the following mechanisms have been implemented to strengthen the security of the system: (i) The use of two different certificates for signature and encryption. (ii) A NONCE created for each session to ensure freshness of the messages. (iii) Both the *session key* and the *NONCE* are generated within the SC, and then sent encrypted to TTP. The fact that TTP uses them to correctly compose the message *R* is a proof that TTP is the one that decrypted the message in the same session (due to the freshness of *NONCE*) and no one else did (the only way would be to get the Private Key of TTP but Public Key Cryptosystems are considered secure and unbreakable). (iv) The HMAC sent within the response is salted with $N_{sc}+1$. The change in the value of the salt introduces variability in the hash making it more unlikely to forge.

References

[1] N. Dragoni, O. Gadyatskaya, and F. Massacci, "Supporting applications' evolution in multi-application smart cards by security-by-contract," in *Proc. of WISTP*, 2010, pp. 221–228.

[2] R. Sekar, V. Venkatakrishnan, S. Basu, S. Bhatkar, and D. DuVarney, "Model-carrying code: a practical approach for safe execution of untrusted applications," in *Proc. of SOSP-03*. ACM, 2003, pp. 15–28.

[3] N. Dragoni, F. Massacci, K. Naliuka, and I. Siahaan, "Security-by-contract: Toward a semantics for digital signatures on mobile code," in *Proc. of EUROPKI*. Springer-Verlag, 2007, pp. 297–312.

[4] L. Desmet, W. Joosen, F. Massacci, P. Philippaerts, F. Piessens, I. Siahaan, and D. Vanoverberghe, "Security-by-Contract on the .NET platform," *Information Security Tech. Rep.*, vol. 13, no. 1, pp. 25 – 32, 2008.

[5] N. Dragoni, O. Gadyatskaya, and F. Massacci, "Security-by-contract for applications evolution in multi-application smart cards," in *Proc. of NODES*, DTU Technical Report, 2010.

[6] D. Vanoverberghe, P. Philippaerts, L. Desmet, W. Joosen, F. Piessens, K. Naliuka, and F. Massacci, "A flexible security architecture to support third-party applications on mobile devices," in *Proc. of ACM Comp. Sec. Arch. Workshop*, 2007.

[7] B. Yee, "A sanctuary for mobile agents," in *Secure Internet Programming*, J. Vitek and C. Jensen, Eds. Springer-Verlag, 1999, pp. 261–273.

[8] G. Necula, "Proof-carrying code," in *Proc. of the 24th ACM SIGPLAN-SIGACT Symp. on Princ. of Prog. Lang.* ACM Press, 1997, pp. 106–119.

[9] D. Ghindici and I. Simplot-Ryl, "On practical information flow policies for java-enabled multiapplication smart cards," in *Proc. of CARDIS*, 2008.

[10] ITU-T, "ITU-T Rec. X.509," 2005.

[11] M. Nystrom and B. S. Kaliski, "PKCS #10: Certification Request Syntax Specification version 1.7," RFC 2986, 2000.

 ISBN: 978-1-61208-100-7

MagiSign: User Identification/Authentication

Based on 3D Around Device Magnetic Signatures

Hamed Ketabdar
Quality and Usability Lab, TU Berlin
Deutsche Telekom Laboratories
Ernst-Reuter-Platz 7,
10587 Berlin
hamed.ketabdar@telekom.de

Kamer Ali Yüksel
TU Berlin
Ernst-Reuter-Platz 7
10587 Berlin
kamer.yuksel@telekom.de

Amirhossein Jahnbekam, Mehran Roshandel, Daria Skripko
Deutsche Telekom Laboratories
Ernst-Reuter-Platz 7
10587 Berlin
{amirhossein.jahanbekam, mehran.roshandel, daria.skripko}
@telekom.de

***Abstract*— In this paper, we present "MagiSign", a new user identification/authentication technique based on 3D magnetic signatures created in the space around a (mobile) device. The main idea is to influence magnetic (compass) sensor embedded in some mobile devices (e.g., iPhone 3GS, G1/2 Android) using a properly shaped magnet. The user draws 3D signatures in the 3D space around the device using a magnet (e.g., pen, rod, ring shaped). This is what we call as "3D Magnetic Signature". The temporal pattern of change in the magnetic field around the device is sensed and registered by the internally embedded magnetic sensor. For authentication/identification, new magnetic signature samples are compared with models created based on registered signatures. As the magnetic signature can be flexibly created in 3D space, it provides a wider choice for authentication. Unlike regular signatures, a hardcopy can not be easily generated resulting in higher security. "MagiSign" technique does not require expensive or complex hardware/algorithm, and does not impose major change in hardware or physical specifications of the device. It can be especially suitable for small mobile devices.**

Keywords*-User Identification/Authentication, 3D Magnetic Signature, Around Device Interaction, Magnet, Embedded Magnetic (Compass) Sensor.

I. INTRODUCTION

Portable personal devices such as PDAs or mobile phones are being widely used during daily life. These devices can be used to store or access sensitive information. They are frequently used in insecure locations with little or no physical protection, and are therefore susceptible to theft and unauthorized access. User authentication/identification therefore seems to be essential for granting access to certain information or services. The authentication is conventionally performed using secret codes and personal identification numbers (PINs) [1]. However, they can be easily compromised, shared, observed, stolen or forgotten. Other techniques such as Fingerprint [2], a face profile [3], voice based verification [4], or combination has been also investigated.

Figure 1: 3D Magnetic Signature

In this work, we propose a new authentication/identification technique, mainly for small mobile devices, based on interaction with embedded magnetic (compass) sensor. This sensor is already embedded in some mobile devices (e.g, Apple iPhone and Google Android) for navigation purposes. We call the new method "MagiSign" or "3D Magnetic Signature". The user simply makes a 3D signature in the space around the device using a properly shaped magnet (rod, ring, pen) (Figure 1). Movement of the magnet changes temporal pattern of magnetic filed sensed by the magnetic sensor integrated in the mobile device. For signature identification/authentication, a new signature sample is matched against models created for signatures of users. The idea is partly inspired by Around Device Interaction (ADI) framework [5][6][7], which propose using space around the device for interaction with the device.

As the magnetic signature can be created in 3D space, it provides a very large flexibility for the choice of signature. Unlike regular signatures, it is not easy to make hardcopies of such a signature. The proposed technique relies only on a magnetic sensor (already embedded in some devices mainly for navigation purposes), and a magnet as external accessory. Compared to camera and fingerprint sensor, a magnetic

 ISBN: 978-1-61208-100-7

sensor can be much simpler, smaller and cheaper, and can be internally embedded. It does not impose major change in hardware or physical specifications of devices, which can be especially important for small mobile devices. Extracting useful biometric information from magnetic sensor data can be algorithmically simpler than computer vision or audio processing techniques and the method is not subject to different sources of illumination, occlusion, and audio noise. Such an approach opens up a new, simple and effective way for user identification/authentication based on 3D Magnetic Signatures. Although we mainly talk about mobile devices, the presented approach can be also used for other platforms in a similar way.

The paper is organized as follows: Section 2 describes the idea behind our approach for user identification/verification in more details. Section 3 explains feature extraction and signature/user classification. Section 4 presents experiments and results. An implementation of the proposed approach as a demonstrator on Apple iPhone is introduced in Section 5, and Section 6 provides conclusions and future work tracks

II. MAGENTIC SIGNATURE : THE IDEA

As already mentioned, the basic idea behind our approach is to provide a new user identification/authentication technique based on the so called "3D Magnetic Signature" or "MagiSign". The user creates his own arbitrary 3D signature using a properly shaped magnet in the 3D space around the device (Figure 1). Movement of the magnet changes the magnetic field sensed (registered) by the built in compass (magnetic) sensor. For identification/verification, temporal pattern of a new signature is compared against a model already created based on pre-registered magnetic signature samples of users. Some mobile devices such as Apple iPhone 3GS, and Google Android are already equipped with compass (magnetic field) sensor. The magnet which should be used for creating signatures is a regular non-powered magnet in a proper shape such as rod, ring or pen. The idea is partly inspired from Around Device Interaction framework which proposes to use space around the device for touch less interaction with the device based on analyzing different sensory information [5][6][7].

The 3D magnetic signature provides a wider choice for authentication as it can be flexibly drawn in 3D space around the device, and can be consequently very difficult to replicate. Additionally, unlike regular signature, no hardcopy of the magnetic signature can be easily produced, resulting in higher security.

Although it is potentially possible to capture gesture based signatures by e.g., camera [10], getting useful information from magnetic sensor is algorithmically and technically much simpler than implementing computer vision techniques. In contrast to accelerometer based gesture recognition techniques [9], our hand gesture recognition approach requires a peripheral magnet. Nonetheless, using a tiny magnet in our case helps the user not to loose the direct sight to the mobile device screen. In addition, for installed Authentication/Identification devices in gates, shaking the entire device is not possible while a tiny magnet can be easily used to draw a signature. Our method does not impose major change in hardware or physical specifications of mobile devices. It does not require installing complex, expensive, and space occupying sensors which can be critical in small mobile devices. It is only based on a magnetic sensor which is internally embedded in some new mobile devices. For mobile devices such as iPhone and G1/2 Android, it is only necessary to have a properly shaped magnet as an extra accessory. Unlike face and audio based authentication techniques, our approach does not suffer from illumination variation, occlusion and audio sources of noise. Since the interaction in our method is based on magnetic field (which can pass through hand, body, clothes and many other different objects), even the space at the back of device can be efficiently used for signing, yet providing more flexibility for authentication. Additionally, the user can interact with the mobile device, even if the device is not in the line of sight, or covered (e.g., mobile device in a pocket or bag). For instance, the user can activate a service or unlock the device without taking it out of his pocket/bag.

We have built a demonstrator called "MagiSign" (presented in Section 5) based on the magnetic signature concept. The demonstrator is built as an application for Apple iPhone 3GS. The application allows recording signature templates, and verifying the user identity based on new signature samples. A confidence score indicating the match between new signature samples and the templates is also provided on the screen.

III. PROCESSING MAGNETIC SIGNATURES

Magnetic signatures are created based on arbitrary moving a magnet (a rod or ring) by hand in the space around the device along different 3D trajectories (Figure 1). The signature can be a simple 3D motion, or the regular signature of the user drawn on the air! or any other combination of even higher complexity which actively uses all 3D space around the device. The rod shaped magnet can be installed in a pen. We have used iPhone 3GS as mobile device for our studies.

The embedded compass (magnetic) sensor provides a measure of magnetic field strength along x, y, and z directions. The values change over a range of -128 to 128.

In our current setup, the user should press a button during performing the magnetic signature, in order to indicate the beginning and end of recording magnetic signals. An alternative would be automatically detecting begin and end of the signature by comparing Euclidean norm of magnetic field strength against a pre-defined threshold.

The embedded magnetic sensor captures temporal pattern of change in magnetic field due to the movement of the magnet.

 ISBN: 978-1-61208-100-7

Some features are then extracted from signals captured by the magnetic sensor. The extracted features are then used to train reference statistical models for different users/signatures. During the test of the system, new signature samples are matched against the reference statistical models. An output score indicating the match between the new signature sample and existing models is then used as a basis for identification/verification.

IV. FEATURE EXTRACTION

Feature extraction allows for preserving information which can be discriminative between signatures/users and removes redundant information. All the features are extracted over samples in an interval marked by the beginning and end of the signature. This interval is divided into two equal length windows, and a feature vector is extracted for each window. The two feature vectors are then concatenated to form a new feature vector to be used for signature classification. Dividing the signature interval to multiple windows allows for capturing temporal pattern of the signature in a more detailed way. Features we have used are mainly based on average or variance of magnetic field strength in different directions, as well as piecewise correlation between field strength in different directions. Features used in this study are listed in the following:

- Average field strength along x, y, and z directions (3 features)
- Variance of field strength along x, y, and z directions (3 features)
- Average of Euclidean norm of filed strength along x, y, z (1 feature)
- Variance of Euclidean norm of field strength along x, y, and z (1 feature)
- Piecewise correlation between field strength along x-y, x-z, and y- z (3 features)

These features form an 11 elements feature vector for each window. The two window feature vectors are then concatenated to form a new 22 elements feature vector for each signature.

Alternatively, all above features can be extracted from a time derivative of magnetic signals, instead of raw magnetic signals. Applying a derivative operator before feature extraction can cancel the effect of magnetic source noises (e.g., earth magnetic field).

V. IDENTIFICATION/AUTHENTICATION

The extracted feature vector is used as input to machine learning algorithms for signature identification/verification. We have studied Multi-Layer Perceptron (MLP) [8] as the classifier.

Multi-Layer Perceptron (MLP) is an Artificial Neural Network which can realize an arbitrary set of decision regions in the input feature space. The feature vectors are used to train the MLP. During testing the system, a feature vector is presented at MLP input. The MLP estimates posterior probability of different signature classes at output (each MLP output is associated with one signature/user class). The signature/user class with highest posterior probability is selected as identification/authentication output.

VI. EXPERIMENTS AND RESULTS

We have set up signature identification/verification experiments in order to evaluate our method. We have invited 15 test users for the experiments. Each user is asked to make a magnetic signature 15 times using a rod shaped magnet. We recorded the signals captured by the embedded magnetic sensor using an application developed for Apple iPhone 3GS.

Features are extracted from magnetic signals as described in Section 3.1. As already mentioned, the input to feature extraction can be raw magnetic signals, or their time derivatives. Both cases are studied in the experiments. The extracted features are used for signature/user classification using MLP. We have used a 10 fold cross-validation scheme for managing training and test data.

Table I shows signature/user identification results using MLP as classifier for features extracted from raw, as well as derivative magnetic signals. As can be seen in the table, the best performance reaches good accuracy of 95.2% for user identification. Using raw magnetic signals slightly outperforms the use of derivatives, however we think if the identification process is performed in a situation that orientation or tilt angle of the mobile device can not be well stabilized, derivatives could be more informative.

Table II shows authentication related measures for the experiment, averaged over different users (raw signals are used for feature extraction). These measures are area under ROC curve, True Positive (TP) rate, and False Positive (FP) rate. The authentication results show a good trade-off between true and false alarms.

We have further investigated the issue of user identification/authentication using simple and identical (among users) 3D gestures instead of personalized signatures. For the experiment, we invited 6 users which are asked to all draw similar and simple gestures shown in Figure 2. These gestures are then used for user identification in the same process as explained for personalized signatures.

Table III shows user identification results for different gestures. As it can be seen in the table, even using very simple identical gestures, users can be identified with relatively high accuracy. This means that the whole process extracts biometric information allowing user identification, using even identical simple gestures.

TABLE I. Signature/user identification results. The first column shows the results when raw magnetic signals are used as input to feature extraction. The second column shows results when derivative of magnetic signals is used for feature extraction.

Source	Raw signals	Derivative signals
Accuracy	95.2%	94.4%

 ISBN: 978-1-61208-100-7

TABLE II. User authentication measures, averaged over users.

Measure	ROC Area	TP rate	FP rate
Value	0.991	0.952	0.003

TABLE III: User identification accuracy for simple identical gestures. Gestures are identical among users.

Gesture ID	1	2	3
Accuracy	90.0%	92.2%	91.2%

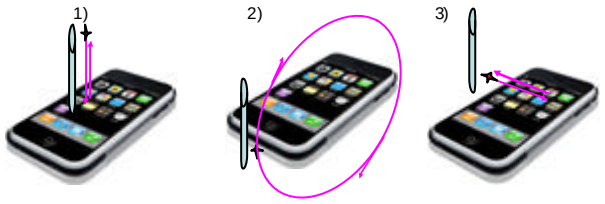


Figure 2: Simple gestures used for user identification.

VII. DEMONSTATOR

Based on the idea presented in this paper, we have implemented a user verification/identification demo for Apple iPhone 3GS. In our demo program, users are allowed to register a few templates of their 3D Magnetic Signature (at least two) around the device. Afterwards in user identity phase, new signature samples can then be recognized/verified against previously recorded patterns. The demo application can also provide a confidence score indicating the level of match between a new signature sample and registered templates. The demo application uses dynamic time warping (DTW), a template matching approach to match between signature samples and registered templates. The accuracy of the signature recognition application using 2 registered templates of users' signature is up to 92%. The application also allows the user to adjust sensitivity of the algorithm for verifying signatures by filtering the magnetic signals.

VIII. CONCLUSIONS AND FUTURE WORK

In this paper, we have studied a new simple yet effective technique for identification/authentication based on what we call as "3D Magnetic Signature" or "MagiSign". The user can use 3D space around the device to flexibly create 3D signatures using a properly shaped magnet. The 3D magnetic signature provides a wider choice for authentication as it can be flexibly drawn in 3D space around the device. Unlike regular signature, no hardcopy of the magnetic signature can be easily produced, resulting in higher security. "MagiSign" is a touch-less way of authentication. It and does not impose major changes in hardware or physical specifications of mobile devices. It is only based on a simple, very small and internally embedded sensor.

There are plenty of possibilities for further improving the current system. For instance, users can create 3D Magnetic Signatures using their own personalized magnet. This personalized magnet can be considered as a physical key. Shape, polarity, angle of usage (during signing), and intensity of the magnet can affect magnetic signature pattern. Therefore, the security level of such a signature can be reinforced using personalized magnets e.g., with custom shape, polarity, intensity, etc. We are also interested to run a survey to analysis attacking possibilities by asking the users to imitate the signature of other people.

REFERENCES

[1] "HP iPAQ Pocket PC h5500 User Guide," Hewlett-packard Company, http://bizsupport.austin.hp.com/bc/docs/support/SupportManual/lpia8006/lpia8006.pdf, July 2010.

[2] P. Gupta, S. Ravi, A. Raghunathan, and NK. Jha, "Efficient Fingerprint-Based User Authentication for Embedded Systems," Design Automation Conf., pp. 244-247, June 2005.

[3] N. Aaraj, S. Ravi, A. Raghunathan, and N.K. Jha, "Architectures for Efficient Face Authentication in Embedded Systems," Proc. Design, Automation & Test in Europe, March, pp. 1-6, 2006.

[4] C.C. Leung, Y.S. Moon, and H Meng, "A Pruning Approach for GMM-Based Speaker Verification in Mobile Embedded Systems," Lecture Notes in Computer Science, vol. 3072/2004, pp. 607-613, 2004.

[5] T. Starner, J. Auxier, D. Ashbrook, and M. Gandy, "The gesture pendant: A self-illuminating, wearable, infrared computer vision system for home automation control and medical monitoring," International Symposium on Wearable Computing, 2000, pp. 87-94.

[6] S. Kratz, and M. Rohs, "HoverFlow: expanding the design space of around-device interaction," In Proc. of the 11th International Conference on Human Interaction with Mobile Devices and Services, Bonn, Germany, pp. 1-8, 2009.

[7] L.S. Theremin, "The Design of a Musical Instrument Based on Cathode Relays," Reprinted in Leonardo Music J., No. 6, pp. 49-50, 1996.

[8] M.L. Minsky and S. Papert, "Perceptrons," Cambridge, MA: MIT Press, 1969.

[9] P. Keir, J. Payne, J. Elgoyhen, M. Horner, M. Naef, and P. Anderson, "Gesture-recognition with Non-referenced Tracking," 3D User Interfaces 3D User Interfaces (3DUI'06), pp.151-158, 2006.

[10] Y. Wu, and T.S. Huang, "Vision-Based Gesture Recognition: A Review," Proceedings of the International Gesture Workshop on Gesture-Based Communication in Human-Computer Interaction, pp. 103-115, 1999.

 ISBN: 978-1-61208-100-7

Classification of Mobile P2P Malware Based on Propagation Behaviour

Muhammad Adeel, Laurissa Tokarchuk, Muhammad Awais Azam
School of Electronic Engineering & Computer Science, Queen Mary University of London,
London E1 4NS, United Kingdom
{muhammad.adeel, laurissa.tokarchuk, muhammad.azam}@elec.qmul.ac.uk

***Abstract*—With a multifold increase in the number of mobile users over past few years, mobile malware has emerged as a serious threat for resource constrained handheld devices. From experience of the Internet malware attacks like *CodeRed* and *Slammer*, it may not be difficult to predict the extent of devastation mobile malware could potentially cause. Numbering around 700 today, detection of mobile P2P malware may prove a serious challenge considering scarce memory, processing and battery resources of handheld devices. Issue may worsen if the detection takes place on mobile devices. Thus there is a strong need of identifying commonalities between various kinds of mobile malware to reduce the detection footprint. As a novel contribution, this work discusses various possibilities of classification of mobile malware and proposes a technical behaviour-based classification that could help detect a range of malware families in real time based on their behaviour during various stages of an attack.**

Keywords- Mobile P2P; Malware classification; Behaviour identification; Mobile malware families

I. INTRODUCTION

There exist over two billion mobile phones in the world today. Statistics from a survey conducted by Dong *et al* [1] reveal that Symbian is the leading operating system in terms of market density with 63% of the market share followed by Windows OS with 16% market density and Palm OS with 10% market penetration. Substantially large penetration of Symbian OS makes it a hot target for mobile worms and viruses. There are over 400 various kinds of mobile malware and around 700 of their variants discovered so far while approximately 90% of this malware targets Symbian-based handhelds [1]. It is difficult to develop an electronic system that detects all of these viruses as they use different strategies to attack the system. Mobile viruses and worms are known to have commonalities in terms of their behaviours however, no technical categorization of such malware exists to-date [2].

Besides common propagation avenues (i.e. MMS & SMS, Bluetooth and Mobile Internet), there are many other ways the malware could propagate in mobile P2P networks. Services like GPRS allow mobile devices to create IP connections with remote servers through cellular vendor's network. This may allow an adversary to take advantage of inherently weak defenses of resource constrained mobile devices. Use of WLAN on handhelds may also put smartphones at risk from various kinds of security threats [3]. Copying files to mobile devices through removable media such as SD cards has proven dangerous with regards to virus replication. Email applications and instant messaging can also act as an avenue for malware propagation while web browsing on handhelds can be dangerous in terms of download and execution of malicious code on mobile device. Damages due to malware propagation through any of the means above can range from loss of privacy and transfer of unsolicited information to the system malfunctioning and failure. Malware causing service disruptions and economic losses can be termed critical though. Figure 1 gives an overview of threat levels of prominent mobile malware by different antivirus companies.

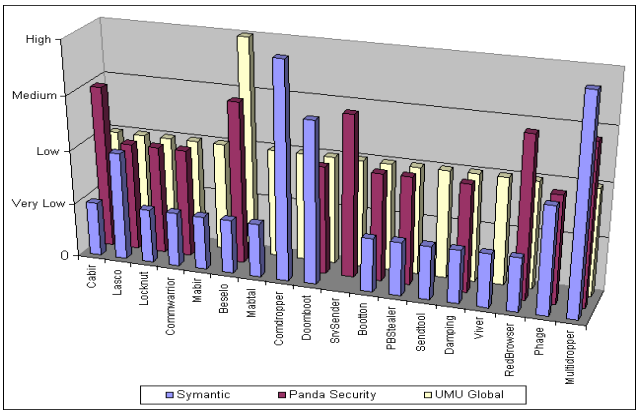


Figure 1: Malware Rating by Antivirus Companies

This work mainly intends at giving new dimensions to the classification of mobile P2P malware and discusses novel contributions in terms of technical classification of malware in Section 4 and the modifications to an existing classification mechanism proposed by Kim *et al* [4] in Section 3. Section 2 discusses another classification of malware based on propagation technologies. Although a very generic and rather theoretical classification of mobile malware could be based on operating system alone however, we keep this work focused on technical classification that could act as a baseline for detection of malware in real time.

II. CLASSIFICATION BASED ON PROPAGATION TECHNOLOGY

This section classifies mobile worm infections in terms of propagation technologies i.e. Bluetooth, MMS/SMS and Internet. It also paves the way for more elaborate technical categorizations in coming sections.

A. Infections through Bluetooth

3G/4G mobile devices are usually equipped with short-range transmission technologies like Bluetooth and Infrared. This allows them to communicate directly with other devices nearby rather than through communication via a cellular services provider's network. Bluetooth

 ISBN: 978-1-61208-100-7

technology can be deemed as one of the contributing factors that gave rise to the concept of mobile P2P networks however, it could also be proclaimed as a major factor behind propagation of peer-related mobile malware in handheld devices. Bluetooth-based malware propagates using Bluetooth capabilities of mobile phones and exploits vulnerabilities of Bluetooth technology to cause catastrophes in mobile P2P networks. Bluetooth technology is known for its inherent security vulnerabilities, Bluetooth and others short-range technologies like Infrared open new avenues of threat dissemination from neighbours. Figure 2 gives a pictorial view of the propagation strategy adopted by the worms like Cabir [5], Metal Gear [6], PBSteal [7] and Lasco [6] mainly using Bluetooth technology. Bluetooth data transfer directly between P2P handhelds makes these resource-constrained devices and the mobile network extremely vulnerable to worm attacks.

Once infected through its mobile peer, a victim will attempt to propagate malware further through the same strategy. Victim not only suffers in terms of battery drain but also in terms of infection to SIS or system files. Infected applications or even operating system may not function properly, hence leaving a mobile functionally dead. Variants of such mobile worms may also propagate secret mobile information to other devices through Bluetooth.

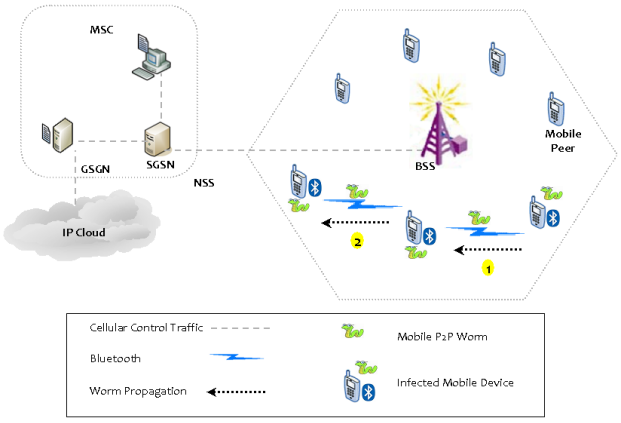


Figure 2: Bluetooth Worm Attack

B. *Infections through SMS & MMS*

MMS and SMS can be considered primary services in terms of mobile usage. Cellular service providers are offering enticing packages to attract customers use more MMS and SMS services. These services however could also be used as launching pad for different kinds of worm attacks in mobile networks. Worms like Mabir [8] and Commwarrior [5] propagate infection through MMS messages while malware like Mquito [9], Wesber [10] and RedBrowser [5] send premium rate SMS messages and incur costs on victim mobiles. An important motive of the attackers is to incur cost on customer. Mobile worms like Mabir and Commwarrior are capable of propagating through MMS thus giving worm propagation a global perspective. Figure 3 illustrates the attack scenario in which an infected mobile node can infect another mobile through a malicious MMS sent via MMS server. Variants of SMS worms besides sending premier-rate SMS messages could also disclose a mobile's private information to its neighbours.

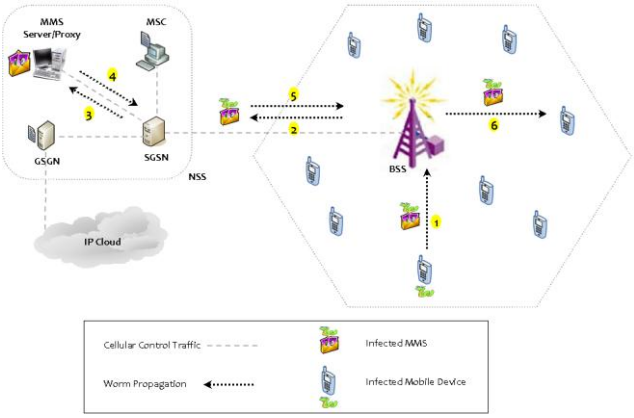


Figure 3: MMS & SMS Worm Attack

C. *Infections through Fixed P2P Networks (Mobile Internet)*

A key attraction in use of mobile P2P networks is a tempting large repository of free downloadable content over World Wide Web. Besides mobile P2P applications like MBit and PeerBox, mobile peers can also interact directly with peers on fixed P2P networks. CDMA and GSM based 3G cellular networks offer higher data rates with rather reduced costs for downloading content. This entices more mobile customers to access P2P content through mobile Internet and hence become vulnerable. Doomboot [12], BBProxy [11], CARDTRAP [14], Metal Gear, PBSteal and RedBrowser are typical examples of malware that is downloaded onto mobile peers this way. Authors in [13] propose an architecture in which mobile peers are no different than fixed peers if a few P2P-specific servers are deployed in cellular vendor's network. Their framework enables mobile users transparently download P2P content from the Internet however it might put mobile peers at a direct risk of security attacks from Internet. Figure 4 illustrates another scenario in which worms originating from fixed P2P network could infect a device after downloading malicious content. Infected device could then infect other devices using data entities of the service provider's network.

This category mainly includes the malware that can be downloaded from the Internet while browsing fixed P2P networks. It then has capability to propagate further on mobile P2P network using different propagation strategies. This category of infections also includes worms from the previous two categories, such as RedBrowser, Metal Gear and Mquito that are downloaded through web browsing or accessing P2P content over the Internet. Victim devices are thus turned into launching pads for further attacks.

 ISBN: 978-1-61208-100-7

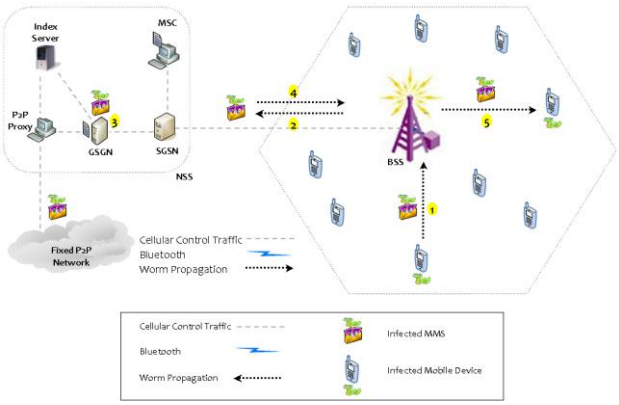


Figure 4: Worm Attack through P2P Core

It was observed that although this classification can act as a baseline for future classifications, it might not be capable of providing a practical categorization of mobile malware capable of detecting malware families in real time. A same worm can potentially use different technologies to propagate during its infection and hence making identification of malware rather infeasible through this way of classification. Thus we proceed to the next section where a rather technical classification of malware is suggested and which to some extent is capable of detecting individual worms rather distinctly during an attack.

III. CLASSIFICATION BASED ON INFECTION SEQUENCE

This section gives a categorization of malware in terms of sequence of operations performed by malware during an infection. Kim *et al* give set of operations for each of the four types of worms they brought under discussion. A set holds operations in a sequence they occur during an attack. Each line in Table 1 gives behaviour of a distinct malware in terms of sequence of operations. We extend the work of Kim *et al* by introducing more operations that enables us to cover further malware types. Rest of this section discuss various operations that could form the basis of detection of distinct malware types.

- *EXEC:* This operation involves execution of malware code on a node to infect it and initiation of the process of propagation of malware. Generic SI model of propagation of malware assumes that every node that downloads the malware should execute it and thus get infected. Every infection starts with EXEC phase however EXEC may have different consequences in different types of malware.
- *CRT_MSG:* To transmit the infection, infected payload is created and in subsequent phases forwarded to the vulnerable devices. Payload for Bluetooth based infections like Cabir differs from SMS based infections like Mosquito.
- *BT_SCAN: In* Bluetooth based infections like Cabir, Lasco & PB Stealers, infected file created through CRT_MSG is propagated to all the Bluetooth devices in coverage are (i.e. in neighbour set). Hence a Bluetooth scan must takes place through Service Discovery Protocol [15] to discover the active Bluetooth neighbours.
- *PB_SCAN:* Infection through MMS propagation like Commwarrior, Mabir and Beselo involve selection of contacts (i.e. victims) from the phonebook for onward transmission of malware. Browsing of the phonebook for this sole purpose is covered under PB_SCAN.
- *RD_PRM:* Worms like RedBrowser, Mosquito & GameSat send premier-rate SMS messages to specific phone numbers for the purpose of financial gains against victim. Attempt to read a premier number from the memory may be another significant operation in terms of detection of such kinds of malware.
- *SND_BT:* After scanning for active Bluetooth devices around, malware is transmitted to those devices. SND_BT records the device IDs and the total number of Bluetooth enabled devices to which the malware is being propagated. Section to come will reveal that the number of devices to which the malware is propagated plays an important role in identification of a family.
- *SND_MMS:* Following *EXEC*, *CRT_MSG* and *PB_SCAN*, next step in the MMS based infections is to send the infected multimedia-message to the selected contacts from the phonebook. Cellular network radio interface is used to get the messages delivered. Again, the number of contacts to which the MMS is being sent plays an important role in identification of malware families in future detections.
- *SND_SMS:* For the SMS based malware families, premier rate SMS messages are sent form the infected device using Short Messaging Service Centre (SMSC) [16] of the cellular network. Apart from the SMS infection scenario above, SND_SMS also executes in other variants of SMS based malware that repeatedly send random messages containing personal information (suppose phonebook entries) of an infected device to its neighbours.
- *CPT_BNRY:* Worm families like Doomboot besides other MMS and Bluetooth related operations also corrupt system binaries on an infected device. Consequently, the victim device could fail on reboot. This unique operation may play a vital role in aimed future detections of this kind of families.
- *LOG_SCAN:* Worm families like Lasco constantly scan the call-logs of infected mobile and reply any incoming message with copies of the malware. LOG_SCAN operation could distinguish this type of malware from others.
- *USR_IDLE:* This operation is aimed at the detection of user idle time by logging data about key-presses and interaction of user with the device. Malware may propagate without the knowledge of phone user and logging such

 ISBN: 978-1-61208-100-7

information may prove very effective in prevention of various malware related attacks.

Building up on the work of Kim *et al*, Table 1 gives sequence of operations for two new families i.e. Call-Loggers & Premier Chargers to be discussed in next section on lines 4 and 5.

TABLE 1. CLASSIFICATION BASED ON SET OF OPERATIONS

1. EXEC › CRT_MSG › BT_SCAN › SND_BT
2. EXEC › CRT_MSG › BT_SCAN › SND_BT › PB_SCAN › SND_MMS
3. EXEC › CRT_MSG › BT_SCAN › SND_BT › PB_SCAN › SND_MMS › CPT_BNRY
4. EXEC › CRT_MSG › BT_SCAN › SND_BT › PB_SCAN › LOG_SCAN › SND_MMS
5. EXEC › CRT_MSG › RD_PRM › SND_SMS

Although we have succeeded in extending the work of Kim *et al* in terms of adding new families and operations, the basic limitation of their work is its inherent incapability of detection of malware families as it purely follows a malware-specific classification model. For instance, their model fails in distinguishing between Commwarrior and Mutational Commwarrior (contains Beselo & Disco worms) families. A detection model based on their classification model may require definition (as in Table 1) for every single malware discovered. Maintaining such a memory-intensive up-to-date database may not be feasible on resource constrained mobile devices and hence we propose a classification model capable of acting as a baseline for detection of malware families.

IV. CLASSIFICATION BASED ON BEHAVIOUR DURING ATTACK

After attempting to classify malware based on the transmission technology and then on the sequence of operations a malware perform under an attack, this section gives a classification of malware based on their behaviour. Table 1 would suggest that although every set distinctly elaborates the behaviour of a particular malware, most of the operations in the database (Table 1) are redundant. Hence rather than selecting the whole set to describe a malware type, under this classification, we select key classification features pertaining to a group of various malware and name them as *flags*. Feature extraction and flagging mechanism is explained through Figure 5. Every malware family exhibits one or many characteristics named as flags in lower part of Figure 5. Sequence of occurrence of flags will eventually determine a malware family. Feature extraction also results in considerable reduction of the malware behaviour storage footprint.

Step 1 of the classification based on behaviours is the identification of the core threat conditions during an attack and setting an appropriate *flag* to *High* if that conditions becomes to true while Step 2 will be to see that to what family this flag or the combination of flags belong to, thus declaring an appropriate alarm. Section below explains some of the extracted behaviours (pertaining to various classes of malware) that would help distinctly identify a malware family. These extracted behaviours are called *flags.*

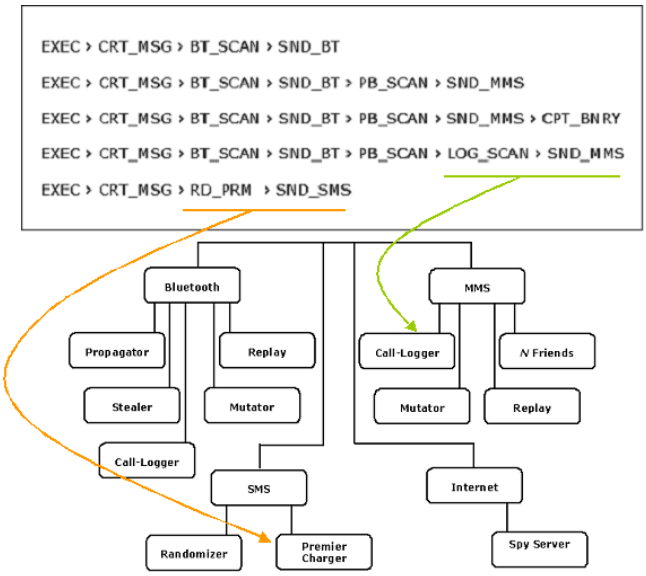


Figure 5: Classification Based on Behaviours

A. BT Propagator

Worms like Cabir, Lasco, PBStealer exhibit this threat condition as do a few malware droppers like Cdropepr [17] and MGDropper [18]. BT_SCAN alongwith SND_BT triggers this flag while the conditions like malware replication to all the active neighbours and repetition of this flag confirms the existence of a malware activity relating to a Bluetooth malware family.

B. BT Mutator

BT Mutator flag corresponds to the malware that uses Bluetooth technology for its replication onto its neighbours and mutates its signature after little iteration. Variants of Cabir family use mutational strategies to remain undetected during attack. Most of the signature-based detection techniques fail in detecting such mutations however the proposed behaviour-based detection is inherently capable of detecting them.

C. BT Replay

This class encompasses the malware that uses Bluetooth technology for propagation to active neighbours of an infected device. Trojans like PBSteal repeatedly send copies of phone-critical information to first connecting neighbour in the list. Alongside stealing phone-critical information, the main motive of such malware is to flush the battery power of infected mobile device as well as the recipient of this transmission. Some dropper like SendTool [19] that eventually drop PBSteal into the target victim's inbox can also be put into this category.

D. BT Stealer

On of the main motive of the malware exhibiting this behaviour is to disclose the phone and user critical

 ISBN: 978-1-61208-100-7

information to the neighbours around. Malware like PBStealer [20] could be put in this category that after infecting a device, makes this devices transmit confidential information to all its neighbours and repeats this behaviour with constant intervals. PBSteal and SendTool could also be put in this category of malware. Relying on various other aspects beyond the scope of this discussion, detections in future stages of this project will be capable of distinctly identifying malware from *BT Replay and BT Stealer* families.

E. MMS N Friends

MMS ***N*** Friends is one of the most important threat conditions in which an infection is propagated to a fixed number of contacts selected from the phonebook of a device. This operation is repeated with constant intervals (in most of the cases interval gap is 10 minutes). 16 Variants of the Commwarrior family and droppers like Commdropper [21] could be put in this category.

F. MMS Replay

One of the consequences of the malware exhibiting this condition is depletion of battery resources of infected and target mobiles. Infected mobiles send repeated copies of the same multimedia message to same mobiles every few minutes. As MMS is a premium service, cost incurred by the victim as a result of this replication could be way too high. Such kind of MMS replay attacks have resulted in various DoS attacks on MMS servers in the past [22].

G. MMS Mutator

Even with MMS based malware families, mutation becomes a key challenge for signature-based techniques. Worms like Beselo and Disco [23] are similar in propagation behaviour to the Commwarrior families but their strategy to mutate makes them qualify to be identified as a distinct family.

H. MMS Call-Logger

Worms like Mabir listen very intelligently to the call logs on an infected mobile phone and reply to the incoming traffic with an infected MMS. Receiving victim device assumes malicious MMS as reply of its message, executes the file attached to the MMS and gets infected.

1. MMS Binary Corruptors

Characteristic that makes this family distinct and rather deadly from the Commwarrior family is that followed by a Commwarrior like infection, it also corrupts the system binaries of the victim mobile thus making it unable to reboot at the next start-up. Worms like Doomboot belong to this family of malware.

J. SMS Premier Charger

Malware families exhibiting this characteristic aim at incurring financial losses to the victim. Making use of SMS technology, the infected devices are made to send an SMS message to premier numbers every few minutes. RedBrowser, GameSat [24] and Mosquito are the most common worms that may fall under this family of malware. Symbian OS Viver [25] is more aggressive as it sends a premier message every 15 seconds and may cause a huge financial loss if remains undetected ever for a shorter time.

K. SMS Randomizer

By constantly listening at the call logs of victim mobile, this class of malware responds to every SMS or call with a random SMS message. Motivation behind this attack is to incur financial loss on the customer. In future though, SMS randomiser can be used to launch more classified attacks like MMS Call-Logger. Symbian trojans like SrvSender [30] belong to this family of malware.

L. Spy-Server

Discovered in January 2010, Ikee worm [29] for the IPhoneOS [26] devices belongs to this category of malware. Ikee makes victim IPhones periodically transmit phone-critical information to a remote server. Cross-platform spyware like Flexispy [27], Mobispy [28] and Blackberry spyware MobiStealth [31] also belong to this category of malware.

V. DISCUSSION & CONCLUSION

Table 2 gives various flags based on which the malware families are identified.

TABLE 2. CLASSIFICATION BASED ON BEHAVIOUR

Flag	Environment	Description
BP	Bluetooth	Bluetooth Propagator
BR	Bluetooth	Bluetooth Replay
BM	Bluetooth	Bluetooth Mutation
MN	MMS	MMS *N* Friends
MR	MMS	MMS Replay
MM	MMS	MMS Mutation
BS	Bluetooth	Bluetooth Stealer
MC	MMS	MMS Call-Logger
MB	MMS	MMS Binary Corruptor
SP	SMS	SMS Premier Charger
SR	SMS	SMS Randomizer
IS	Internet	Internet Spy-Server

Every flag represents a distinct behaviour and characteristic of malware during an attack. A distinct set, sequence or pattern of flags represents the very core functionality of a malware family and thus forms the basis of its detection. Some of the families might not be detectable through one flag alone. BP flag alone if TRUE means alarms about an underway Cabir family infection while a specific pattern of repetition of BP flag confirms this infection. If both BP and MN flags are TRUE, it will prompt a Commwarrior family infection. Similarly if BP and MR flags are TRUE, it represents a Mutational Commwarrior family infection while MR flag alone if TRUE will indicate an MMS Replay family attack.

It was observed that the malware classification based on communication technology alone was not appropriate because different kinds of malware propagating even though similar technology may have varying characteristics in terms of motives, damages and infection strategy. Extensions to the operation database of Kim *et al* although resulted in detection of more malware types however, the model fails to detect the malware families due to its inherent incapabilities. As their solution requires

 ISBN: 978-1-61208-100-7

logging definition of every malware in the detection database, its size may also prove a major concern for resource constraint mobile devices. Based on the example of Commwarrior and Mutational Commwarrior families in Section 3, it was also realized that the traditional signature based techniques may fail in identification of malware families while even some behavioural detection techniques like the one proposed by Kim *et al* may not prove effective in identification of mobile malware families. Alongside considerable reduction the in size of detection database, novel classification based on behaviours proposed in Section 4 has also proved capable of distinctly identifying 12 malware families that accommodate over 200 worms and make about 25% of the total detected mobile malware [32].

REFERENCES

[1] William Aiello & Steven M. Bellovin et al, "Efficient, DoS Resistant, Secure Key Exchange for Internet Protocols CCS'02, November, 2002, Washington, DC USA

[2] Mikko Hypponen, "Mobile Malware", Invited talk delivered at 16th Usenex Security Symposium, Boston, USA, August 2007. http://www.usenix.org/events/sec07/tech/ hypponen.pdf

[4] Hahnsang Kim, Joshua Smith, Kang G. Shin, "Detecting Energy-Greedy Anomalies and Mobile Malware Variants", MobiSys'08, June 17–20, 2008, Breckenridge, Colorado, USA

[3] Dong-Her Shih, "Security Aspects of Mobile Phone Virus: A Critical Survey"

[5] Mikko Hypponen, " Malware Goes Mobile", Proceedings of Scientific America Inc., 2006

[6] Ajay Sharma, " Bluetooth Security Issues, Threats And Consequences", Proceedings of 2nd National Conference on Challenges & Opportunities in Information Technology (COIT-2008), March 29, 2008

[7] http://www.mobile-antivirus.org/Anti-virus-Articles/PBsteal-fix.html [Last accessed 29 April 2009]

[8] http://www.f-secure.com/v-descs/mabir.shtml [Last accessed 29 April 2009]

[9] Neal Leavitt, "Mobile Phones: The Next Frontier for Hackers?"

[10] http://vil.nai.com/vil/content/v_140595.htm [Last accessed 29 April 2010]

[11] http://www.umuglobal.com/encyclopaedia.php [Last accessed 29 April 2009]

[12] http://www.f-secure.com/v-descs/doomboot_a.shtml [Last accessed 29 April 2009]

[13] Andersen F. and Kappler C. et al., "An Architecture Concept for Mobile P2P File Sharing Services", Lecture Notes on Informatics (LNI) P-51, ISBN 3-88579-380-6, Bonner Köllen Verlag, 2004

[14] http://www.f-secure.com/v-descs/cardtrap_a.shtml [Last accessed 29 April 2009]

[15] Service Discovery Protococl http://people.csail.mit.edu/albert/bluez-intro/x290.html[Last accessed 29 April 2010]

[16] N.J Croft and M.S Olivier, "A Silent SMS Denial of Service (DoS) Attack". TechRepublic White Paper, October 2007

[17] http://www.antivirusprogram.se/virusinfo/Cdropper.A_8390.html [Last accessed 29 April 2010]

[18] http://www.f-secure.com/v-descs/mgdropper.shtml [Last accessed 29 April 2010]

[19] SendTool: http://vil.nai.com/vil/content/v_137772.htm [Last accessed 29 April 2010]

[20] PBSteal:http://www.symantec.com/security_response/writeup.jsp?docid=2006-011915-4557-99 [Last accessed 29 April 2010]

[21] http://www.antivirusprogram.se/virusinfo/SymbOS.Commdropper.A_10377.html [Last accessed 29 April 2010]

[22] Stefan Andersson, "MMS Security Considerations", 3GPP TSG SA WG3 Security, Munich, Germany. 18-21 November 2003

[23] http://threatcenter.smobilesystems.com/?p=1180[Last accessed 29 April 2010]

[24] http://www.umuglobal.com/encyclopaedia/114[Last accessed 29 April 2010]

[25] http://www.f-secure.com/v-descs/trojan_symbos_viver_as html [Last accessed 29 April 2010]

[26] www.apple.com/iphone [Last accessed 29 April 2010]

[27] http://www.f-secure.com/v-descs/flexispy_a.shtml [Last accessed 29 April 2010]

[28] vil.nai.com/vil/content/v_139178.htm [Last accessed 29 April 2010]

[29] http://www.f-secure.com/v-descs/worm_iphoneos_ikee.shtml [Last accessed 29 April 2010]

[30] http://www.f-secure.com/v-descs/trojan_symbos_srvsender.shtml [Last accessed 29 April 2010]

[31] http://threatcenter.smobilesystems.com/?p=1868[Last acces-sed 29 April 2010]

[32] Jamshed Sadiq, "Classification of Mobile Viruses", MSc Thesis Report, School of Electronic Engineering & Computer Science, Queen Mary University of London, August 2009.

 ISBN: 978-1-61208-100-7

Link Stability in MANETs Routing Protocols

Crescenzio Gallo, Michele Perilli, and Michelangelo De Bonis
Dept. of Economics, Mathematics and Statistics
University of Foggia, Italy
c.gallo@ieee.org, m.perilli@unifg.it, m.debonis@ieee.org

***Abstract*—Mobile ad hoc networks (MANETs) is a promising communication paradigm for the emerging collaborative environments, which do not need an underlying stable, centralized routing and management infrastructure. In this paper we propose a particular approach for the design of a mobile routing protocol focused on the stability (measured as the transmission intensity change rate) of network links instead of speed and path length, and simulate its adoption in a random network analyzing the corresponding communication graphs.**

***Keywords*-MANET, routing protocol, link stability.**

I. Introduction

Starting from the routing protocols developed in the seventies, valuable work has recently been done in the field of routing protocols for mobile ad hoc networks (consisting of interconnected mobile hosts with routing capabilities), especially since the advent of wireless networks based on UMTS/LTE and WiFi/WiMAX protocols [5] where it is possible to deal with a variable-speed link going from 1 to about 300 Mbps up to 120 Km/h. Because mobile networks can have very unstable links, stability of routes (instead of the only link speed/intensity and path length) becomes a main target in the development of a mobile routing protocol.

In the following section the state-of-the-art is examined, pointing out the fundamental routing protocol issues. Next, Section III introduces the link stability concept in high mobility networks. In Section IV we present the best-path analysis related to link stability, then stating our proposal of an algorithm for route discovery in Section V. A practical network simulation is illustrated in Section VI together with some relevant parameters. Finally, Section VII summarizes and provides some directions for future work in this area.

II. Background and Related Work

In mobile (and hence wireless) ad hoc networks, instead of wireline networks, every node acts both as a router and a host, so the classical "wired" routing protocols are not applicable at all to MANETs.

Existing routing protocols may be classified based on:

- the logical organization through which the protocol "describes" the network. From this point of view they may be divided in *uniform* (all nodes have the same function) and *non uniform* (the way nodes generate and/or answer path control messages may be different for different group of nodes) routing protocols;
- the way routing information is obtained. From this point of view, protocols may be divided in "Proactive", or Table-Driven (such as DSDV Destination-Sequenced Distance-Vector [10] and WRP Wireless Routing Protocol [11]), "Reactive", or On-Demand (such as AODV Ad hoc On-Demand Distance Vector [10], DSR Dynamic Source Routing [11], and TORA Temporally Ordered Routing Algorithm [11]), and "Hybrid" (such as ZRP Zone Routing Protocol [12]);
- how the routing path is created.

There can be a considerable network overhead and computing resource use in a MANET in order to keep track of frequent changes in topology. Protocols of reactive type were designed for these environments, with the aim of not keeping track of network topology [9]. If a node needs to reach a destination, it starts a discovery process [8] to find the path by transmitting broadcast messages of Route Request (RREQ) type, with TTL set to 1 [13]. Each message has a sequence number, so that only the first message is considered, while its subsequent copies are discarded. When a node receives the first copy of a RREQ from a source node, it stores the address, thereby establishing a return path (reverse route). When the first RREQ reaches the destination, a reply message of type Route Reply (RREP) is sent to the source through the return path. This type of protocol is generally efficient for a single rate network; in a multi-rate network, however, what counts is not to minimize the number of jumps to reach a destination, but the total throughput on a given routing.

An existing technique taking into account, instead of the number of hops, the throughput is the MTM (Medium Time Metric) [1]. In this technique a cost inversely proportional to the speed of the link is established; hence, the minimum cost link is chosen. Instead of considering only the cost of the link, its stability should also be considered [4].

A simple model for computing link stability and route stability based on received signal strengths is proposed in [7]. A comparison of various proposed link stability models is made in [8], stressing *route lifetime*. A different approach is carried out by [4], where *signal stability* is used to define link's connection strength. In [2] a mobility metric (*link duration*) is defined, attempting to quantify the effect of node movement in order to develop an adaptive ad hoc network protocol. This last idea is in part also adopted in the

 ISBN: 978-1-61208-100-7

present paper, going further towards the more comprehensive concept of *link stability*, taking into account transmission intensity rate change, too.

III. Routing Instability in Mobile Networks

Although existing routing techniques are of indisputable validity, as a result of lengthy trials conducted in wired networks, a problem causing the performance loss in wireless ad hoc networks (and impacting on the route discovery processes) is the same routing instability, given that we are dealing with high mobility networks. What is "routing instability"? Let us consider a node represented by a mobile phone transmitting while in movement and think how variable is the signal received from a surrounding node as the issuer node moves in a closed or open environment. The level of the received signal, changing constantly, causes a continuously variable ratio of Signal-to-Noise (S/N), altering the bit-rate and consequently the "cost" of the link. This variability would lead to a continuous instability of routing, causing a continuous search of the "best path". This implies an increase in transmission overhead, impacting greatly on the entire network performance and throughput. A technique that keeps track of link stability is now presented, so as to avoid too unstable links in the route discovery process.

A. Keeping track of routing stability

Keeping memory of stability means understanding how stable are connections between nodes; the idea is to have a table maintaining information associated with each link on its state transitions. With the word "transition" you can consider the link's moving from one transmission intensity (measured in dBm and equal to the signal/noise ratio) to another. Table I illustrates each link associated with its number of transitions.

Table I
Link transitions

Link #	No. of transitions
L_1	n_1
L_2	n_2
L_3	n_3
...	...

B. Stability index and thresholds

Let us now analyze what causes the increase in the number of transitions associated with the link. In order to record link's stability, omit all transitions lying within a defined tolerance (those without a significant loss in link performance). The key idea is to record a transition whenever the link's transmission intensity "oscillates too much", i.e., the difference between the new I_i and the previous I_{i-1} sampled transmission intensity relative to I_{i-1} (in absolute value) falls outside a predefined threshold τ. So you keep track of a transition when

$$\left| \frac{I_i - I_{i-1}}{I_{i-1}} \right| > \tau$$

In order to correctly keep track of transition frequency, it is advisable to sum the number of transitions of a link compared to an observation period. For example, if C is the number of transitions in the time interval ΔT, the frequency F will be

$$F = \frac{C}{\Delta T}$$

C. Observation's time interval

To establish a statistical time interval ΔT is not simple. You can guess it to be inversely proportional to the mobility rate of nodes and directly proportional to the number of nodes. Thus, given a network of N nodes, with average nodes' mobility rate μ, you can say that

$$\Delta T = k \cdot \frac{N}{\mu}$$

After this interval the various counters (column "No. of transitions" in Table I) are zeroed.

At the end, a maximum threshold $C_{\max}$ for the number of transitions in the time interval remains to be defined. Consider, for example, a time interval $\Delta T = 300$ milliseconds and a possible maximum value $F_{\max}$ for the transition frequency F of one transition every 60 milliseconds. From that, you may establish for example $C_{\max} = 3 < \Delta T \cdot F_{\max} = 5$. In a nutshell, if there are more than three transitions within an observation period of 300 milliseconds (i.e., $C > C_{\max}$) you will say that the network link is unstable.

In order to practice an effective implementation of the mechanisms given above, one can follow two approaches. The first is to monitor the stability of the link, the second provides for the updating of the link stability table only after a route discovery request. Given the high overhead required by the first approach, it seems preferable to implement the second as better detailed later.

IV. Best Path Choice in Route Discovery

When deciding on the best path, two alternative approaches called "Link Stability" and "Link Rate" are considered and described below in detail.

A. Link Stability

This technique – as the term shows – prefers link stability, and then in the choice of the route to be built it excludes a priori all links having a transition frequency F above a certain threshold. Returning to our example, if $F_{\max} = \frac{1}{60}$ is the threshold corresponding to a transition every 60 milliseconds, all links having $F > \frac{1}{60}$ will be excluded from the choice. Note that, albeit being true that the stablest link has to be chosen, a stable link could also be one with a zero (i.e., not working) signal intensity. Therefore a minimum

 ISBN: 978-1-61208-100-7

threshold $I_{\min}$ should be set for the link intensity, below which the choice cannot be done even if the link is very stable. So, considering threshold values $F_{\max}$ and $I_{\min}$, a network link is stable when $I_i \geq I_{\min}$ (for all $i = 1 \ldots n$, being $n = \Delta T \cdot F_{\max}$ the total number of samples) and $F \leq F_{\max}$.

B. Link Rate

In this technique, stability becomes of secondary importance: link speed is in any case to be preferred. So, when choosing routes for the construction of the best path, the stablest link will be chosen only within those of equal cost (at an equal speed). But what does *equal speed* mean? First, it should be noted that from a practical point of view having two links of the same speed may not correspond to reality, if not for a purely random case. Therefore, two links are of equal speed if the difference in speed between them is no more than 20%. E.g., if the link L_1 has a bit rate $V_1 = 100$ Mbps you can say that a second link L_2 has the same speed V_2 if 80 Mbps $\leq V_2 \leq$ 120 Mbps.

Coming back to the above sketched technique the algorithm, among two links of equal speed, will choose (only under such conditions) the stablest one. To define this stability the same considerations outlined in the previous technique can be done.

V. Our Proposal

From what said, the focus is here on the approach called *link stability*, where the characteristic parameters for network monitoring are highlighted, i.e., the transition frequency of the received signal intensity (dBm) and the signal intensity itself. In the protocol design and implementation a crucial role is played by the link stability table. To optimize efficiency, the table will be updated at the beginning of every route discovery process, and used in the same process to identify the route.

A. The routing process

Each node manages a routing-path table keeping track of all incoming and outgoing connections. The table contains, with respect to a classic routing table, not the mere next hop off an interface but the entire route (that is, all node addresses belonging to the route to destination) and a time-stamp field used to delete the obsolete routes not used for more than a threshold time limit, as shown below.

Observing Figure 1, the application layer sends a route request to the network layer. The routing protocol (interior protocol) checks if the destination address already exists in the device routing-path table. If it does not, the route discovery process is activated in order to enter the destination node address and its path in the routing-path table. If the destination address exists in the device routing-path table, the relative path (included in the routing-path table) will be used and updated by the real time-stamp value of the sender. In MANETs, the paths included in the routing-path table cannot be static since the network topology changes very frequently; the responsibility of route availability is demanded to the upper layer (because there may be an expired timeout waiting for an acknowledgment). The upper layer must still decide whether to delete a route from the table, even if its time limit has not expired: this means that it will request a route discovery every time a packet nondelivery happens.

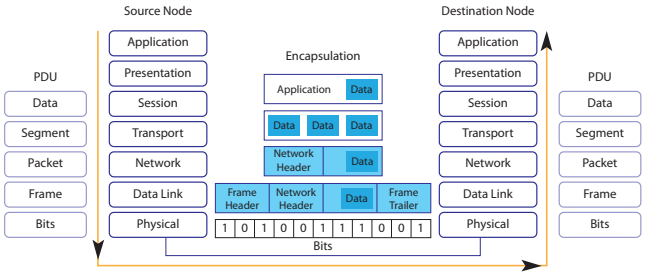


Figure 1. Encapsulation process in communication between nodes

B. Identification of stablest links

To define the stablest links during the route discovery process, a node must collect n transmission intensity values I_i (expressed in dBm) during a statistical time interval ΔT. During this interval, the minimum value I_S of the sampled data I_i will be stored, and the transition counter C will be updated every time the absolute value of the relative difference between the two last observed values is outside the predefined percentage threshold τ.

These values, at the end of ΔT, will be used to check that the received signal intensities I_i are all greater than or equal to a minimum acceptable threshold $I_{\min}$ and if they overcome the percentage threshold τ not too often (i.e., $C \leq C_{\max}$).

If the link is declared stable, a better indicator could be given by a "stability index" $0 < s \leq 1$ given by 1 minus the ratio between the transition counter C with respect to the total number of samples n

$$s = 1 - \frac{C}{n}$$

where $s = 1$ means "highly stable" ($s > 0$ being $C < n$). If the link is unstable, put $s = 0$.

In the example of Table II we assume a maximum transition frequency $F_{\max} = \frac{1}{60}$ and an observation period $\Delta T = 300$ milliseconds. Five intensity signal measurements in dBm are sampled over ΔT (being $\Delta T \cdot F_{\max} = 300 \cdot \frac{1}{60} = 5$) and checked against a predefined minimum acceptable intensity $I_{\min} = -85$ dBm and a relative transition percentage threshold $\tau = 20\%$. A maximum acceptable transition counter $C_{\max} = 3 < 5$ is also established. The transition counter C is increased every time the absolute value of the

 ISBN: 978-1-61208-100-7

relative transition percentage (determined by the last two transmission intensities) overcomes the threshold τ.

Note that the minimum intensity I_{min} is never violated (i.e., $I_S = -80 > I_{\text{min}} = -85$). So, the link is stable because the transition counter $C = 3$ does not exceed the maximum C_{max} too, and its stability index is $s = 1 - \frac{3}{5} = 0.4$.

Should the transmission intensity I_i of the current sample fall under the predefined minimum acceptable intensity I_{min}, then no more samples are collected and the link can be declared "unstable", as shown in Table III.

Table II
LINK STABILITY TABLE – A STABLE LINK

S#	Link transmission intensity I_i (dBm)	Minimum intensity I_S (dBm)	Relative transition % $\left\lvert \frac{I_i - I_{i-1}}{I_{i-1}} \right\rvert$	Transition counter C
1	-50	-50	–	0
2	-70	-70	40.00%	1
3	-80	-80	14.29%	1
4	-40	-80	50.00%	2
5	-70	-80	75.00%	3

$I_{\text{min}} = -85$ dBm, $\tau = 20\%$, $C_{\text{max}} = 3$, $s = 0.4$

Table III
LINK STABILITY TABLE – AN UNSTABLE LINK

S#	Link transmission intensity I_i (dBm)	Minimum intensity I_S (dBm)	Relative transition % $\left\lvert \frac{I_i - I_{i-1}}{I_{i-1}} \right\rvert$	Transition counter C
1	-50	-50	–	0
2	-70	-70	40.00%	1
3	-90	-90	21.43%	2

$I_{\text{min}} = -85$ dBm, $\tau = 20\%$, $C_{\text{max}} = 3$, $s = 0$

C. Route Discovery packet fields

The Route Discovery packet contains the following fields:

- destination node address;
- sender node address;
- sender node time-stamp;
- hop-count (number of links, or nodes, passed through);
- number of stable links;
- pointer to a stack containing addresses of nodes traversed from the sender (bottom) to the recipient (top).

D. Route Discovery algorithm

The Route Discovery process can be summarized as follows.

1) Every node initiating a transmission activates a route discovery process.
2) The transmitter node sends a packet including the destination address and the above mentioned fields.
3) Every node receiving the packet checks if the destination address matches itself.

   Matching. The receiving node stores the return path, and the percentage of stable links over all links traversed. Later, after receiving the first packet, it waits for any other route discovery packet related to the pair sender-timestamp for a specified time ΔT_B. If in this time another route discovery packet arrives, the node compares the percentage of stable links over all links passed through with the previous stored percentage. If it is higher, the new relative path and new percentage will be stored, otherwise the packet will be ignored. All other arriving route discovery packets will be treated the same manner until ΔT_B expires. The recipient, after selecting the best route among all considered in the above said interval:

   a) sends an ACK using the final reverse route. This acknowledgment will be uniquely associated with the route discovery packet transmitted by sender with its time-stamp included;
   b) inserts the reverse route (the winning route) in its routing-path table, binding it to a local time-stamp.

   Reverse-route will be used as long as the routing is valid, i.e., while the recipient is reachable.
   We can consider an improvement in despite of two paths with same percentage value of stable links. In fact, the best path can be better chosen based on the sum of the costs of the links of every path (column *"Cost" of path* of Table VI). The link's cost is the stability index s previously defined and showed in column *Weight* of Table IV. We remember that a link is unstable if $s = 0$. So, the best path will be chosen based on the maximum value of all path costs considered. For this solution, besides the percentage of stable links, the node will store the sum of costs for every path too, and the Route Discovery packet will have a further field reserved for the path cost.

   No Matching. The receiving node checks whether its address is in the stack of traversed nodes:

   a) if yes, it drops the packet, since it is a broadcast route discovery packet previously handled by itself, so the broadcast storm effect will be excluded;
   b) if not, it sends a broadcast route discovery packet to the same destination address adding the node address from which the packet is coming, plus a stable link counter increased

 ISBN: 978-1-61208-100-7

by 1 (in case the receiving node has detected a stable link) and a counter storing all links traversed.

4) Return to Step 3.

VI. Network Simulation

In order to test the assumptions made in our proposed mobile routing protocol we adopted a software network simulator, CNET [3], implementing in the internal layer the route discovery behavior (see Figure 2). CNET employs a simple free-space-loss (FSL) model of signal propagation, with the signal's propagation loss determined by the transmission frequency and the distance between nodes as shown by the formula

$$\mathrm{FSL} = 20\log(f) + 20\log(d) + 92.467$$

where f is the transmission frequency (in GHz) and d is the distance (in km).

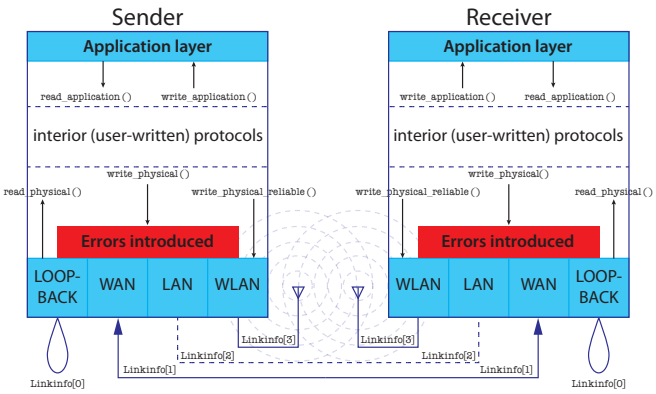


Figure 2. CNET simulation model

We carried out a simulation of a mobile ad hoc network with 18 links and 20 nodes transmitting at 2.45 GHz, with a Tx power of 14.771 dBm and a receiver signal-to-noise ratio of 16.0 dBm. After the simulation, we obtained a set of routes as the result of the *route discovery* processes randomly generated by the hosts of the mobile network. These routes are represented in the undirected weighted graph shown in Figure 3, where edge weight (obtained through a simulation of the transition counter depending on the FSL and Rx values) corresponds to link's stability index as shown in Table IV. The simulated network's properties have been analyzed with the graph utilities available in Mathematica (http://www.wolfram.com/products/mathematica/) and NetworkX (http://networkx.lanl.gov/), according to the indicators introduced by Hanneman [6] and are illustrated in Tables V and VI (note that node M16 has been excluded in the connectivity analysis of the graph). In Table VI, we show the list of paths (routes) for a specific node (M11) along with the "cost" of path, i.e., the sum of edge (link) weights, where an higher value means a better (more stable) path.

In Table V, the parameter *average shortest path length* summarizes the "stability" behavior of the entire network. The average shortest path length is the sum of path lengths $d(u, v)$ between all pairs of nodes (assuming the length is zero if v is not reachable from u) normalized by $n \cdot (n-1)$, where n is the number of graph nodes.

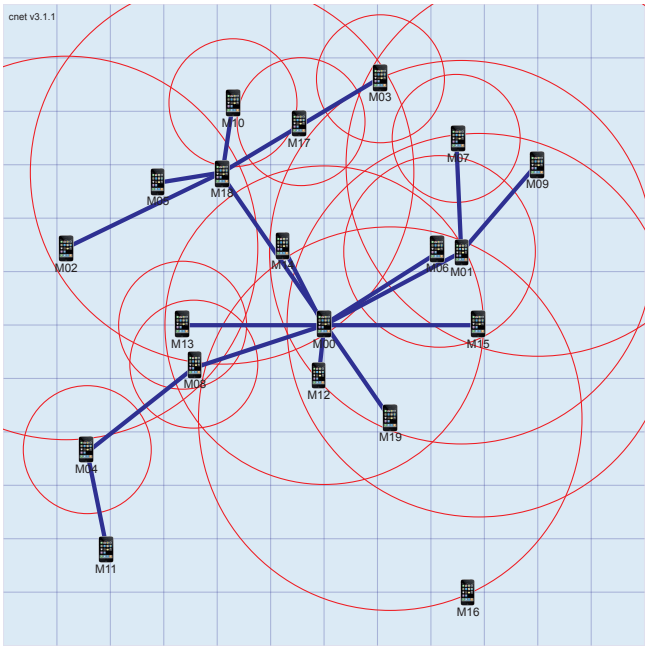


Figure 3. A route discovery simulation with CNET

Table IV
Simulated network links

Link	Distance (km)	FSL	Rx (dBm)	Weight (stability index)
{M00, M01}	0.146	83.561	-64.510	0.184
{M00, M06}	0.128	82.449	-63.398	0.259
{M00, M08}	0.126	82.281	-63.230	0.271
{M00, M12}	0.047	73.760	-54.709	0.848
{M00, M13}	0.132	82.668	-63.617	0.245
{M00, M14}	0.084	78.801	-59.750	0.506
{M00, M15}	0.145	83.484	-64.433	0.189
{M00, M18}	0.171	84.950	-65.899	0.090
{M00, M19}	0.107	80.880	-61.829	0.366
{M01, M07}	0.106	80.771	-61.720	0.373
{M01, M09}	0.108	80.964	-61.913	0.360
{M02, M18}	0.162	84.441	-65.390	0.125
{M03, M18}	0.172	85.001	-65.950	0.087
{M04, M08}	0.128	82.458	-63.407	0.259
{M04, M11}	0.095	79.878	-60.827	0.433
{M05, M18}	0.060	75.904	-56.853	0.702
{M10, M18}	0.066	76.753	-57.702	0.645
{M17, M18}	0.086	79.033	-59.982	0.491

VII. Conclusion

The simple techniques exposed here are suited to any type of mobile ad hoc network and any kind of speed,

 ISBN: 978-1-61208-100-7

Table V
SIMULATED NETWORK'S GRAPH: NETWORK PARAMETERS

Network density	0.105
Network diameter	1.811
Network radius	0.963
Edge-connectivity	0.087
Degree histogram	0:0, 1:14, 2:2, 3:1, 4:0, 5:0, 6:1, 7:0, 8:0, 9:1
Neighbor connectivity	1:2.556, 2:0.389, 3:0.306, 6:0.389, 9:0.472
Average shortest path length	0.813
Center nodes	M00
Peripheral nodes	M11, M12
Articulation nodes	M00, M01, M04, M08, M18

Table VI
SIMULATED NETWORK'S GRAPH: PATHS FOR NODE M11

From	Path	To	"Cost" of path (sum of weights)
M11	M04, M08, M00	M19	1.329
M11	M04, M08, M00	M15	1.152
M11	M04, M08, M00, M01	M09	1.507
M11	M04, M08, M00, M01	M07	1.520
M11	M04, M08, M00	M06	1.222
M11	M04, M08, M00	M14	1.469
M11	M04, M08, M00, M18	M03	1.140
M11	M04, M08, M00, M18	M10	1.698
M11	M04, M08, M00, M18	M05	1.755
M11	M04, M08, M00, M18	M02	1.178
M11	M04, M08, M00	M13	1.208
M11	M04, M08, M00	M12	1.811

by the definition of the indicated parameters. Therefore, this methodology can probably be implemented in any type of network environment, even in networks with very high density of nodes, as wireless networks in delimited environments such as university campus, airports, shopping malls, etc.

It would be useful to conduct proper simulations to test the described algorithm and obtain significant values for the following parameters:

- the statistical time interval ΔT;
- the number n of samples considered in the time interval;
- the minimum acceptable signal intensity I_{min} in dBm;
- the percentage threshold τ of transmission intensity variation;
- the allowed frequency oscillatory limit F_{max} to define a stable link;
- the % of stable links over total traversed links for routing-path table;
- the sender node (waiting for acknowledgement) time-out to initiate a new route discovery;
- the time-out to declare an "old" route in the routing-path table;
- the recipient node wait-time ΔT_B to receive the route discovery.

It would also be useful to study how these techniques, when implemented, impact on the energy consumption of nodes (as a percentage of the generated network overhead).

REFERENCES

[1] B. Awerbuch, D. Holmer, and H. Rubens. The medium time metric: High throughput route selection in multirate ad hoc wireless networks. *Mobile Network and Applications*, 253–266 (2007).

[2] J. Boleng, W. Navidi, and T. Camp. Metrics to enable adaptive protocols for mobile ad hoc networks. In *Proceedings of the International Conference on Wireless Networks (ICWN '02)*, 293–298, Las Vegas, Nev, USA, June 2002.

[3] *The CNET network simulator (v3.2.1)*. School of Computer Science and Software Engineering, University of Western Australia, http://www.csse.uwa.edu.au/cnet/index.html, accessed 7/Jul/2010.

[4] R. Dube, C. Rais, K. Wang, and S. Tripathi. Signal Stability-Based Adaptive Routing (SSA) for Ad Hoc Mobile Networks. *IEEE Personal Communications*, 36–45, February 1997.

[5] M. Ergen. *Mobile Broadband – Including WiMAX and LTE*. Springer, NY, ISBN 978-0-387-68189-4 (2009).

[6] R.A. Hanneman and M. Ridle. *Introduction to Social Network Methods*. University of California, Riverside, http://faculty.ucr.edu/∼hanneman/nettext/index.html, accessed 7/Jul/2010.

[7] M.-G. Lee and S. Lee. *A Link Stability Model and Stable Routing for Mobile Ad-Hoc Networks*. Lecture Notes in Computer Sciences, Springer Berlin–Heidelberg, Vol.4096:904–913 (2006).

[8] G. Lim. Link stability and route lifetime in ad-hoc wireless networks. In *ICPPW '02: Proceedings of the 2002 International Conference on Parallel Processing Workshops*. Washington, DC, USA: IEEE Computer Society, 116 (2002).

[9] R. Paoliello-Guimaraes and L. Cerda-Alabern. *Adaptive QoS reservation scheme for ad hoc networks*. Lecture Notes in Computer Science, 102–112 (2007).

[10] A.H.A. Rahman and Z.A. Zukarnain. Performance Comparison of AODV, DSDV and I-DSDV Routing Protocols in Mobile Ad Hoc Networks. *European Journal of Scientific Research*, 31(4):566–576 (2009).

[11] E.M. Royer and Chai-Keong Toh. A Review of Current Routing Protocols for Ad Hoc Mobile Wireless Networks. *IEEE Personal Communications*, pp. 46–55, April 1999.

[12] P. Sinha, S. Krishnamurthy, and S. Dao. Scalable Unidirectional Routing with Zone Routing Protocol (ZRP) extensions for Mobile Ad-Hoc Networks. In *Proceedings of IEEE Wireless Communications and Networking Conference (WCNC)*, September 2000.

[13] S. Zou, S. Cheng, and Y. Lin. Multi-rate aware topology control in multi-hop ad hoc networks. In *Proceedings IEEE Wireless Communications and Networking Conference (WCNC'05)*, 2207–2212 (2005).

 ISBN: 978-1-61208-100-7

A Device-aware Spatial 3D Visualization Platform for Mobile Urban Exploration

Matthias Baldauf
User-centered Interaction & Communication Economics
FTW
Vienna, Austria
Email: baldauf@ftw.at

Przemyslaw Musialski
Computer Vision
VRVis
Vienna, Austria
Email: musialski@vrvis.at

***Abstract*—Mobile devices displaying 2D map representations are already commonly used for the exploration of urban surroundings on the move. Even though mobile detailed 3D visualizations promise to be an attractive way to cope with the increasing amount of georeferenced information, their widespread use is hampered by the fragmentation of today's mobile device landscape. In this paper, we tackle this real-world problem by introducing a device-aware 3D visualization service platform. Its core is composed of a rule engine selecting and tailoring a suitable visualization process for a requesting device. While we apply remote-renderings for resource-restrained mobile devices, real-time on-device renderings are applied for high-end devices. Following this device-aware approach, a variety of mobile devices with different technical capabilities can be provided with tailored environmental 3D representations for mobile urban exploration.**

***Keywords*—3D visualization; mobile rendering; location-based service; device-awareness**

I. INTRODUCTION

Mobile urban exploration, the exploration of our local or remote surroundings through spatially-aware mobile devices, is starting to become an everyday-activity. While tourists use their mobile phones to learn more about unknown sights, inhabitants familiar with the environment might be more interested in dynamic, personal information attached to places. The amount of available georeferenced information increases steadily, mainly driven by a phenomenon named 'Volunteered Geographical Information' (VGI) [10]. VGI refers to the creation of geographic information such as place-bound reviews or georeferenced photos by individuals made possible by technological advances such as mobile phones with built-in GPS receivers.

Whereas in common location-based services (LBS), 2D maps are the currently most applied method to visualize a user's surroundings and present so-called points-of-interest (POIs), mobile 3D environment representations are attaining increasing interest both in academia and industry. This development is driven by several emerging trends affecting all necessary components of a mobile 3D LBS: the end devices, the digital content, and the environmental models.

First, mobile devices able to render complex 3D scenes in reasonable time are available and affordable for mass market. Second, with more and more georeferenced content available, the usage of the third dimension is one consequential step to better organize and display this information. Efforts are made towards meaningful semantic models where digital information is not simply bound to a coordinate pair but instead attached to real-world objects. Third, only today, the underlying 3D environmental models are easily available. Buildings and complete urban models reconstructed by laser scans and photogrammetry can be obtained from commercial providers of map data such as NAVTEQ [23] or Tele Atlas [33]. Besides such professional activities, even user-driven VGI approaches such as Google's SketchUp [12] enable the creation of 3D building models.

Even though all necessary components are ready, one circumstance still hinders mobile 3D LBS from a widespread deployment: the fragmentation of today's mobile device landscape. Users are equipped with a variety of different client devices reaching from high-end smartphones to cellphones with low-quality displays and very limited processing power. To address this challenge and pave the way for a broad usage of mobile 3D urban representations, this paper introduces a device-aware 3D spatial visualization platform. In contrast to related projects focusing on performant visualizations for one special device, our approach considers the aforementioned device fragmentation and offers appropriate adaption mechanisms enabling advanced spatial visualizations for a variety of mobile devices.

The remainder of this paper is structured as follows. Section II summarizes related work in the field of mobile spatial visualization. In Section III, we define several requirements to enable the aimed widespread deployment. Section IV introduces the components of the proposed visualization platform. The different visualization techniques applied to different end devices are explained in Section V. Finally, we summarize the presented work and draw conclusions in Section VI.

II. MOBILE SPATIAL VISUALIZATION

Due to mobile devices' inherent limitations such as smaller displays and limited processing power, the design and implementation of efficient mobile visual interfaces for spatial information is a challenging task [7]. Related work in the field of mobile spatial visualization can be divided in third person

 ISBN: 978-1-61208-100-7

views making use of 2D maps and 3D representations, and first person views.

A. 2D maps

The projects Cyberguide [1] and GUIDE [6] were one of the first mobile location-aware exploration tools using abstract 2D maps for displaying the user's location and additional spatially-referenced information. Over the last years, a lot of special-purpose mobile guide systems based on 2D maps were developed (e.g., [4], [17], [18], [28]). Today, one of the most well-known public 2D map tools for mobile phones is "Google Maps for Mobile" [11].

B. 3D representations

Bridging the gap to 3D representations, 2D map tiles can be used as ground textures to enable a so-called bird's eye view, i.e., the point of view is elevated and tilted by 45 degrees. This type of visualization is especially favored in car navigation solutions. In addition, researchers used such 2D map textures enhanced with exposed 3D cuboids symbolizing conspicuous landmarks [19].

In the last years, technological advances enabled 3D representations of urban surroundings even on mobile devices (e.g., [5], [9], [20], [25], [30]). Some researchers combined the 3D model with an additional 2D map ([20], [30]) providing a hybrid view. In the meantime, similar products reached the mass market. One of the first public available 3D city guides is "Mobile 3D City" [24].

The project that is most related to the work presented in this paper is NexusVis [22], a distribution visualization framework which also addresses the challenge of adapted spatial representations. However, NexusVis leaves the decision about a suitable visualization process to the client devices whereas the platform proposed in this paper applies a rule-based selection on the server side and thus, enables an easy integration of novel devices. Additionally, NexusVis focuses on portable computers and does not consider the highly fragmented handheld market with the devices' manifold peculiarities.

C. First person views

In addition to map-based representations in third-person views, egocentric first-person approaches to present spatial information have been developed. ViewRanger [3] is one example that provides mobile users with a simplified rendered 3D panoramic view of mountain landscapes. Google's Streetview [13] is another panoramic approach but relies on ready-made 360° photos expensively collected by cars equipped with special cameras.

A latest approach to first-person views are mobile augmented reality systems (e.g., Layar [21], [31]). Here, the live video-stream delivered by the device's camera is overlaid by referenced appropriate information. As such applications solutions may augment only the user's currently visible surroundings they are not capable to support the mobile exploration of remote places.

III. Requirements

Mobile 3D LBS pose a variety of requirements to the underlying technical infrastructure ([34], [35]). To enable a widespread usage and a large penetration of such advanced mobile spatial representations, even additional requirements have to be met. By analyzing the aforementioned related projects and surveying the current telecommunication landscape we identified necessary features. In discussions with the local municipal GIS department as a central stakeholder that provided us with sample data for the prototype proposed in this paper, we completed and verified the following list of key requirements:

- *Adaptive representations*. A practical visualization platform should support different visualization possibilities and select the most suitable one with regard to the requesting device's hardware capabilities. An easy integration of new visualization techniques has to be ensured.
- *Location sensing*. Whereas modern smart phones come with built-in GPS receivers to determine their current location, a lot of former mass market phones lack any localization feature. Therefore, a service platform has to include an appropriate localization method to provide such a device with an estimation of its location.
- *Content hosting*. To keep the memory requirements for a mobile device down, maps and models should be hosted at the platform server. Complete map sets and 3D city models require too much space for an installation on low-end devices and complicate an installation on smartphones. Furthermore, a centralized hosting eases the maintenance and updating of the content.
- *Data standards*. In recent years, standards for 3D city models emerged, e.g., CityGML [8]. A practical service platform must support existing data standards to enable an easy integration of additional content such as environmental models.
- *Data protection*. 3D city models are still expensive to create and maintain and are often subject to copyright restrains. Appropriate visualization methods must take this issue into account and appropriate mechanism must prevent unwanted access to the platform's services.

This list is not intended to be exhaustive. Additional relevant aspects such as privacy issues are beyond the scope of the paper focusing on an universal, device-aware 3D visualization platform. In the following section, the proposed platform's overall architecture and its components are described.

IV. Rule-enhanced spatial service platform

Considering the aforementioned requirements, we designed a rule-enhanced spatial service platform for 3D LBS. Figure 1 depicts the proposed platform's architecture. It provides all the necessary components in order to support a variety of differently equipped mobile devices with 3D urban representations: a 3D model with device-aware rendering possibilities, a database with georeferenced content, as well as auxiliary services exposed via a Web interface protected by HTTP

 ISBN: 978-1-61208-100-7

basic authentication. All service components and the mobile application are written in Java, the server-side components processing 3D model data are implemented using the .NET framework.

A. 3D model management

A fundamental element of the platform is the hosted 3D city model. Currently, we use a detailed, yet untextured 3D model of Vienna's first district, which was provided in the CityGML format by the local municipal department for urban surveying. CityGML is a XML based format designed to manage and store entire urban areas. It provides respective tags to define and store 3D meshes of cities in a hierarchical manner. The buildings are composed of five levels of details, which span from the simple block model of the footprint up to façade details. In our case, we use the first three levels of the model, which provides all buildings, roofs as well as the coarse façade features.

In order to efficiently handle the model for later processing, the provided CityGML model is read into a custom data structure at the platform's startup or at intended later updates, respectively. Importers for further 3D city formats can easily added. Internally, the model then is held in the memory as an enhanced triangle-mesh with full vertex-topology and triangle kd-tree and BSP tree computed. Hence, this data structure is capable of fast intersection or occlusion tests as well as of extracting pieces of the model in real-time on demand.

B. Rule-based rendering

The core functionality of the proposed platform is a rule-based rendering process. Dependent on the capabilities of the requesting mobile device, the client is provided either with a server-side rendered panorama image or an appropriately extracted 3D tile for on-device rendering. Furthermore, the device model and its capabilities have impact on the chosen rendering process itself. When a device with limited processing power is provided with a pre-rendered image, the dimensions of the image as well as its compression are adapted to the requesting device's display size. In case of a smartphone, the format of the provided 3D tile is adapted to the requesting device model.

The integration of a rule engine allows an external fine-grained definition of how different requesting end devices are provided with spatial representations. The rule engine and the currently supported visualization techniques are described in detail in Section V.

C. POI query

The POI query service enables the search for georeferenced information. By passing a location and a radius in meters, appropriate POIs can be fetched. The POI information consisting of data such as the item's title, a unique identifier, its coordinates, its media type and a short description is flattened in a simple comma separated text format that can easily be extracted by a mobile device.

D. Visibility detection

The included visibility detection engine [32] is able to restrict a set of POIs to the ones visible at a passed location. An environmental block model is applied to determine the free lines of sight to POIs and remove those items currently hidden by buildings. Each returned visible POI is annotated both with its distance in meters and its direction in degrees with regard to the passed location.

E. Content adaptation

Locative media files may be accessed by their corresponding IDs via a content adaption service. This device-aware service considers the requesting device's display size and accordingly adapts requested images on-the-fly before passing them to the client device.

F. Localization

A network-based localization service enables the detection of a device's location without any built-in localization features such as a GPS receiver. In cooperation with a mobile network operator, this service returns (after the user's agreement) the device's estimated location with an additional value specifying the inaccuracy in meters. If the accuracy is satisfying, the returned location may be used a direct input parameter for querying location-based information. Otherwise, the rough localization result may be refined by the user e.g., on a 2D map to specify her current location.

G. Map service

A third party mapping service is integrated to provide 2D map tiles used for ground textures, manual location refinement and combined 2D/3D views.

V. Device-aware visualization

The first step of the proposed rule-based rendering process is the detection of the requesting device model. We follow the approach of device-adaptive Web sites where the incoming HTTP request's user-agent header is examined to identify the device model and its features and thus, the site's appearance can be tailored (e.g., [2]). In our engine, we make use of WURFL [27] to derive a model's technical capabilities from the device's user-agent string. WURFL provides a comprehensive database containing information about capabilities and features of current mobile devices such as details about the hosted operating system and the screen dimensions. To create, update and evaluate rules specifying which devices should be provided with which environmental visualization, we utilize the rule engine Drools [15]. Originally, Drools aims at the implementation of flexible business logic but can be useful in any dynamic environment where it should be easy to add new conditions.

 ISBN: 978-1-61208-100-7

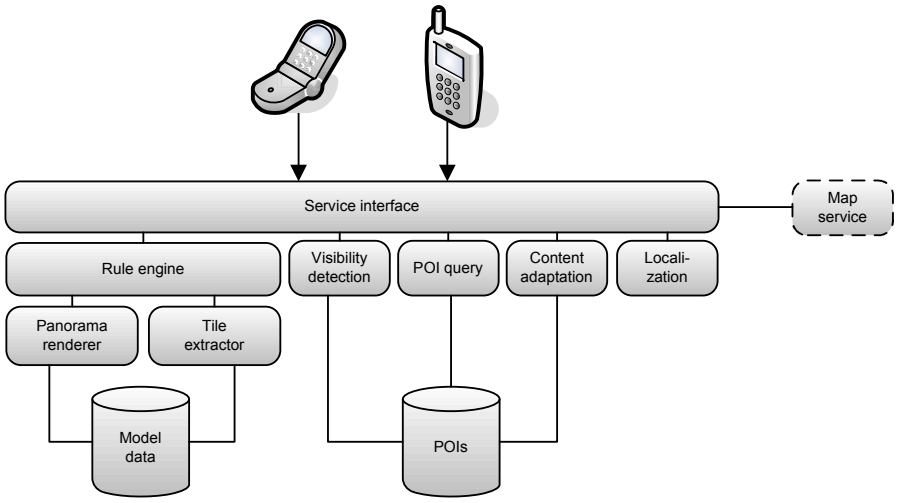


Fig. 1. Components of the proposed device-aware visualization platform.

A. Remote rendering

If our rule-base determines that the requesting device is not capable to perform a real-time-rendering, e.g., based on the involved operating system, the device is provided with a remote (i.e., server-side) rendered 360° panorama image for the desired location. This component provides the limited functionality of a Web Terrain Service (WTS) specified by the Open Geospatial Consortium (OGC) [26].

The panorama is created in two steps: First, we generate a cube-mapping by placing the camera at the desired location in the server-side model and by rendering the six possible cube faces in 3D space. The cameras have the horizontal and vertical field of view angles of exactly 90°. This mapping results in six squares, which cover the entire visible space at the camera position. In the second step we remap the cube faces onto the cylinder side surface of some desired resolution and a vertical angle. Too large or too small angles cause the over or under amplification of the sky and the floor respectively. Usually, it is suitable to use an angle between 100° and 140° to generate a panorama with enough detail. Finally, this rendering can be easily represented as an ordinary bitmap image and can be stored, compressed and transmitted as a PNG stream (Figure 2). The image's dimensions and compression are again determined by the rule engine and tailored to the end device.

To augment the downloaded panorama image with POIs, the mobile device queries the visibility detection engine for georeferenced items in the user's current or (in the case of a remote place's exploration) potential field of view. Passing again the desired location, the client receives a set of items with appropriate distance and orientation values. Having scaled the appropriate semi-transparent POI symbols according to the distances, the icons can be correctly placed onto the panorama regarding the POIs' calculated angles (Figure 3). Scrolling the panorama is possible via the cursor keys. In case of a built-in compass the view is automatically rotated as the user turns.

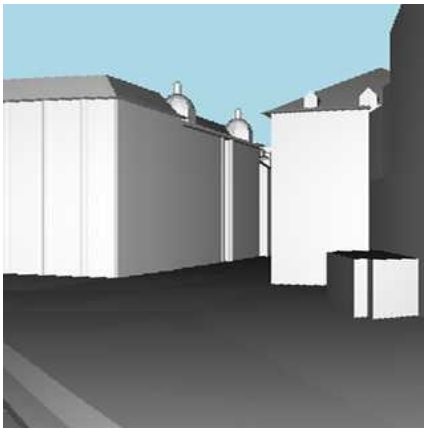


Fig. 2. A small part of the server-side rendered panorama image.

Fig. 3. Device-adapted panorama augmented with selectable POIs.

B. On-Device rendering

If the requesting device is considered to be able to display 3D models at reasonable frame rates, a client-side rendering

(a) (b) (c)

Fig. 4. Tile-based real-time rendering on a smartphone.

technique is applied. Thus, the user is provided with a detailed 3D model where the viewing point can be changed in real-time. By default, the camera shows the user's position (or any other desired city location) in bird's eye perspective elevated by 45 degrees (Figure 4).

This real-time 3D visualization is realized using a tile-based rendering approach, i.e., a displayed urban 3D scene is made up of several single tiles. One 3D tile consists of the appropriate model snippet, a 2D map part used as ground texture as well as a corresponding set of POIs. Each of the three components is downloaded on-demand via the appropriate platform service. Model snippets can be extracted from the model in real-time and can be cached on the server for faster access. Finally, the snippet is exported in a 3D format suitable for the requesting mobile device. Currently, we support exports in the M3G [14] format as well as in a custom text format for OpenGL ES ([16], [29]).

The complete tile is displayed on the device when all three parts are loaded correctly. Currently, one tile spans an area of 100x100 meters. During tests, this size turned out to provide a suitable tradeoff between the covered model area and the arising loading times in a 3G network. Again, locative information and media files are symbolized by accordingly placed semi-transparent icons.

Figure 4(a) shows the client-rendered 3D visualization with its viewing point at the default height. Zooming out reveals the tile-based rendering approach. The scene in Figure 4(b) consists of four tiles. In case, the end device is equipped with a GPS receiver the urban scene is constantly updated while the user is moving and new tiles are loaded when the user approaches the border of the currently displayed scene as depcited in Figure 4(c). The most distant, i.e., no more visible tiles are continuously discarded to efficiently use the device's memory. The availability of a built-in compass enables the scene's automatical alignment with the user's orientation. Additionaly, interaction with the model is possible via the numeric keys or a touchscreen.

VI. Conclusions and Outlook

In this paper, we tackled remaining issues hindering the widespread penetration of mobile 3D LBS. In particular, we addressed the fragmentation problem of today's mobile device landscape by introducing a device-dependent visualization approach.

The platform proposed in this paper includes all the necessary components to provide different end devices with tailored urban visualizations relying on one central 3D city model. The platform's core is composed of rendering modules, which are invoked by a rule engine analyzing the requesting device's technical capabilities. Not only the decision about the appropriate rendering process is device-dependent but so is the actual process itself. In our current prototype we support a server-side rendered 3D panorama and real-time on-device 3D rendering. While the pre-rendering approach takes into account the device' screen dimensions and adopts the panorama's height and compression, the on-device rendering approach exports 3D tiles in a format suitable for the requesting device.

Promising future work includes the device-aware adaption of the 3D model geometry. Automatically reducing a 3D tile's complexity by intelligently removing vertices would allow real-time renderings on even more mobile devices at feasible frame rates. Finally, modern mobile Web browsers promise to provide a future environment for advanced spatial visualizations. Whereas some modern browsers already are location-aware, 2D and 3D drawing functionalities have just started to become included in desktop browsers.

 ISBN: 978-1-61208-100-7

ACKNOWLEDGMENTS

This work has been carried out within the project *WikiVienna* financed in parts by Vienna's WWTF funding program, by the Austrian Government and by the City of Vienna within the competence center program COMET. Additionally, the authors would like to thank Vienna's municipal department for urban surveying (MA41) for providing the 3D city model used in the project.

REFERENCES

[1] G. D. Abowd, C. G. Atkeson, J. Hong, S. Long, R. Kooper, and M. Pinkerton. Cyberguide: a mobile context-aware tour guide. *Wirel. Netw.*, 3(5):421–433, 1997.

[2] A. Artail and M. Raydan. Device-aware desktop web page transformation for rendering on handhelds. *Personal Ubiquitous Comput.*, 9(6):368–380, 2005.

[3] Augmenta. Viewranger. http://www.viewranger.com/vrproductinfo.php. Accessed July 7, 2010.

[4] J. Baus, C. Kray, and A. Krüger. Visualization of route descriptions in a resource-adaptive navigation aid. *Cognitive Processing*, 2(2–3):323–345, 2001.

[5] S. Burigat and L. Chittaro. Location-aware visualization of VRML models in GPS-based mobile guides. In *Web3D '05: Proceedings of the tenth international conference on 3D Web technology*, pages 57–64. ACM, 2005.

[6] K. Cheverst, N. Davies, K. Mitchell, and A. Friday. Experiences of developing and deploying a context-aware tourist guide: the guide project. In *MobiCom '00: Proceedings of the 6th annual international conference on Mobile computing and networking*, pages 20–31. ACM, 2000.

[7] L. Chittaro. Visualizing information on mobile devices. *Computer*, 39(3):40–45, 2006.

[8] CityGML. http://www.citygml.org. Accessed July 7, 2010.

[9] V. Coors and A. Zipf. MoNa 3D – mobile navigation using 3D city models. In *Proceeding of the 4th international symposium on location based services and telecartography*, 2007.

[10] M. Goodchild. Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69(4):211–221, 2007.

[11] Google Maps Mobile. http://www.google.com/mobile/products/maps.html. Accessed July 7, 2010.

[12] Google SketchUp. http://sketchup.google.com. Accessed July 7, 2010.

[13] Google Streetview. http://maps.google.com/help/maps/streetview. Accessed July 7, 2010.

[14] Java Community Process. JSR 184: Mobile 3D Graphics API for J2ME. http://www.jcp.org/en/jsr/detail?id=184. Accessed July 7, 2010.

[15] JBoss Community. Drools - Business Logic Integration Platform. http://www.jboss.org/drools. Accessed July 7, 2010.

[16] Khronos Group. OpenGL ES Overview. http://www.khronos.org/opengles. Accessed July 7, 2010.

[17] C. Kray. *Situated Interaction on Spatial Topics*. PhD thesis, Computer Science Department, University of Saarland, Saarbrücken, Germany, 2003.

[18] J. Krösche, J. Baldzer, and S. Boll. Mobidenk-mobile multimedia in monument conservation. *IEEE MultiMedia*, 11:72–77, 2004.

[19] A. Krüger, A. Butz, C. Müller, C. Stahl, R. Wasinger, K.-E. Steinberg, and A. Dirschl. The connected user interface: realizing a personal situated navigation service. In *IUI '04: Proceedings of the 9th international conference on Intelligent user interfaces*, pages 161–168. ACM, 2004.

[20] K. Laakso, O. Gjesdal, and J. Sulebak. Tourist information and navigation support by using 3d maps displayed on mobile devices. In *Proceedings of the Workshop on Mobile Guides, Mobile HCI*, 2003.

[21] Layar. http://layar.com. Accessed July 7, 2010.

[22] C. Lübbe, A. Brodt, N. Cipriani, and H. Sanftmann. NexusVIS: A distributed visualization toolkit for mobile applications. In *Proceedings of the IEEE Pervasive Computing and Communications Workshops*, pages 841–843, 2010.

[23] NAVTEQ. http://www.navteq.com. Accessed July 7, 2010.

[24] Newscape Technology. Mobile 3D City. http://www.mobile3dcity.com. Accessed July 7, 2010.

[25] A. Nurminen. m-LOMA - a mobile 3D city map. In *Web3D '06: Proceedings of the eleventh international conference on 3D web technology*, pages 7–18. ACM, 2006.

[26] Open Geospatial Consortium. http://www.opengeospatial.org. Accessed July 7, 2010.

[27] L. Passani. WURFL. http://wurfl.sourceforge.net. Accessed July 7, 2010.

[28] G. Pospischil, M. Umlauft, and E. Michlmayr. Designing LoL@, a Mobile Tourist Guide for UMTS. In *Mobile HCI '02: Proceedings of the 4th International Symposium on Mobile Human-Computer Interaction*, pages 140–154. Springer-Verlag, 2002.

[29] K. Pulli, T. Aarnio, V. Miettinen, K. Roimela, and J. Vaarala. *Mobile 3D Graphics with OpenGL ES and M3G*. Morgan-Kaufmann, 2007.

[30] I. Rakkolainen and T. Vainio. A 3D city info for mobile users. *Computers & Graphics*, 25(4):619–625, 2001.

[31] G. Schall, E. Mendez, E. Kruijff, E. Veas, S. Junghanns, B. Reitinger, and D. Schmalstieg. Handheld augmented reality for underground infrastructure visualization. *Personal Ubiquitous Comput.*, 13(4):281–291, 2009.

[32] R. Simon and P. Fröhlich. A mobile application framework for the geospatial web. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 381–390. ACM, 2007.

[33] Tele Atlas. http://www.teleatlas.com. Accessed July 7, 2010.

[34] F. Wang and J. Liu. Towards 3D LBS - challenges and opportunities. In *Proceedings of ISPRS Congress*, 2008.

[35] S. Zlatanova and E. Verbree. Technological developments within 3D location-based services. In *Proceedings of International symposium and exhibition on geoinformation*, 2003.

 ISBN: 978-1-61208-100-7

QoS Aware Mixed Traffic Packet Scheduling in OFDMA-based LTE-Advanced Networks

Rehana Kausar, Yue Chen, Kok Keong Chai, Laurie Cuthbert and John Schormans
School of Electronic Engineering and Computer Science
Queen Mary University of London
London E1 4NS, UK
rehana.kausar,yue.chen,michael.chai,laurie.cuthbert,john.schormans@elec.qmul.ac.uk

***Abstract*— In this paper, a packet scheduling framework is proposed for LTE-Advanced downlink transmission. The proposed framework adds the new functionality of an adaptive TD scheduler with built-in congestion control to the existing conventional quality of service (QoS) aware packet scheduling algorithms. It optimizes multiuser diversity in both the time and frequency domains by jointly considering the channel condition, queue status and the QoS feedback. The framework aims to improve the system spectral efficiency by optimizing the use of available resources while maintaining QoS requirements of different service classes and a certain degree of fairness among users. The results show an improved QoS of Real Time traffic and a fair share of radio resources to Non Real Time traffic types.**

Keywords- Packet scheduling; OFDMA; QoS; LTE-A.

I. INTRODUCTION

Long Term Evolution Advanced (LTE-A) is an all-IP based future wireless communication network that is aiming to support a wide variety of applications and services with different quality of service (QoS) requirements. It is targeting superior performance in terms of spectral efficiency, system throughput, QoS and service satisfaction when compared with existing 3GPP wireless networks [1].

As one of the core functionalities in radio resource management, packet scheduling (PS) plays an important role in optimizing the network performance and it has been under extensive research in recent years. Different PS algorithms have been deployed aiming at utilizing the scare radio resource efficiently. The classic PS algorithms exploiting multiuser diversity are the MAX C/I and Proportional Fairness (PF) algorithms. MAX C/I algorithm allocates a physical resource block (PRB) to a user with the highest channel gain on that PRB, and can maximize the system throughput [2]. The PF algorithm takes fairness among users into consideration and allocates resources to users based on the ratio of their instantaneous throughput and its acquired time averaged throughput [3]. However these algorithms aim only at improving resource utilization based on channel conditions of users; QoS requirements "e.g." delay requirements of real time (RT) service or minimum throughput requirements of non-real time (NRT) service are not considered at all. In the next generation networks, apart from system throughput and user fairness, the crucial point is to fulfill users' QoS requirements in a multi-service mixed traffic environment. This is because different service types are competing for radio resources to fulfill their QoS requirements. To allocate radio resources efficiently and intelligently in such complex environments is challenging. Various methods have been proposed aiming to use radio resources efficiently to fulfill QoS requirements of different traffic types [4][16][17].

In [4], a service differentiation scheme is used which classify mixed traffic into different service classes and grants different scheduling priorities to them. Two types, VoIP and BE are considered and the results show an improvement in RT QoS at the cost of system spectral efficiency, when the RT queue is granted the highest priority. In [5], an urgency factor is used to boost the priority of a particular service. When any packet from a service flow is about to exceed its upper bound of QoS requirement, its priority is increased by adding an urgency factor. Although most of the packets are sent when they are nearly ready to expire, a lower packet loss is achieved thus improving the performance of system by guaranteeing QoS requirements to different services. In mixed traffic scenarios, queue state information (QSI) becomes very important in addition to channel state information (CSI) [6] [15]. A time domain multiplexing (TDM) system based Modified Largest Waited Delay First (M-LWDF) is presented in [6] which takes into account both QSI and CSI. This algorithm serves a user with the maximum product of Head of Line (HOL) packet delay, channel condition and an arbitrary positive constant. This constant is used to control the packet delay distribution for different users. This algorithm is applied in a frequency domain multiplexing (FDM) system in [7] to optimize sub-carrier allocation in OFDMA based networks. It shows improved performance in terms of QoS but like M-LWDF updates the queues state each TTI rather than after each sub-carrier allocation. In [8], M-LWDF is modified by updating the queue status after every sub-carrier allocation. It takes into account RT and NRT traffic types and provides better QoS for both services. The results show an improvement in delay for RT and throughput for NRT service. However this idea can be extended to more intelligent scheduling framework by adding more traffic types and making resource allocation more adaptive based on QoS.

 ISBN: 978-1-61208-100-7

In a multi-service environment, the crucial point is to clearly define the QoS requirements of different services, their demands for radio resources and their channel conditions and queue status to support their demands. Combined consideration of this information can lead to a more efficient PS algorithm, which can be further optimized for network level congestion control by giving QoS feedback.

The work in this paper addresses the scheduling problem in a multi-service wireless environment where the competition to get radio resources is keen and there are strict QoS requirements. A novel PS framework is proposed with added functionalities, to achieve better QoS of different traffic types, a fair share of throughput among users and improved spectral efficiency. The proposed PS framework segregates different types of traffic and sorts users in the service specific queues based on different queue sorting algorithms. A built-in congestion control Adaptive Scheduler is introduced in the TD which makes the system more adaptive to meet QoS guarantees of RT traffic and prevent NRT traffic from starvation. Multiuser diversity in both time and frequency domain are exploited by frequent updating of queue state information and channel condition which leads to a balance prioritizing among users of different traffic types.

The rest of the paper is organized as follows. Section 2 gives a detailed description of the proposed PS framework. Section 3 presents system model and its performance metrics. The simulation model and results are described in Section 4, and finally, conclusions are presented in Section 5.

II. PROPOSED SCHEDULING FRAMEWORK

A schematic diagram of proposed scheduling framework is shown in Figure 1.

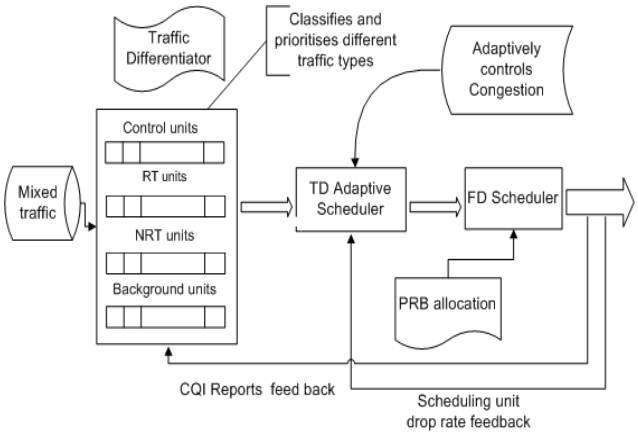


Figure 1. Proposed scheduling framework

The framework is composed of three main units: a traffic differentiator and prioritizing unit, a TD adaptive scheduler with built-in congestion control and a frequency domain (FD) scheduler where resources are mapped to users according to priority order selected in TD. Compared with other PS algorithms, the novelty of the proposed framework lies mainly in the TD adaptive scheduler. However in the traffic differentiator and prioritizing unit, delay-dependent queue-sorting algorithms make a difference compared with the schemes used in reference paper.

The detailed description of the functionality of each unit, the algorithms and policies used in each unit is presented below.

A. Traffic differentiator and prioritizing unit

The need for a differentiator arises when there are different traffic types demanding radio resources with different QoS requirements. In such an environment it becomes very important to classify traffic in service queues to enable queue specific prioritizing schemes to be applied flexibly. Service classification is in fact the first step towards optimizing utilization of available radio resources while dealing with mixed traffic. This is because with complete knowledge of QoS requirements of each class, just enough radio resources can be allocated to these classes. The QoS guarantees become more feasible when radio resources are allocated according to the well-defined demands of traffic types rather than by estimation.

In the proposed scheduling architecture mixed traffic is classified in four queues; Control (control information), RT conversational traffic (voice), NRT streaming (video file download) and background (email, SMS). These queues are chosen for the present study because they cover most of the common data types including low latency, high throughput and low priority. The Background traffic represents the best effort (BE) class of traffic and does not have any QoS requirements. The control traffic is the most important traffic type so it is put into a dedicated queue and served before other traffic types. In the present work control information for downlink (DL) scheduling is considered only as this study is for downlink transmission of LTE-A networks.

In the proposed PS framework, one user is assumed to have one service type and one scheduling unit (SU) carries the information about user, service type and buffer status. The queues in the differentiator are prioritized from top to bottom that is Control, RT, NRT and Background respectively. After differentiation, SUs are sorted within the queues using different queue sorting algorithms. The Control queue SUs are sorted by Round Robin (RR) algorithm because all control information has to be equally important, meanwhile RT, NRT and Background queue SUs are sorted by using queue specific priority metrics.

RT Traffic

The QoS requirement for RT traffic is defined as $d_k < DB_{RT}$ where d_k is delay of user k, DB^{RT} is the delay budget for RT traffic. The delay budget for RT traffic is 40ms [8] [17] in OFDMA-based networks. If this condition is not met then the SUs will be dropped from the queue. A

delay dependent queue sorting algorithm is used for RT users and the priority metric is formed by the product of normalized Head of Line (HOL) delay and the complex channel gain of the users. The Normalized HOL delay is a ratio of user's waiting time and the delay budget for RT traffic. The waiting time of a user is equal to number of transmission time intervals (TTIs) during which the user has not been allocated. The priority of user *k* at time *t*, $p_k(t)$ is

$$p_k(t) = F_k^{RT}(t) \times H_k^{RT}(t) \quad (1)$$

where $H_{k\in K}^{RT}$ is the channel gain of user *k* and $F_{k\in K}^{RT}$ is normalized waiting time of user *k* at time *t* given by.

$$F_{k\in K}^{RT}(t) = \frac{T_{Waiting}}{DB^{RT}} \quad (2)$$

where $T_{waiting}$ is the waiting time, DB^{RT} is upper bound of delay for RT traffic.

In each TTI, the user with the highest priority value is sorted at the front of the queue followed by users with priority value in descending order.

NRT Traffic

The priority metric for NRT streaming video traffic is the product of normalized HOL delay of each user and the ratio of its instantaneous throughput and the average throughput over a given time interval. In this queue the throughput ratio is used instead of channel gain to provide a balance between throughput and fairness. SUs are arranged according to the highest value of this priority metric thus not only satisfying their QoS requirements but also exploiting multiuser diversity in TD. The priority of user *k* at time *t*, $p_k(t)$ is

$$P_k(t) = F_{k\in K}^{NRT}(t) \times \frac{r_k}{R_k} \quad (3)$$

where $F_{k\in K}^{RT}$ is normalized waiting time of user *k* at time *t* and is given by

$$F_{k\in K}^{NRT}(t) = \frac{T_{Waiting}}{DB^{NRT}} \quad (4)$$

where $T_{waiting}$ is the waiting time, DB^{NRT} is upper bound of delay for NRT streaming video traffic, r_k is instantaneous throughput and R_k is average throughput of user k.

The time average throughput of user k is updated by the moving average as below as used in [9] and many other papers,

$$R_k(t+1) = \left(1 - \frac{1}{t_c}\right) R_k(t) + \frac{1}{t_c} \sum_{m=1}^{M} r_{k,m}^{/}(t) \quad (5)$$

where t_c is the length of time window to calculate the average data rate, $1/t_c$ is called the attenuation co-efficient with the widely used value 0.001, $r_{k,m}^{/}(t)$ is the acquired data rate of user k at PRB m if m is allocated to k else it is zero.

Background Traffic

Background traffic has no QoS requirements so priority is given to BE users based only on channel conditions. However to maintain some fairness between users, the proportional fairness (PF) algorithm is used as the queue sorting algorithm for Background queue. The priority of user *k* at time *t*, $p_k(t)$ is

$$P_k(t) = \frac{r_k}{R_k} \quad (6)$$

where r_k is instantaneous throughput, R_k is average throughput of user k as defined previously.

After prioritizing users in the queues the TD adaptive scheduler picks specific proportion of users from the queues.

B. Time Domain adaptive scheduler

This unit aims at guaranteeing the QoS of RT traffic and at the same time ensuring fairness for NRT traffic. It allocates just enough resources to meet the QoS requirements of RT and remaining resources are allocated to NRT services based on the requirements of service types. This scheduler unit enhances the adaptability of the whole framework by collecting the QoS feedback, such as SU drop rate, as its input to make decisions on the TD adaptive scheduling policy selection. The system is said to be in congestion when the QoS of the RT service is not met and due to system load RT SUs are dropping frequently. The TD adaptive scheduling unit is integrated with a built-in policy based congestion control that controls congestion of the system in the network.

The TD adaptive scheduling algorithm works as follows.

Let the total number of available PRBs be denoted by C. If λ denotes the proportion of PRB assigned to RT users and $(1-\lambda)C$ is assigned to NRT users then λ can be adaptively adjusted according to the practical user distribution or QoS of RT traffic. The proportion of capacity given to the NRT traffic for this paper is further divided in different types of the NRT traffic (control, NRT streaming video and Background) such that first the control queue is allocated enough PRBs to deliver control information of all users and then rest of the PRBs are allocated to the NRT and the Background queue. In this way control queue is at the top and is always allocated enough PRBs.

 ISBN: 978-1-61208-100-7

In this paper, three built-in congestion control policies are chosen to exemplify the adaptive capability of TD adaptive scheduler in which the value of λ is changed according to network conditions. The value of λ is changed based on a threshold χ which is set using the drop rate of SUs of RT traffic. When the number of SUs dropped exceeds the threshold χ, the built-in congestion control policy changes accordingly to reduce SUs drop rate. The distribution of the NRT capacity is adjusted according to the buffer status and requirements of NRT service types such as streaming traffic is more important and more frequently requested service than Background traffic and control information is always less than actual data to be sent.

In this paper, the PS algorithm in [4] with fair TD scheduling is considered as reference algorithm. The TD scheduler in [4] uses conventional channel dependent queue sorting algorithms and gives priority to different queues from top to bottom based on fair scheduling or by strict priority. In fair scheduling one user is picked from each queue at a time, starting from top queue and in strict priority queues are emptied completely one by one. In FD, resources are mapped in priority order to the users selected in TD.

C. Frquency Domain scheduler

Resources are actually mapped to SUs in the FD scheduler according to the priority selected in TD. Multiuser diversity is exploited by using channel dependent proportional fair (PF) algorithm in FD. For each SU, the best PRB (with highest throughput) is selected out of available PRBs and is allocated to this SU.

III. SYSTEM MODEL AND PERFORMANCE METRICS

In this work, an OFDMA system with minimum allocation unit as 1 PRB containing 12 sub-carriers in each TTI is considered. The DL channel is a fading channel within each scheduling drop. The received symbol $\mathrm{X}_{k,m}(t)$ at the mobile user k on sub channel m is the sum of White Gaussian Noise and the product of actual data and channel gain as shown below,

$$\mathrm{X}_{k,m}(t) = \mathrm{H}_{k,m}(t)\mathrm{I}_{k,m} + Z_{k,m}(t) \qquad (7)$$

where, $\mathrm{H}_{k,m}(t)$ is the complex channel gain of sub channel m for user k, $\mathrm{I}_{k,m}(t)$ is data symbol from eNB to user k at sub channel m and $Z_{k,m}(t)$ is complex White Gaussian Noise [8]. It is assumed as in [4], [5], [8] and [14] that the power allocation is same, $P_m(t) = P/M$ on all sub channels. Where, P is the total transmit power, $P_m(t)$ is the power allocated at sub channel m and M is total number of sub channels. At the start of each scheduling drop, the channel state information $\mathrm{H}_{k,m}(t)$ is known by the eNodeB.

The channel capacity of user k on sub channel m can be calculated by using Eq. (8) as used in [5][8],

$$C_{k,m}(t) = \mathrm{B}\log_2\left(1 + \frac{|H_{k,m}(t)|^2}{\sigma^2\Gamma} P_m(t)\right) \qquad (8)$$

where B is the bandwidth of each PRB, σ^2 is the noise power density and $\Gamma = -\ln(5\mathrm{BE}R)/1.5$ is the SNR gap determined by bit error rate BER.

In the proposed framework, users are served by one of the differentiated queues depending on their QoS requirements. For example RT users must not exceed their delay bounds, NRT users must achieve their minimum data rate and there should be fairness among Background users. At a given time *t*, PRBs are allocated to users by the following algorithm.

Step 1: Initialize queues for all traffic types and the number of PRBs.
Step 2: Sort users in these queues according to queue sorting algorithms given in equations 1, 3 and 6 for different traffic types.
Step 3: Select a number of users from these queues according to built-in policies in TD adaptive scheduler.
Step 4: Allocate PRB to the user with the highest priority.
Step 5: Remove the allocated PRB from the PRB list and the allocated user from the user list.
Step 6: Go to step 4 if the PRB list is not empty else go to next TTI.

Resource allocation is completed when all PRBs are allocated. The proposed PS framework is analyzed under performance metrics of system throughput, user fairness and QoS of different traffic types.

The system average throughput is the sum of average throughput across all users. To measure the fairness among users, Raj Jain fairness index is adopted that is given as below as used in [10][11],

$$Fairness = \frac{\left[\sum_{k=1}^{K} \tilde{R}_k\right]^2}{K\sum_{k=1}^{K}\left(\tilde{R}_k\right)^2} \qquad (9)$$

The value of fairness index is 1 for the highest fairness when all users have same throughput. In Equation (9), K is the total number of users and $\tilde{R}_k$ is the time average throughput of user k.

The value of SU drop rate and the average delay of RT traffic are used to evaluate QoS of RT traffic. SU drop rate is calculated by the ratio of number of RT SUs dropped to total number of RT SUs. In addition, the average delay for

 ISBN: 978-1-61208-100-7

all NRT traffic is also calculated to prevent NRT traffic from starvation.

IV. SIMULATION RESULTS AND DISCUSSION

Simulation model, results and analysis will be presented in this section.

A. Simulation model

A single cell OFDMA system with total system bandwidth of 10 MHz and PRB size of 180 kHz has been considered. Total system bandwidth is divided into 55 PRBs. The simulation parameters used for system level simulation are based on [12] and these are typical values used in many papers. The wireless environment is typical Urban Non Line of Sight (NLOS) and the LTE system works with a carrier frequency of 2GHz. The most suitable path loss model in this case is the COST 231Walfisch-Ikegami (WI) [13] as used in many other papers on LTE.

Users are assumed to have a uniform distribution and the total number of RT users is assumed to be equal to total number of NRT users as in [8]. Each TTI is 1 ms and the delay upper bound for RT traffic is taken 40 ms which is equivalent to 40 time slots. Total eNB transmission power is 46dBm (40w) and BER is 10^{-4} for all users.

B. Simulation results

The performance of the proposed framework is evaluated by comparing it with the stand alone PF and QoS aware PS algorithm in [4] referred to as the reference algorithm hereafter. All simulations are done in Mat lab.
Figure 2 shows the average delay of RT users with different adaptive TD scheduler policies for 80 active users.

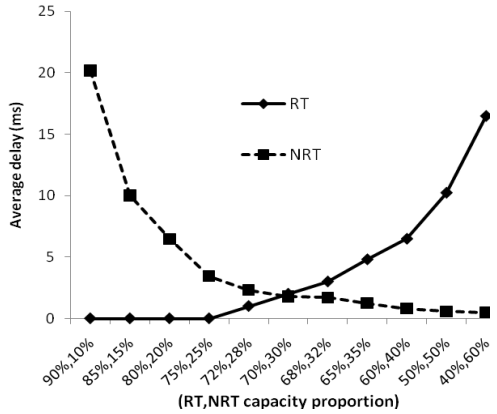


Figure 2. Comparison of TD adaptive scheduler policies

The average delay for RT traffic decreases as RT capacity proportion is increased and increases as RT capacity proportion is decreased. This change in average delay of RT traffic is shown by solid line in Figure 2. On the other hand, by increasing RT capacity, the average delay of NRT traffic does get very high. The average delay of RT and NRT traffic is analyzed under a number of TD adaptive scheduling policies to find a good trade-off so that RT traffic may not exceed its delay upper bound and at the same time the QoS of NRT may be satisfied. For this particular user distribution, the policy (70%, 30%) shows a balance point where both RT and NRT can get reasonable capacity proportion and it is adopted as the default policy in the next results. The proposed algorithm will start with (70%, 30%) policy and will be able to switch to other policies depending on network conditions.

Figure 3 shows the average delay of RT traffic under different system load when reference and the proposed algorithms are used. The standalone PF algorithm has no functionality for QoS of RT traffic that is why it is not included in this analysis.

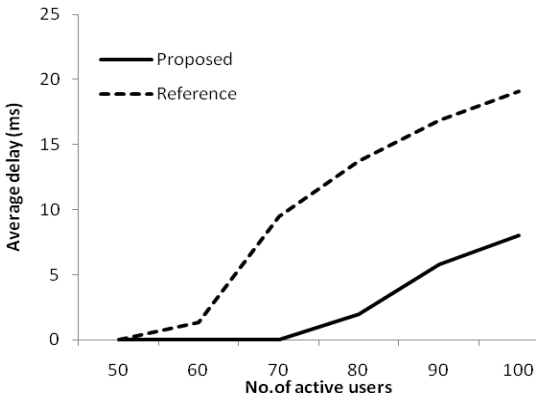


Figure 3. Delay under different system loads

The RT delay increases with system load for both reference and the proposed algorithm. However delay with proposed algorithm remains lower than the reference algorithm as shown. This is because the adaptive TD scheduler in the proposed algorithm adaptively controls the delay of RT traffic. In Figure 4 SUs drop rate for RT traffic under different system load is shown for the proposed and the reference algorithms.

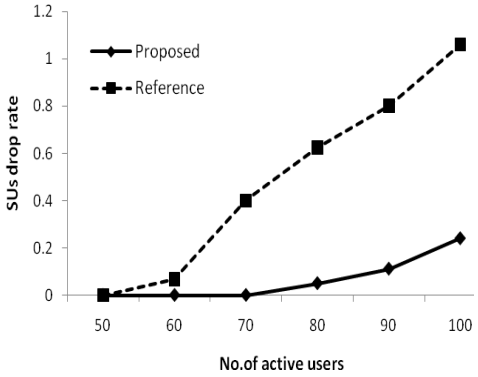


Figure 4. Sceduling unit drop rate Vs system load

There is no SU drop up to a load of 70 active users with proposed algorithm; however after that SU drop rate increases at a tolerable rate. The SUs drop rate for reference algorithm is zero when total number of users is 50 which is lower than the available number of PRBs (55). However with the increase in system load, SUs drop rate for reference algorithm increases significantly as shown.

Figure 5 shows throughput and fairness comparison of reference, proposed and PF algorithm. These simulations are done under the same system load of 110 users.

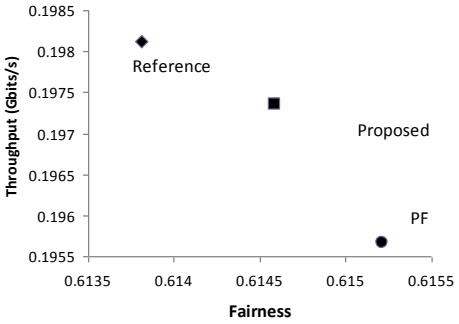


Figure 5. Trade-off between fairness and throughput

The system overall throughput for the proposed algorithm is lower than the reference algorithm by only 0.4%. This is because in the proposed algorithm, a delay-dependent queue-sorting algorithm is used and users with relatively low channel conditions but more waiting time are scheduled to guarantee QoS of RT traffic. This lowers the system overall throughput by a slight amount compared to the reference algorithm but more than PF algorithm.The fairness of proposed algorithm is improved as compared to the reference algorithm and is slightly less than PF algorithm as shown.In the three algorithms fairness of the PF algorithm is the highest with value 0.615213 as PF being an algorithm designed for user fairness and is taken as a reference for fairness analysis. The fairness index with the proposed algorithm is 0.61452 and with the reference algorithm fairness index is 0.61357.

In this way, the proposed algorithm sacrifices a little throughput (compared with reference algorithm) and fairness (compared with PF algorithm) but presents an better trade-off between throughput and fairness (compared with both reference and PF algorithms) as shown.

V. CONCLUSION

In this paper, we have presented a QoS aware PS framework that is composed of three main units for the resource allocation in DL transmission for OFDMA-based networks. These units use different queue sorting, TD adaptive scheduling and FD scheduling algorithms to guarantee better QoS to different traffic types. It is able to improve system spectral efficiency by optimizing the use of given radio resources and maintains a certain degree of fairness among users at the same time. This is achieved by adaptively providing just enough resources to RT traffic and distributing remaining resources efficiently to NRT services. The results show an improved QoS of RT traffic and a better trade-off between user fairness and system overall throughput.

REFERENCES

[1] Harri H. and Antti T.,"LTE for UMTS OFDMA and SC-FDMA Based Radio Access," John Wiley and sons ltd 2009, pp. 181-190.

[2] M. Sauter.(2008,April 23). Wireless Moves, 3GPP Moves on: LTE-Advaced.Website: http://mobilesociety.typepad.com/mobile_life/2008/04/3gpp-moves-on-html 29 .05.2010.

[3] Stefania S., Issam T., and Matthew B., "The UMTS Long Term Evolution Forum Theory to Practice," 2009 John Willey & Sons Ltd.ISBN: 978-0-470-69716-0.

[4] Jani P., Niko K., tero H., Martti M. and Mika R., "Mixed Traffic Packet Scheduling in UTRAN Long Term Evaluation Downlink," IEEE 2008, pp. 978-982.

[5] Gutierrez, F.Bader, and J.L.Pijoan, "Prioritization function for packet scheduling in OFDMA systems," Wireless internet conference 2008, Nov. 08, Maui, USA. http://dx.doi.org/ICST.WICON2008.5002.

[6] M.Andrews et al., "Providing quality of service over a shared wireless link," Communication magazine, IEEE, vol.39, 2001, pp. 150-154.

[7] P.Parag, S.Bhashyam, and R.Aravind, "A subcarrier allocation algorithm for OFDMA using buffer and channel state information," Vehicular Technology Conference. VTC-2005-Fall.2005 IEEE , pp. 622-625.

[8] Jun S., Na Yi, An Liu and Haige X., "Opportunistic scheduling for eterogeneous services in downlink OFDMA system," School of EECS, Peking University Beijing, P.R.China, IEEE computer Society 2009, pp. 260-264.

[9] G.Song et al, "Joint channel-aware and queue-aware data scheduling in multiple shared wireless channels," Communication Magazine,IEEE, vol.39,2001, pp. 150-154.

[10] B. Chisung, and C. Dong, "Fairness-Aware Adaptive Resource Allocation Scheme in Multihop OFDMA System," Communication letters IEEE, vol.11, pp. 134-136, Feb. 2007.

[11] Lin X., Laurie C. "Improving fairness in relay-based access networks," in ACM MSWIM, Nov. 2008, pp. 18-22.

[12] page 3GPP TSG-RAN, "TR25.814: Physical Layer Aspects for Evolved Utra," Version 7.0.0, June, 2006.

[13] IEEE 802.16j-06/013r3: "Multi-hope Rehay System Evaluation Methodology (Channel Model and Performance Metric)," IEEE 802.16 Broadband Wireless Access Working Group, 2007-02-19.

[14] Jiho J. and Kwang Bok L., "Tranmit power adaptation for multiuser OFDM systems," Selected Areas in Communication, IEEE Journal on vol.21, 2003, pp. 171-178.

[15] Suleiman Y.Yerima and Khalid Al-Begain, "Dynamic buffer management for multimedia QoS beyond 3G wireless networks," IAENG International Journal of computer science, 36:4, IJCS_36_4_14, Nov.2009.

[16] Won-Hyoung P., Sunghyun C., and Saewoong B., "Scheduling design for multiple traffic classes in OFDMA networks," IEEE 2006, pp. 790-795.

[17] T. Janevski, " *Traffic Analysis and Design of Wireless IP Networks"*, Artech House, Norwood, MA, 2003.

 ISBN: 978-1-61208-100-7

M2Learn Open Framework: Developing Mobile Collaborative and Social Applications

Sergio Martin, Gabriel Diaz, Elio Sancristobal, Rosario Gil, Manuel Castro, Juan Peire
Computer and Electrical Engineering Department
UNED – Spanish University for Distance Education
Madrid, Spain
{smartin, gdiaz, elio, rgil, mcastro,jpeire}@ieec.uned.es

Ivica Boticki
Learning Sciences Laboratory, National Institute of Education
Nanyang Technology University of Singapore
Singapore, Singapore
ivica.boticki@nie.edu.sg

***Abstract*— This paper presents M2Learn framework as an open platform, which facilitates the development of mobile learning and ubiquitous applications. The main features of this framework are: (a) transparent management of multiple location-based technologies including GPS and cell towers; (b) identification of objects through RFID; (c) support for motion sensors (e.g., G-Sensor); (d) interoperability with Moodle e-learning platform; and (e) support for widely accepted e-learning standards, including LOM for learning objects, and IMS-QTI for assessment. To demonstrate the benefits of the M2Learn system, a MobileTwitter microblogging application had been developed. The benefits of the application relying on the M2Learn's infrastructure are discussed and experience in designing the system presented.**

Keywords - mobile learning; collaborative learning; M2Learn framework; microblogging.

I. INTRODUCTION

Mobile computing is one of the fastest growing technology industry areas worldwide. Mobile device adoption rate in western countries is 90%, with the youngest generations as the leading users [1]. One of the reasons for this success is the improvement in the devices' technical features and their low prices. The new generations of mobile devices have wider and touch screens, built-in digital cameras, and support Wi-Fi and 3G web connectivity. Most devices are equipped with GPS (Global Positioning System) receivers, RFID (Radio Frequency IDentification), NFC (Near Field Communication) readers or smartcards. All these new embedded technologies make a platform for the new generation of applications to be used in all kinds of environments. These applications are context-aware and benefit from the internet connection anytime anywhere.

There is a significant body of research witnessing these devices can be used is education in order to enhance the learning and teaching processes. Coupled with the proliferation of mobile devices, this causes the increase of the demand for mobile learning application development. Due to heterogeneous nature of the devices, frameworks which ease the development tasks have emerged. These systems are created to decrease the inherent complexity of the technologies brought into the mobile learning development. In this paper we present such a framework named M2Learn and an exemplary application for microblogging, which can be used to support a diverse set of learning and teaching scenarios.

The paper is organized in three main parts: introduction (chapter II), which describes how social interaction can improve the learning experience; architecture description (chapters III and IV), where an open framework to simplify the development of mobile applications is described; and finally the results; where an application built over the proposed framework is shown.

II. ENHANCING LEARNING EXPERIENCES WITH SOCIAL LEARNING

The term „Web 2.0" was introduced in 2005., denoting a shift in perception of the World Wide Web, which becomes regarded as a lot more than just a passive source of information - it becomes a platform, a user-oriented environment where people interact and actively participate in content creation [2]. One of the most important and well known components of the Web 2.0 concept is the so called "social software." [3]. Examples of social software can be found in tools and applications such as wikis, applications designed for collaborative work allowing a number of users to edit online content; blogs; online diaries; podcasting, a new type of online media transferring using syndication feeds; sites like del.icio.us [4] and Flickr [5], which use a concept called "folksonomy," a style of a collaborative categorization of content using tags; content sharing tools such as YouTube, MySpace and RSS, allowing users to subscribe to a website's content and receive notification each time the page changes; e-portfolio applications etc.

The impact of the Web 2.0 concepts on e-learning is summarized in the term e-learning 2.0 [6]. Tools like wikis, blogs, podcasting, e-portfolios etc. are used both in formal education and informal learning. The rise of importance of student-centred learning can be noticed: the learner is no longer a passive consumer of information but an active and engaged participant in the learning process who creates his or her own content and constantly interacts with other users.

 ISBN: 978-1-61208-100-7

Control of the learning process has been placed „into the hands of the learner," communication and collaboration being the key components [6].

In the era of pervasive and ubiquitous computing, where learning is not restricted to one single place and has become integrated into our daily lives, the importance of mobile learning is rapidly rising due to the numerous possibilities mobile devices have to offer as means of supporting the learning process [7]. Due to their unique characteristics they provide new possibilities for interaction, collaboration amongst learners, informal learning and are, in the same time, user – centred and personal. Therefore, they naturally coexist and supplement the idea of e-learning 2.0.

III. OBJECTIVES: TOWARDS AN OPEN FRAMEWORK FOR THE DEVELOPMENT OF MOBILE LEARNING APPLICATIONS

Mobile devices present an attractive tool for learners since they integrate several well known functionalities: videogames; watching videos; communication via mobile devices; and collaborative technologies (e.g., blogs, wikis, mash-ups and social networks). All these, supplemented with an adequate learning activity design, transform learners from mere listeners towards active participants in the learning process engaged in interaction in the way are already used to. Therefore, learners feel more motivated and engaged in learning.

M2Learn project makes a step further in the state of the art design of mobile learning applications. It supports the development of innovative mobile learning applications that complement and enrich learning experiences encouraging learning "anywhere, anytime", improving the social interactions, providing personalized educative experience to each learner, and reaching to places where traditional or on-line learning cannot reach. M2Learn middleware is devoted to helping mobile learning best supporting education, complementing traditional and on-line learning instead of replacing them, thereby promoting blended approaches.

The main contributions of the M2Learn should be: it is an open framework which simplifies and facilitates the development of complex mobile learning applications both "lowering the floor" (i.e., makes it easier for designers to build applications) and "raising the ceiling" (i.e., increases the ability of designers to build more sophisticated applications). The M2Learn's framework should:

- Provide an interface towards some services in the existing e-learning platforms (currently Moodle). This feature allows external applications, such as mobile learning applications and games, to interact with the services. Currently, the supported services are focused on collaboration, communication, and assignments methods.
- Gather students' e-portfolio from an external e-learning platform regardless the application source (e.g., games, mobile applications, e-learning platforms, and etc).
- Enable transparent sensor data acquisition, supporting location GPS, cell towers, Wi-Fi; identification through RFID; and motion recognition thanks to accelerometers, and digital compasses.
- Be released as source code, under a GNU GPL v3 licence.

It is our main goal to create a framework with inherent reduced inner complexity of location techniques, easy accessible learning services, context-awareness and management of e-learning resources, facilitating the development of m-learning projects within educational institutions.

IV. THE M2LEARN FRAMEWORK

Successful adoption of mobile devices in education requires the development of applications which adequately support mobile learner. Thus, authors have committed to create the M2Learn framework, supporting the development of the new generation of mobile learning applications [8]. Being scalable and reusable, the framework supports plug-and-play configuration by the consistent use of standards and definitions of public interfaces. This allows for adding of new services to the system without changing the software structure (Figure 1).

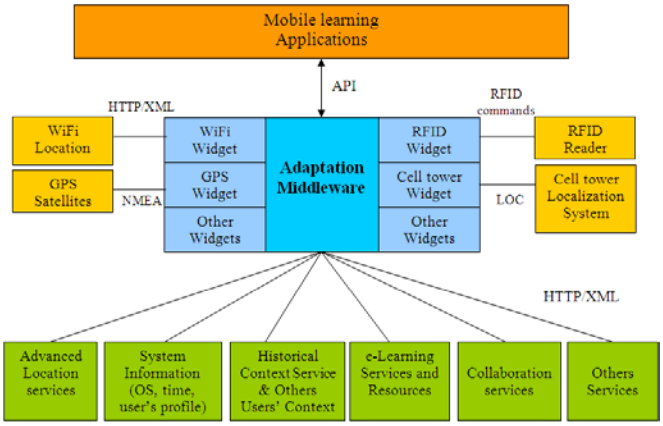


Figure 1. M2Learn architecture includes location technologies and integrates various services, some of them coming from e-learning platforms.

The framework supports user-driven collaboration and communication by integrating content into blogs, chats, and forums already supported by the existing e-learning platforms in organizations (at this stage only support for Moodle had been developed). M2Learn promotes the P3 social networking (i.e., People-People-Place) driving the participation in mobile social learning-oriented communities, mainly due to its interface towards e-learning platforms and location technologies [8].

A. A Context-aware approach

M2Learn provides easy access to sensors and multi-modal interfaces (e.g., accelerometers), having the potential of improving student engagement in content and activities. It enables invisible switching between various location-based technologies (e.g., GPS, cell towers-triangulation, WiFi) and supports the "Internet of Things" paradigm by integrating a RFID reader management module. All this sensor data is aggregated into users' context and complemented with the

access to services devoted to translating identifications (e.g., coming from RFIDs) or latitude and longitude coordinates (from GPS or cell towers) into an area name (e.g., room, building, street, city, and country) which can then be associated to services or contents. The contextual information can be also used to simplify the development of augmented reality applications since developers use the provided API (location and motion) to acquire the information to be superimposed to the camera images.

All the contextual information (e.g., location, time, profiles, schedule, surrounding people or preferences) can be used to personalize the access to content and services. The M2Learn framework includes a module for content and service discovery based on spatial and temporal variables (associating resources to time and places); yields personalized context-aware search results and integrates mainstream and social media through the use of data feeds (RSS).

B. *Towards integrated learning services and mashups*

The M2Learn framework supports fundamental e-learning standards, such as LOM for Learning Objects and IMS-QTI for assessment. It is able to communicate with the services offered by the existing e-learning platforms such as calendar, chat, forum, blog, assignments, and wiki. Logs from external applications can be used to put together students' e-portfolios and triangulate it with the educational experience coming from a mobile application, game, or an e-learning platform.

What is more, the framework allows the creation of mashups by offering its contextual information through an external publicly available interface. Figure 2 represents an example of such an environment where two mobile applications are built on top of the M2Learn Framework interact via the services offered by an e-learning platform. These applications compile contextual information from different sensors and send it to the Context Hub, which then distributes it to the rest of the group or other external applications to support the creation of mashups. A key element in this ecosystem is the Contextual Service Directory which provides information about the available services for each user according to the location and time.

This architecture simplifies the development of mobile learning mashup applications considerably. For example, users will be able to create a mash-up system using the predefined location information APIs instead of dealing with intricacies of the NMEA protocol and GPS protocols; or dealing with complex communication mechanisms through a serial port with an RFID controller in order to read information from an RFID tag and interpret it accordingly.

As an example of added value and the ease of use, users will be able to create a mobile blog using the services provided by an e-learning platform, but they will not need to create any web service in the language of the platform or understand how the database is structured. They will be able to reuse simple existing interface with information and services provided by the M2Learn middleware.

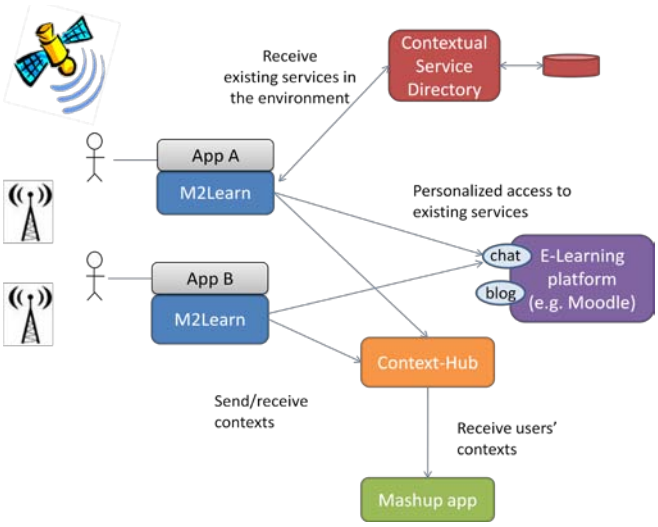


Figure 2. Overview of a mobile ecosystem built on top of the M2Learn Framework.

Although the M2Learn framework is already developed and deployed, mobile learning applications that rely on it are currently being designed. It has to be noted that every new mobile learning application cannot be conceived as an isolated entity, but as a live ecosystem of both services and users. With M2Learn this is indeed possible since it facilitates the coexistence of multiple users that interact, collaborate and communicate with each other minding group's context.

V. DEVELOPING A MICROBLOGGING APPLICATION USING THE M2LEARN FRAMEWORK

In order to evaluate the ease, effectiveness and the appropriateness of the M2Learn framework, 7 students taking a distance learning course in mobile programming were each assigned a task of developing an application using different kinds of mobile technologies (i.e., Web, .NET, Android). All these applications have one thing in common: they are based on the M2Learn framework and reused its APIs in order to build higher level application services.

One of the applications deals with microblogging and is inspired by ever-so-popular web service Twitter [9]. It is based on M2Learn's context aware modules and is connected to an e-learning platform. From an educational point of view, the application could be used for out-of-the-classroom activities in which students microblog their opinions or answers in different context and locations. Students collaborate *in situ* depending on the task assigned by the teachers and are latter debriefed in the classroom by the teachers who reviews their contributions through an e-learning platform

The MobileTwitter microblogging application created by the students encapsulates the following functionalities: (a) a user is able to post a message; (b) the posted message is

 ISBN: 978-1-61208-100-7

stored into the central server space; (c) other users are able to read the message by connecting to the system.

These mechanisms are similar to the well known web service Twitter. In order to put the educational benefits in the first place the following modification were introduced. Each time a user posts a message the application automatically associates a location tag with it. The location appears regardless of the position of users posting the message – they can be indoor or outdoor. As an example, a microblogged sentence ”I’m going to start the practice!” is transformed into “Sergio@Juan del Rosal 12, Madrid: I’m going to start the practice!” prior to the submission into the centralized online repository. In this case M2Learn framework is used to contextualize student generated artifacts and to store them into the Moodle platform ready for the teacher and students post-activity debriefing session. Throughout the process, all the activity is logged into the M2Learn platform step-by-step contributing to building students’ e-portfolios.

The MobileTwitter application is composed out of three main modules: configuration, message publishing (used by the students in order to send messages), and message review (reviewing others’ messages) (Figure 3).

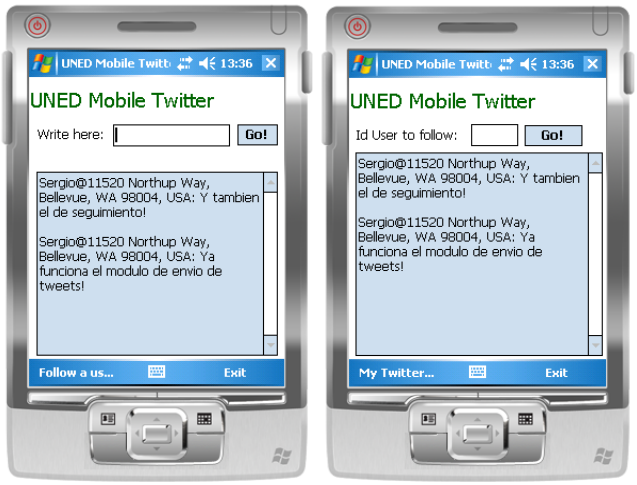


Figure 3. The MobileTwitter microblogging application developed by the students of Mobile Programming distance learning course

VI. CONCLUSION

The paper presented M2Learn framework which provides an infrastructure for the development of mobile learning applications. The infrastructure makes the use of context-aware information (such as location information) straightforward and the access to external e-learning repositories much easier due to the definition of externally accessible interfaces.

These benefits can be used by the developers of mobile learning system in order to build mobile learning applications quicker and to benefit from the centralized online repositories which not only store user artefacts, but which automatically create on-line user portfolios and aggregate contextual information to provide higher level application services. As an example we present a microblogging application built on top of the framework making use of the contextual information in enhancing mobile learning experiences.

The authors nevertheless acknowledge that the design of learning environments should lead the development of mobile learning applications. In that sense, M2Learn framework can contribute by structuring and simplifying the usually troublesome mobile learning activity design procedure.

VII. ACKNOWLEDGMENT

Authors would like to acknowledge to the Spanish Science and Innovation Ministry for the support in the project TIN2008-06083-C03/TSI “s-Labs – Integración de Servicios Abiertos para Laboratorios Remotos y Virtuales Distribuidos” and to the CYTED-508AC0341 “SOLITE-SOFTWARE LIBRE EN TELEFORMACIÓN” project support.

Also authors would like to acknowledge the support of the Project 142788-2008-BG-LEONARDO-LMP mPSS – mobile Performance Support for Vocational Education and Training Project and IPLECS Project – Internet-based Performance-centered Learning Environment for Curricula Support Project ERASMUS 141944-LLP-2008-1-ES-ERASMUS-ECDSP. Finally authors want to acknowledge the support provided by e-Madrid Project, S2009/TIC-1650, “Investigación y Desarrollo de tecnologías para el e-learning en la Comunidad de Madrid”.

REFERENCES

[1] Vidal, F. and Mota, R., “Encuesta de Infancia en España 2008”. Fundación SM, Universidad Pontificia Comillas-ICAI-ICADE y Movimiento Junior, pp. 1-16, 2008

[2] T. O'Reilly, “What Is Web 2.0 - Design Patterns and Business Models for the Next Generation of Software”, 2005, http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html

[3] B. Alexander, “Web 2.0: A New Wave of Innovation for Teaching and Learning?”, EDUCAUSE Review, vol. 41, no. 2 (March/April 2006): 32–44, 2006.

[4] Delicious. On-line resource, accessed on February 15, 2010. URL: http://delicious.com/

[5] Flickr. On-line resource, accessed on February 15, 2010. URL: http://www.flickr.com/

[6] S. Downes “E-learning 2.0”, eLearn, 2005, ACM Press, http://www.elearnmag.org/subpage.cfm?section=articles&article=29-1

[7] Jacobs, J., and Polson, D., “Mobile learning, social learning”, Proceedings of Learning On The Move OLT Conference, 26 September 2006, Queensland University of Technology, Brisbane.

[8] Martín, S., Diaz, G., Sancristobal, E., Gil, R., Castro, M., and Peire J., “Supporting M-learning: The location challenge”, Proceedings on the 2009 IADIS Mobile Learning Conference, January 2009, Barcelona.

[9] Twitter. On-line resource, accessed on February 15, 2010. URL: http://twitter.com/

 ISBN: 978-1-61208-100-7

Next Generation Network Architecture for Integration of Wireless Access Networks

Fazal Wahab Karam

Center for Quantifiable Quality of Service, Norwegian University of Science and Technology, Norway

fwkaram@q2s.ntnu.no

Terje Jensen

Center for Quantifiable Quality of Service, Norwegian University of Science and Technology, Norway

Telenor Group, Norway

terje.jensen1@telenor.com

***Abstract*—Motivation for the development of Next Generation Networks (NGN) concept is not only the success of mobile technology and growing popularity of IP-based multimedia services but also required cost savings, limited address space and fueling competition and collaboration. This paper proposes a solution for seamless interworking between network domains for the NGN concept. More specifically, seamless interworking is described for WLAN, WiMax and UMTS/LTE networks. The proposed architecture is an all-IP design. We introduce a heartbeat mechanism between WLAN, WiMax and UMTS/LTE using the IEEE 802.21 Media Independent Handover (MIH) Information Service, which enables low handover latency by reducing the target network detection time.**

Keywords- NGN; WLAN; WiMax; UMTS; LTE; QoS; IEEE 802.21 MIHF; interworking; domain coupling.

I. INTRODUCTION

Not only the huge success of mobile technology and growing popularity of IP-based multimedia services are motivating the development of Next Generation Networks (NGN), but also saving costs, open for more address space and lowering threshold for more actors. NGN is described as the convergence of public switched telephone network, the wireless networks and the data networks.

The combination of cellular networks and wireless networks meets the need for wide range and high data rate. In effect, this provides better service to users. To approach such a combination this paper proposes a next generation scheme for seamless interworking between WLAN [1], WiMax [2] and UMTS/LTE [3] networks hereby called WLAN-WiMax-UMTS/LTE Interworking Architecture. This paper is a continuation to [4] which proceeds with proposing a framework to provide service mobility following the NGN architecture and uses the IEEE 802.21 MIH services to exchange information by introducing a control plane with MIH information severs to guarantee connectivity while using best network selection (see Figure 1). This paper proposes how to integrate wireless access networks in the access plane of the proposed framework.

The paper is structured as follows. Section II summarizes the types of integration and pros and cons of those types of integration. Section III proposes the next generation WLAN-WiMax-UMTS/LTE Interworking Architecture. Section IV describes Heartbeat Messages and Section V discusses how QoS parameters are measured in the proposed scheme. Section VI gives analytical evaluation of the heartbeat mechanism. In the last section, the paper is concluded.

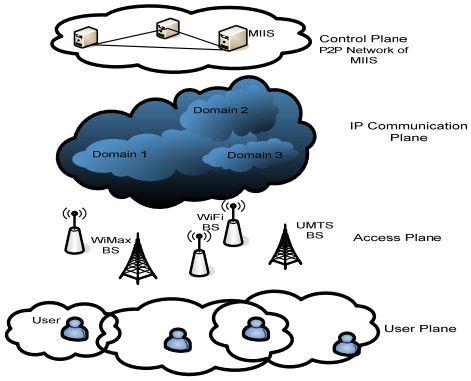


Figure 1: Proposed NGN Architecture [4]

II. TYPES OF ACCESS NETWORKS INTEGRATION

Two approaches for integration of wireless (e.g., WLAN) and cellular (e.g., UMTS) networks have been defined by ETSI – loose coupling and tight coupling. In loose coupling (Figure 2), wireless and cellular networks are not directly connected. They do not share a common protocol stack. Hence, this results in fairly long handover latency and potential packet loss.

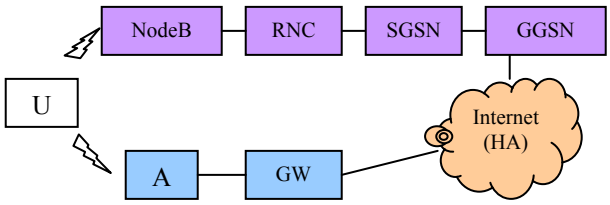


Figure 2: Loose Coupling WLAN-UMTS Interworking

One advantage is that it allows independent deployment and traffic engineering of WLAN and UMTS. Roaming agreements with partners can allow widespread service enabling subscribers to use a single service provider for all network access. Another advantage is that it allows a WISP (wireless internet service provider) to

 ISBN: 978-1-61208-100-7

provide its own public WLAN hotspot, interoperate through roaming agreements with public WLAN and UMTS service providers or manage a privately installed enterprise WLAN. Loose coupling is commonly implemented by use of Mobile IP.

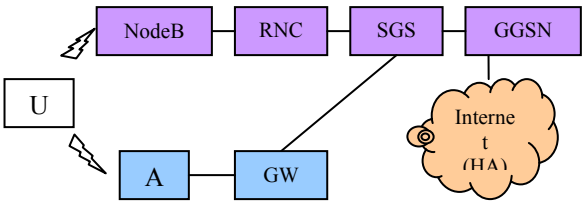


Figure 3: Tight Coupling WLAN-UMTS Interworking

In tight coupling, a wireless network is connected to cellular network just like any other radio access network (see Figure 3). Wireless network emulates the functions similar to GPRS functions [7]. Advantages of tight coupling include seamless handover across WLAN and UMTS, reuse of AAA, reuse of infrastructures, increased security, common provisioning and customer care and access to core UMTS services like SMS, MMS and location based services. One disadvantage is that tight coupling needs to be tailored for WLANs owned by cellular operators and does not easily support third party WLANs. The same operator must manage both WLAN and UMTS parts since the core network interfaces are exposed. Another disadvantage is that tight coupling does not straight forwardly support legacy WLAN terminals that do not implement the UMTS protocols. Cost and capacity of the SGSN associated with the connection of WLAN may also be a disadvantage. While throughput capacity of traditional SGSNs is sufficient to support thousands of low-bit-rate GPRS terminals, it may not be sufficient to support hundreds of high-bit-rate WLAN terminals. Thus SGSN could become a bottleneck for high data rate applications.

The integration architectures, as explained in [5], differ for each of the interworking options, as shown in Figure 4.

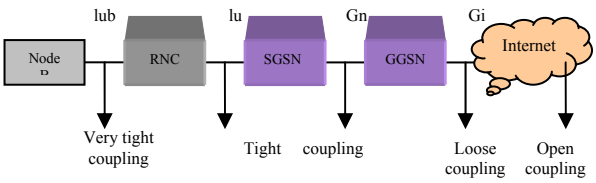


Figure 4: Interworking points with the relevant architecture names (based on [5])

WLAN can be integrated with UMTS at PS core. Depending on which point it is attached to – it is loose, tight, very tight or open architecture.

Note that in this section, WLAN and UMTS are used as examples of wireless and mobile systems. The description can be made more general including systems such as WiMax, LTE and others.

III. NEXT GENERATION PROPOSED ARCHITECTURE

The design of an architecture that efficiently integrates WLAN, WiMax and UMTS/LTE is a rewarding task. In the following a scheme for integration of cellular and wireless networks is described. The case of WLAN, WiMax and UMTS is chosen for illustration as shown in Figure 5.

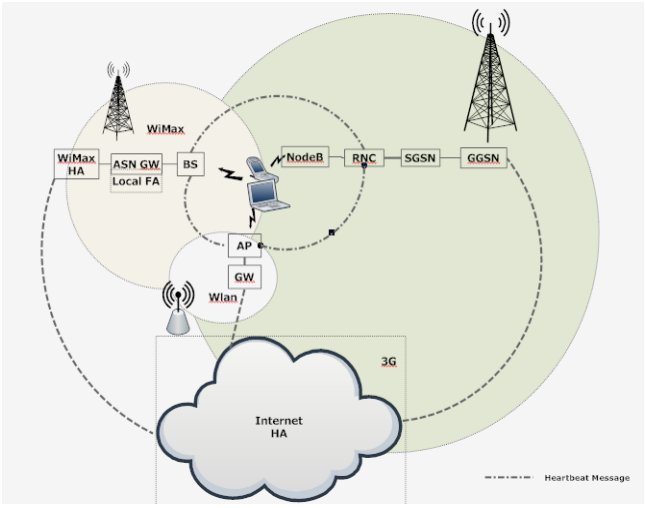


Figure 5: Proposed Integration of Access Networks

A. Architecture

We assume that neighboring WLAN Access Point (AP) and UMTS Base Station (BS)/ Radio Network Controller (RNC) and WiMax BS are connected via a backbone network. The APs and nodes are capable of verifying the identity of neighbor APs and nodes. The User Equipment (UE) has multi radio interfaces and protocol stacks for WLAN, WiMax and UMTS. Note that different classes of UEs can be present, some able to make use of multiple radios while others are not.

The significant characteristic of the proposed architecture is that the UE needs not have all the interfaces on all the time. The list of all available networks can be retrieved from any one of the interfaces – WLAN, WiMax or UMTS using the IEEE 802.21 MIH Information Service Request.

AP, BS and RNC are connected, i.e., they can identify each other and can share information with each other, either directly or via a mediator. They share information with each other by way of periodic heartbeat messages using IEEE 802.21 Media Independent Handover (MIH) Information Service Messages. A UE initiating a session can get relevant network and QoS information depending on type of service. The network selection can then be either made by UE or by network node (including AP/ BS/ RNC) to provide seamless handover and service continuity.

 ISBN: 978-1-61208-100-7

B. WLAN-UMTS interworking

One motivation for WLAN-UMTS interworking is to extend UMTS services and functionality to WLAN access environment [11]. Additional capacity and higher data rates for end users on WLAN and operation of WLAN on unlicensed frequency band makes it all the more suitable for WLAN-UMTS interworking. Figure 6 summarizes the UMTS-WLAN interworking characteristics [12].

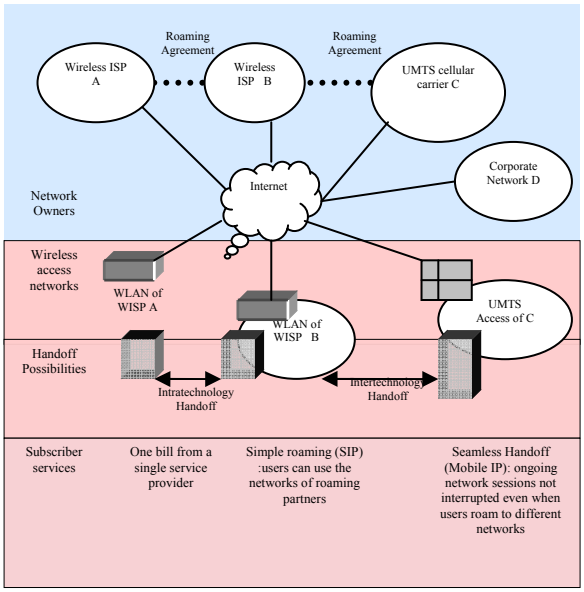


Figure 6: WLAN–UMTS Interworking (adapted from [12])

We propose a scheme where AP and RNC (Radio Network Controller) forward packets to each other intelligently. WLAN and UMTS are indirectly connected through an IP-based network which means a loosely coupling architecture. Mobile IP can be used to roam between WLAN and UMTS networks. An enhancement of the existing architecture is the introduction of communication mechanism between APs and RNCs/network nodes using IEEE 802.21 MIH with heartbeat messages as explained in Section IV. One option is to allow APs and network nodes to obtain information via multi-radio interfaces. Another option is to allow UEs reporting on the conditions observed. In the former case, AP and network nodes periodically broadcast heartbeat messages to all neighboring nodes including other radio interfaces using IEEE 802.21 MIH messages.

On receiving the heartbeat messages, AP and network nodes update the network map information. This way, all relevant nodes know the existence of other access networks including the relevant network attributes. This information is later used for making optimum network selection before handover.

C. WLAN-WiMax interworking

WLAN offers high data rates within a 100 m range whereas WiMax offers lower data rates in an 8 km range [6]. Instead of selecting one network to provide access to network services, interworking both networks can use each network's advantages. Service providers can provide bundled services to users in either access network thereby using both licensed and license-exempt frequency bands. In this way, service providers can sell attractive devices supporting WiMax and WLAN capabilities taking advantage of device cost savings [8].

In the proposed architecture, WLAN and WiMax are integrated at the IP layer following the NGN concept [7]. In the proposed integration, a WiMax BS or WLAN AP periodically broadcasts heartbeat messages giving neighbor network information to relevant nodes. Any AP or BS in the vicinity will receive these messages. It is assumed that AP and BS will stay on and listen on those interfaces. In effect, AP and BS neighbors can assist when selecting the best available network. This may reduce target network detection time. The heartbeat messages follow the format of IEEE 802.21 MIH messages. The heartbeat message format is given in Chapter IV.

The difference between this proposal and those given in [8][10] is that the target network detection time is reduced. The IEEE 802.21 MIH Information Service heartbeat messages minimize the target network detection time. Since a current network admission controller (e.g., residing in AP/BS) is aware of the target network (post handoff), the handover latency is reduced.

D. WiMax - UMTS interworking

WiMax – UMTS may be partially overlapping as opposed to fully overlapping commonly seen for UMTS – WLAN (see Figure 7), [14]. In the latter case, the UE may maintain the Packet Data Protocol (PDP) context of UMTS while simultaneously being connected to WLAN. Hence, when the UE leaves the WLAN spot, it can reconnect immediately to UMTS without reactivating PDP context. However, since WiMax coverage may be partially overlapping UMTS, the handover needs to be fast enough to maintain service continuity.

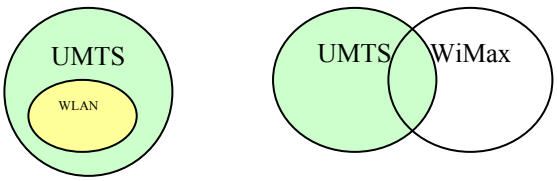


Figure 7: WLAN-UMTS Vs WiMax-UMTS

To achieve this, Mobile IP can be used as common interconnection protocol. The WiMax BS and UMTS node must then interwork. The WiMax Access Network (ASN) provides the WiMax access services for the UE. WiMax Home Agent (HA) manages the mobility inside WiMax network. The WiMax HA is not included in UMTS core network to keep its independence (loose/open coupling). The Foreign Agents (FAs) located in ASN Gateway are considered as the local FAs in the interworking architecture. A common AAA network could be utilized. GGSN manages

 ISBN: 978-1-61208-100-7

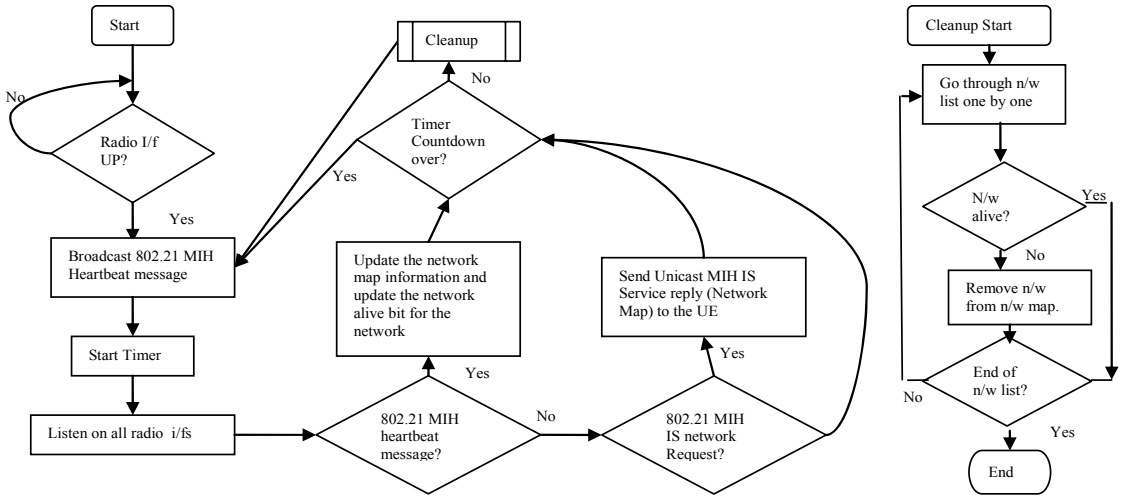


Figure 8: Flowchart for Heartbeat Mechanism at AP, BS, RNC

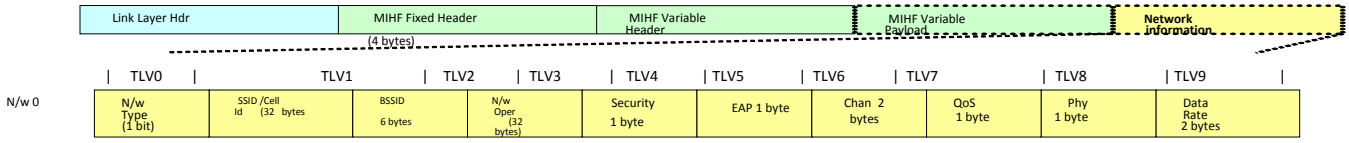


Figure 9: Heartbeat (Network Information) Message

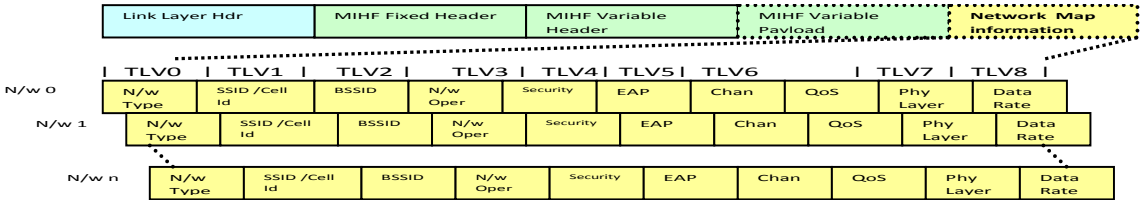


Figure 10: Heartbeat Message Information Service Reply (Network Map) Format

the mobility and FA functions in UMTS network. The major advantage of this proposal over the proposal in [14] is the seamless handover and low packet loss. It is achieved through IEEE 802.21 MIH based heartbeat messages. This method guarantees both the independence of between networks, and low handover latency.

IV. HEARTBEAT MESSAGES

IEEE 802.21 MIH messages make the heart beat message uniform across different access networks. The timing of these messages will also be independent. The message format is as shown in Figure 9. This is in TLV format as required by IEEE 802.21 Variable Payload. Network attributes are periodically distributed to every relevant node. Each of the nodes listens on for heartbeat messages from other nodes. They form a network map based on the information received through heartbeat messages. On receiving the heartbeat message, they update their network map information periodically. UE can either retrieve the network list from any of them or can issue an MIH Information Service Request for Network Map Information on all its radio interfaces. As shown in Figure 8, the node multicasts its network information using IEEE 802.21 MIH messages on relevant interfaces. Being a periodic message, the AP will send the heartbeat message every *n* time units. So a timer is used to count for this purpose. Meanwhile, AP also listens on relevant interfaces for heartbeats from other nodes. On receiving an IEEE 802.21 MIH packet on any interface, AP checks what type of packet it is. If it is a heartbeat message, AP will update the network map information and marks the entry alive. If it is a network request message from a UE, AP sends the network map information to the UE. If the timer expires it goes through the network list and checks whether the network is still active. If the network is not active, that entry is deleted from the network map. This way all the entries in the network map

 ISBN: 978-1-61208-100-7

are checked. Responses are shown in Figure 9 and Figure 10 which is in TLV format as required by IEEE 802.21 Variable Payload. This communication mechanism enables faster handovers by referring to the available network map information.

V. QoS Network Performance Parameters

ITU recommendation for the service classes and QoS parameters mapping for various access technologies are given in [4]. QoS parameters are collected by every network. These include parameters like delay, throughput, packet error ratio, average packet transfer delay and jitter. For example, transfer delay is measured by measuring the time required by a packet to travel from ingress to egress of a node (e.g., AP/BS). The IN timestamp and OUT timestamps are stored and their difference gives the packet delay.Average packet delay is calculated as accumulated packet delay divided by number of packets transmitted successfully.

Similarly, packet error ratio is given by the number of failed transmissions over total transmissions. Throughput is the maximum sustained traffic rate for WiMax, Maximum Bit rate and Peak rate for UMTS/LTE and Peak Data Rate for WLAN. This is kept track of by relevant network nodes. Using the standard formulas for M/M/k/m queue model, [4], delay and throughput parameters can be calculated by the individual network nodes. Similarly, call blocking probability can be calculated using analytical model proposed in [15] in terms of number of virtual channels (*N*), user arrival rate (λ), arrival rate of type1 call (λ1), arrival rate of type2 call (λ2) arrival rate of type3 call (λ3) and service time of the user (μ)].

Handoff Call dropping probability may be calculated as given in [16] [17].

$$P_d(n,g) = \frac{\frac{A^{n-g}}{n!} A_1^g}{\sum_{i=0}^{n-g-1} \frac{A^i}{i!} + \sum_{i=n-g}^{n} \frac{A^{n-g}}{i!} A_1^{i-(n-g)}}$$

Note: if we set $g = 0$, the above expressions reduces to the classical Erlang-B loss formula [17]. Where Pd - Call dropping probability, g – number of channels reserved for handoff calls, n – Number of idle channels and A – Call arrival rate / (call completion rate + handoff departure rate). The QoS network performance measurement parameters (Delay, Throughput, Call Blocking probability and Handoff Dropping probability) are piggybacked to heartbeat messages as shown in Figure 11.

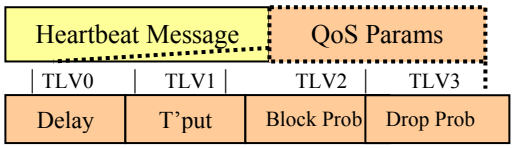


Figure 11: QoS Params

The QoS Network Performance Measurement parameters can then be used to rank the access networks and finally select the best available network. A network ranking algorithm will require the QoS parameters to make optimum decision suiting the network requirements of UE. For example, if UE is connected to WiMax and is moving another network may become available, such as UMTS and WLAN. Say, if call dropping probability of UMTS is higher than for WLAN, and the ranking algorithm should place higher desirability to WLAN and rank it better than UMTS. These parameters are the key factors that decide the desirability of a network when making a network selection decision.

VI. Analytical Evaluation

In this section, we analyze the costs involved to find whether the proposed system is better than existing loose coupled architecture. Consider a UE is connected to one of the access networks – WLAN, WiMax or UMTS. Say there are x networks available to UE. The UE sends x requests to find available networks to perform a handover and receives x responses about available networks (if all are still available).

So, for detecting the target networks in conventional loose coupling architecture, the UE will have to exchange at least $2x$ packets to find characteristics about x access networks. This is best case scenario. The ACK packets and retransmissions due to error are not even considered here.

Handoff Latency (LH) can be expressed in terms of detection delay (LD), handoff request delay (LHR) and handoff response (LHS) delay.

$$LH = LD + LHR + LHS$$

The proposed architecture aims at minimizing the detection delay thereby reducing the handoff latency. [18] Shows, by simulation, that detection delay is a significant component of Handoff latency. The data rates differ, however, for WLAN, WiMax and UMTS. Together with the different network solutions implies that responses will arrive at the UE at different times and reflect different time intervals. Some examples of data rates are given in Table I.

TABLE I. Data rates for WLAN, WiMax & UMTS

WLAN	***Mobile WiMax***	***UMTS***
54 Mbps	40 Mbps	2Mbps

In the proposed architecture, UE sends MIH Network Request Message and receives the network Map information through the IEEE 802.21 MIH Network Map Message. So, in all 2 packets are exchanged to get the list of available networks. Let Dwlan –Time to transfer a packet on WLAN, Dwimax – Time to transfer a packet on WiMax and Dumts– Time to transfer a packet on UMTS. In Loose coupling architectures given in [8][10], [14]; Time to detect x networks when UE is connected to (serial request - response) WLAN= 2 * x * Dwlan. Time to detect x networks when UE is connected to WiMax = 2 * x * Dwimax. Time to detect x networks when UE is connected to UMTS= 2 * x * Dumts.

 ISBN: 978-1-61208-100-7

For the architecture described in this paper, time to detect x networks when UE is connected to WLAN = 2 * Dwlan, time to detect x networks when UE is connected to WiMax = 2 * Dwimax and time to detect x networks when UE is connected to UMTS= 2 * Dumts. So, to detect available networks on each of the access networks, the cost involved in terms of time units is summarized in Table II.

TABLE II. CALCULATIONS SHOWING DETECTION DELAY LD

	Number of Access networks	***Time to detect on WLAN***	***Time to detect on WiMax***	***Time to detect on UMTS***
Loose coupling Architecture	x	$2 * x * D_{wlan}$	$2*x* D_{wimax}$	$2 *x * D_{umts}$
Proposed architecture	x	$2* D_{wlan}$	$2*D_{wimax}$	$2*D_{umts}$

As shown in Table II, the detection delay LD is only a multiple of the packet transfer delay in each of the access networks. However, for other architectures, LD is a multiple of both the number of access networks (serial detection) as well as the packet transfer delay. With increase in the number of networks available, the detection delay will also increase. The numbers of messages are also increasing correspondingly. For serial detection, the delay will be equal to the one proposed in this paper only when $x = 1$. That is, only when one network is available.

VII. CONCLUSION

In this paper, we have proposed a next generation WLAN-WiMax-UMTS interworking architecture. This is based on NGN architecture (which follows 3GPP standards) and proposes a novel heartbeat mechanism using IEEE 802.21 MIH Information Service. The architecture promises a low target network detection time during the switching of the communication. The mobility between two access networks is achieved by the Mobile IP at the network layer.

The QoS Network Performance Measurement Parameters are piggybacked to heartbeat messages and shared with other access networks. We have shown how proposed architecture can minimize detection delay, and ultimately reduce the handover latency. Our future work will focus on handover algorithms and ranking algorithm for access networks based on the proposed architecture and their performance evaluation through simulation.

REFERENCES

[1] R. O'Hara and T. L. Cole, ''Local and metropolitan area networks-Specific requirements,'' IEEE Std 802.11™-2007: http://standards.ieee.org/getieee802/download/802.11-2007.pdf. [Accessed: May 10, 2010]

[2] R. B. Marks and J. Puthenkulam, ''Local and metropolitan area networks—Part 16: Air Interface for Broadband Wireless Access Systems,'' IEEE Computer Society, IEEE Std 802.16™-2009: http://standards.ieee.org/getieee802/download/802.16-2009.pdf . [Accessed: May 05, 2010]

[3] The UMTS Forum, ''3G/UMTS Evolution: towards a new generation of broadband mobile services,'' December 2006: http://www.umtsforum.org/component/option,com_docman/task,cat_view/gid,327/Itemid,214/. [Accessed: May 04, 2010]

[4] F. W. Karam and T. Jensen, "On Schemes for Supporting QoS and Mobility in Heterogeneous Networks," FIT09, December 2009: http://q2s.ntnu.no/publication?publsearch=Karam,%20Fazal%20Wahab&publsearchoption=authors&publsearchsubmit=Search. [Accessed: April 01, 2010]

[5] P. Dini, J. Mangues-Bafalluy and M. Cardenete-Suriol, ''On the Interworking among Heterogeneous Wireless Networks for Seamless User Mobility,'' IEEE Transactions on Magnetics, 2007: http://www.cttc.es/resources/doc/071023-heterinterw-camera-ready-33406.pdf. [Accessed: April 14, 2010]

[6] H. Haffajee and H. A. Chan, "Low-cost QoS-enabled Wireless Network with Interworked WLAN and WiMAX", IEEE AusWireless'06, Australia, March 2006: http://utsescholarship.lib.uts.edu.au/iresearch/scholarly-works/handle/2100/177. [Accessed: April 05, 2010]

[7] K. Sutherland, ''Next Generation Networks (NGN), ACMA/ITU International Training Program,'' October 23, 2007: http://165.191.2.22/webwr/_assets/main/lib310475/next_generation_networks.pdf. [Accessed : March 15, 2010]

[8] Motorola and Intel, ''WiMax and Wifi Together: Deployment Models and User Scenarios.'' White Paper, 2007: http://whitepapers.zdnet.com/abstract.aspx?docid=350149. [Accessed: March 20, 2010]

[9] J. Guo, R. Yim, T. Tsuboi, J. Zhang and P. Orlik, ''Fast Handover BetweenWiMAX and WiFi Networks in Vehicular Environment,'' ITS World Congress 2009: http://www.merl.com/reports/docs/TR2009-063.pdf. [Accessed: May 05, 2010]

[10] T. Yahiya, H. Chaouchi, A. Kassler and G. Pujole," Seamless Interworking of WLAN and WMAN Wireless Networks" MSPE'06, RWTH Aachen University, Germany, Nov 2006: http://www.cs.kau.se/~andreask/papers/2006/MSPE2006.pdf. [Accessed: March 15, 2010]

[11] 3rd Generation Partnership Project, ''Technical Specification, 3GPP system to Wire-less Local Area Network (WLAN) interworking; System description (Release 6),'' 3GPP TS 23.234, v6.0.0, March, 2004.

[12] Buddhikot, S. Han, Y. W. Lee, S. Miller and L. Salgarelli, ''Design and implementation of a WLAN/cdma2000 interworking architecture,'' Communications Magazine, IEEE, Volume: 41, 2003, pp. 90 – 100.

[13] J. Y. Song, S. W. Lee and D. Cho, ''Hybrid Coupling Scheme for UMTS and Wireless LAN Interworking,'' vol. 4, IEEE (VTC' 03), 2003, pp. 2247-2251.

[14] Q. Thinh, N. Vuong, L. Fiat and N. Agoulmine, ''An Architecture for UMTS-WiMax Interworking,'' IEEE Broadband Convergence Networks (BcN' 06), 2006, pp. 1-10.

[15] R. Babu and P. S. Satyanarayana, ''Call Admission Control performance model for Beyond 3G Wireless Networks,'' International Journal of Computer Science and Information Security,Vol. 6, No. 3, 2009, pp. 224-229.

[16] S. Xie, ''Vertical Handoff Decision Algorithm Based on Optimal Grade of Service,'' IETE Journal of Research, March 2010, pp. 44-51.

[17] G. Haring, R. Marie, R. Puigjaner and K. S. Trivedi, ''Loss formulae and their optimization for cellular networks,'' IEEE Trans. on Vehicular Technology, ,May 2001, pp. 664-673.

[18] M. B. R. Murthy and F. A. Phiri, ''Performance analysis of downward handoff latency in a WLAN/GPRS interworking system,'' Journal of Computer Science, Jan, 2005, pp. 24-27.

 ISBN: 978-1-61208-100-7

A Mobile Internet Service Consistency Framework

Yangyi Wen, Chunhong Zhang
Key Laboratory of Universal Wireless Communications, Ministry of Education,
Mobile Life and New Media Laboratory,
Beijing University of Posts and Telecommunications,
Beijing, China
wenyangyi74@gmail.com, zhangch@bupt.edu.cn

Yabo Du, Yang Ji
Key Laboratory of Universal Wireless Communications, Ministry of Education,
Mobile Life and New Media Laboratory,
Beijing University of Posts and Telecommunications,
Beijing, China
blessdyb@gmail.com, jiyang@bupt.edu.cn

***Abstract*—As more and more Internet services migrate to mobile terminal, it is very necessary to carry out the service consistency between traditional Internet and mobile internet when service convergence. Unfortunately, current frameworks can not satisfy the requirements of consistency of service, for the difficulties in realize the coherence and continuity of user experience. In this paper, we bring forward a service consistency framework to achieve the inter-service consistency by the introduction of service consistency platform and widget manager. This service consistency framework mainly adopts W/S framework on mobile terminal, while maintains B/S framework on traditional terminal to retain the existing advantage. This paper is concerned to utilize the framework to make user experience the "anytime, anywhere, anyway" Internet service.**

Keywords-Internet service consistency; service consistency framework; inter-terminal

I. INTRODUCTION

Recently, the traditional Internet corporations speed up the process of service migration to mobile field. As the same time, terminal equipment manufacturers and operators also join the mobile Internet industry [1]. These participants gradually become the core forces in this huge Internet industry.

These leaders capture the market rapidly by putting forward various mobile terminal solutions, integrating the resource of data, optimizing the user experience and developing mobile Internet service actively, adding to the advantages of brand, technique and capital [2].

However, all the enterprise will face an extremely difficult problem: how to migrate existing services to the mobile terminal or how to extend the mobile Internet service inheriting the advantages of traditional Internet. Some provided WAP or WEB sites that can be visited with mobile browser to end-user. Some launched online store to sell applications. And some are still in exploration.

Every attempt converges on the realization of inter-terminal Internet service. A method or platform may make Internet service and mobile technology better fusion, optimize cooperation mechanism and realize the effective utilization of resources.

Current frameworks are unable to meet the requirement of coherence and continuity of experience because of the limit of platform or technique or others. We put forward our framework, not only settling down the difficulty but also maintaining existing advantage.

This paper is structured as follows, besides this introduction and the future work: Section 2 briefly discusses related work, which mainly proposes the concept of Internet service consistency and compares the realization used by three usual frameworks. Section 3 introduces the service consistency framework, including the design, the structure and the work flow. Section 4 presents a practical case based on the proposed framework.

II. RELATED WORK

A. Internet service consistency

Internet service should have three important elements, service data, service process and service display. Service data is the most interested thing to users. Service process is the work flow of service. Service display is the interface of service. All these are perceived by user as user experience.

It is inevitable to realize Internet services across different terminal. The display of traditional Internet service recurs to the computer terminal, called desktop Internet. The development of Internet is not limited to one terminal only. In fact, terminals that support access to Internet are rapidly increasing, including mobile terminals [3].

In this instance, inter-terminal Internet services that refer to one Internet service realized and appeared in different terminal.

It is necessary to realize service consistency of inter-terminal Internet services. End-users expect service can be consistent using different terminals. It means users can complete some parts of service in one terminal, and then continue completing rest parts in another. And at the same, there should be nearly no difference between terminals in the display of service [4].

Internet service consistency can be defined as inter-terminal Internet service maintaining the identity and

 ISBN: 978-1-61208-100-7

continuity of user experience. So we can conclude the characteristics of Internet service consistency:

- The terminal adaptation of service. The same service should be displayed and used cross different terminals. It is the fundamental of service consistency.
- The identity of service. It is composed of two elements, identity of service data, service process and service display. The identity of service data ensures that the inter-service is the same. The identity of service process ensures the flow of service is the same. And the identity of service display ensures the interface of service is the same.
- The continuity of service. It focuses on the continuity of service data, service process. The continuity of service data means that the data end-users is using or submitting on one terminal can migrate smoothly to another terminal. For example, we have seen a film for three minutes in computer, and then we can continue seeing the rest part of film in mobile phone automatically. The continuity of service process means that the process end-users are going alone can complete inter-terminal. Taking shopping for example, it is allowed users order a commodity in mobile phone, and then pay in computer.

Unfortunately, the increasingly emerging inter-terminal Internet service usually can not achieve the service consistency.

B. *Current frameworks for traditional mobile Internet service*

During the migration of Internet service from desktop to mobile terminal, the problem of service consistency also exits. According to the great advantage of traditional desktop Internet, we are inclined to discuss and modify the accessing framework in mobile terminal. Mobile Internet services should adapt all mobile terminals and maintain the identity and continuity of user experience with traditional service. Three frameworks are usually chosen to realize Internet service.

B/S (Browse/Server) framework refers to use browse to experience Internet service deployed in the remote server [5]. It is a classical framework accessing Internet service. Most mobile Internet service also prefers B/S framework no matter realizing with WAP or WEB technique.

C/S (Client/Server) framework is an accessing Internet service way using particular client [6]. Having a long magnificent history, C/S framework is also been applied widely on mobile terminal.

W/S (Widget/Server) is not a new framework that visits remote server using widget applications [7]. As a matter of fact, Widget applications appeared in computer desktop originally.The widget has been widely applied in mobile because of its small sizes.

However, as TABLE 1 showed, these frameworks can not achieve service consistency, especially in the continuity of user experience.

In terminal adaptation of service, B/S framework is very outstanding in respect that almost every mobile terminal has browser to visit Internet service. But C/S framework and W/S framework have to face this difficult problem. Services adopted C/S framework need to develop many clients to adapt different terminal, while services using W/S framework demand a widget engine installed in terminals [8].

In identity of service, services of B/S framework can be considered as a simpler version of traditional desktop Internet service. C/S framework usually made some modification of service process in view of the small screen and difficulty operation of mobile terminal. But the interface is friendlier because client technique has considered the display capacity of terminal. W/S framework is mostly for small services which are always parts of traditional services. Yet the interface of widget applications is closer to traditional one because widget technique is similar to web [9].

In continuity of service, no framework succeeds to realize because no method has been introduced to achieve communication between relatively independent terminals.

TABLE I. FRAMEWORK COMPARISON

Service consistency		B/S framework	C/S framework	W/S framework
Adaptation		very satisfied	Barely satisfied	Satisfied
Identity	Data	satisfied	satisfied	satisfied
	Process	More direct	Made some modification	Part of whole process
	Display	simpler	More friendly	closer to traditional service
Continuity	Data	Not satisfied	Not satisfied	Not satisfied
	Process	Not satisfied	Not satisfied	Not satisfied

From the above analysis, it is just the time to put forward a new framework to realize Internet service integration on mobile terminal.

III. SERVICE CONSISTENCY FRAMEWORK

Service consistency framework should resolve the problems caused by inter-terminal and maintain the same user experience. At the same time, Service consistency framework should not change the existing traditional Internet service.

A. *The design of service consistency*

Service consistency framework adopts B/S framework to keep the advantage in traditional terminal, and choose widget to realize the service in mobile terminal. In order to achieve Service consistency, some modifications are inevitable. A service consistency platform in charge of saving user service data and pushing the service to terminal and widget manager with responsibility for communication with platform and management of widget are introduced.

 ISBN: 978-1-61208-100-7

The service will break up to lots of parts. A corresponding widget is responsible for each part in term of contract. To provide inter-terminal service, a mapping table between service process and widget application should be offered to register in service consistency platform by service provider. A database or XML file (format mentioned as Part B) can be used to describe the mapping table.

Consistency framework consists of five parts: Internet service platform, service consistency platform, desktop Internet service display, mobile Internet service display, and access network. The components of framework are showed as Figure 1:

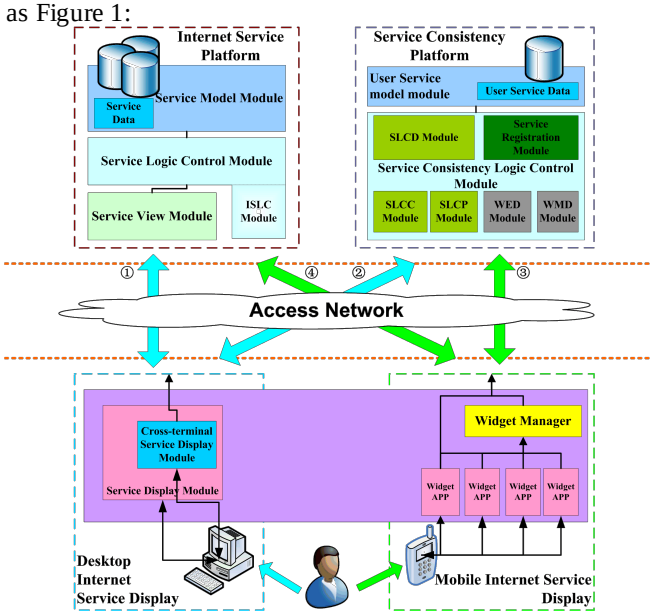


Figure 1. The components of service consistency framework

1) Internet service platform

Internet service platform can be divided into three modules: service model module, service logic control module and service view module.

Service model module is primarily to complete the encapsulation of service data.

Including inter-service logic control module (ISLC module) disposing inter-service, service logic control module based on service process deals with service data. After running successfully, widget application will access service logic control module directly to receive service data so that the identity of service data is realized.

Service view module takes charge of desktop Internet service display, while widget will display the mobile Internet service

2) Service consistency platform

Service consistency platform includes user service model module and service consistency logic control module.

User service model module packs user service data. User service data involved user information and service information what user just done and will do. Once having user service data, to realize the continuity of service is easier.

Service consistency logic control module comprises six sub-modules.

Service registration module handles the registration of inter-service. Only after registration, widget is available to download.

Service logic control of service customization module (SLCC module) realizes customization of service. User service data is gathered by this module.

Service logic control of user service data module (SLCD module) takes charge to save, and read, and modify, and delete user service data.

Service logic control of service pushes module (SLCP module) will push the related information to mobile terminal at appointed time. By this module, user service data reaches mobile terminal successfully.

Above modules accomplish the migration of user service data from desktop Internet service to mobile service. So the realization of continuity of service data achieves.

Widget engine download module (WED module) provides online download of engine. The widget engine is developed as software which can be downloaded. The problem of terminal adaptation of service is settled down.

Widget manager download module (WMD module) offers download of manager. Widget download module supply download of widget application. Widget manager can mange widget applications to complete the whole service process so as to the achievement of identity of service process. And widget manager makes the continuity of service process feasible.

3) Desktop Internet service display

Inter-service display module is added to achieve the display of service customization, while other modules of this part conform to existing framework.

4) Mobile Internet service display

Widget manager was introduced to communicate with service consistency platform and manager widget application.

Service is showed by widget application in mobile terminal. This ensures the identity of service display.

It is worth noting that widget manager starts widget application and transfer user service data to application. After that, application can directly communicate with Internet service platform.

5) Access network

Access network sees to data transmission between platform and terminal.

In conclusion, service consistency framework has the capacity of realizing Internet service consistency.

B. The work flow of integration framework

In Figure 1, the label 1 means that end-user firstly browses Internet service in traditional terminal. The communication in label 2 happens when user decides to customize service. Then, service consistency platform will interact with mobile terminal, just as label 3 noted. Finally,

 ISBN: 978-1-61208-100-7

mobile widget accesses Internet service platform to gain service data, showed as label 4.

We can notice that the data transmission of label 1 and 4 are just the same as B/S framework and W/S framework. So the most important flow is how terminals communicate with service consistency platform. And this flow can be divided into three parts and described as follow:

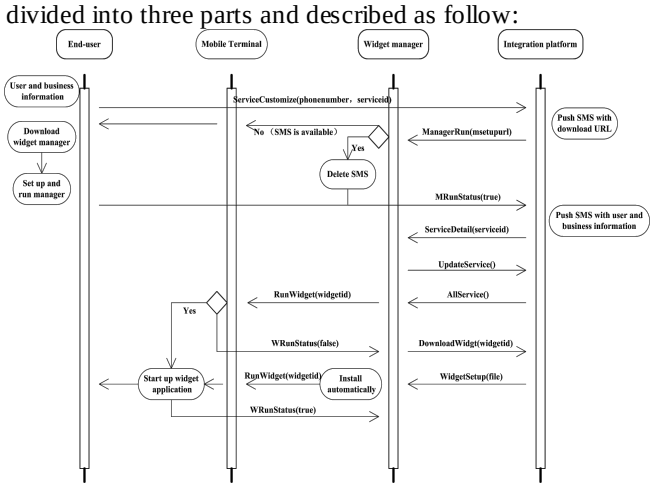


Figure 2. The work flow of integration framework

Firstly, it is very important to make sure whether widget manager has set up in the terminal. Service consistency platform pushes SMS to mobile terminal and waits confirmation message. Widget manager will send successful message to platform immediately after intercepting SMS if manger has been running in the terminal. Else user can download widget manager according to the URL providing by SMS. When set up successfully, widget manager shall send confirmation message to the platform then. The format of SMS should be:

http://<serviceurl>/Download/<widget manager>

Secondly, updating mapping table is completely necessary. Mapping table is one-to-one mapping of widget application information and service information. Only known the latest mapping table, widget manager can accurately arrange widget application to show service. Service consistency platform will push SMS with user service data and application identifier to mobile terminal when receiving confirmation message. Widget manager will obtain this SMS, and then request to update the mapping table. A sample XML can be used to update the map with the following table format

<widget application information:
WidgetId(Unique),
address of download,
app description;
service information:
ServiceId(Unique),
service description; index1, index2…>

The format of SMS should be:

http://<serviceurl>/WidgetID/index1&index2&...

At last, widget manager dispatches widget application to display service. Manager finds the right application identifier from the mapping table in terms of application identifier pushed and tries to start it. We deem that the failed message means the application has not been set up. Manager will refer to the mapping table and download the application. Then installation and startup will accomplish automatically with the help of widget manager. The platform will receive the application running successfully message from manager.

From the above work flow, inter-terminal Internet service based on service consistency framework does not bring more burdens to end-user. The complete automatic of process makes the service realize seamless inter-terminals. End-user can feel nothing different especially after widget manager and application have been installed.

IV. IMPLEMENTATION BASED ON SERVICE CONSISTENCY FRAMEWORK

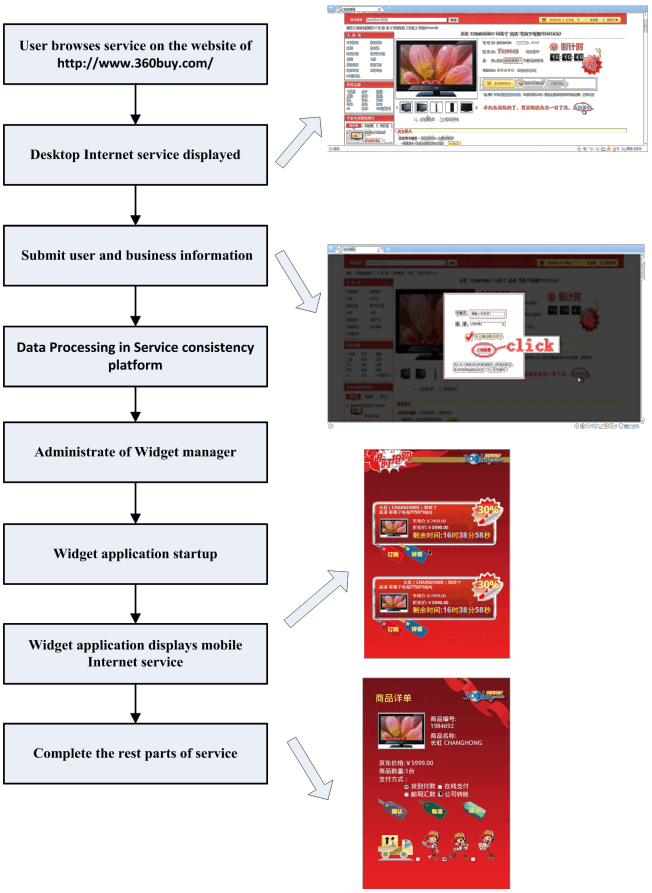


Figure 3. Implementation of 360buy based on service consistency framework

360buy widget application just realizes a service based on service consistency framework. User browses service on the website of 360buy, when he is willing to buy some new discount commodity. However, we have to wait an hour to buy for justice, while we will be offline that time. Then we can customize the service. After all user and service

information is submit, service consistency platform will push the user service data to mobile terminal at the due time. With the help of widget manager, widget application of 360buy successfully runs in user terminal. User can also follow the flow of community distribution. From now on, any user favorite things can be bought by this mobile widget application from 360buy.

To test the service consistency of 360buy widget application, we install it in some mobile terminals with different platforms, such as S60, Windows Mobile and OMS. All these terminals have a pre-install BAE runtime environment (a kind of widget engine).

The running success ratio of the application is more than 95%. The terminal adaptation of service is upstanding. The identity of service is close to 90% thank to the similar technique between web and widget. After the introduction of service consistency framework which resolves the consistency problem fundamentally, the service of continuity is nearly 100%.

TABLE II. PERFORMANCE COMPARISON

Service consistency		360buy widget application	Others (http://m.360buy.com)
Adaptation		>=95%	≈100%
Identity	Data	100%	100%
	Process	>=90%	≈60%
	Display	>=80%	<=70%
Continuity	Data	≈100%	0
	Process	≈100%	0

V. FUTURE WORK

Under the background of network and service convergence, Internet plays a more and more important role. In this case, it is of great importance to realize the Internet service consistency when traditional Internet service is migrated to other terminal [10].

Whether widget is supported by terminals is the key of popularize of Integration framework put forward. However, it is hard to realize that widget can be run int all terminals with different system structure, although more and more terminals are concerned about widget and have done some research or development, such as PDA, TV, sensor, and mobile phone of course and so on. A universal widget engine should be developed to run inter-terminal widget application.

Another problem is to determine the service granularity. It is difficult to decide how many parts service should be divided. And then we have no way to know that the widget should realize which services.

In current service consistency framework, widget manager is only considered as background software. In fact, widget manager is not a widget but a software. So we have to face the adaptation problem.

In future work, the service consistency platform should have the ability of distinguishing different terminals and communicating with each one. Apart from that, service consistency platform can also be regard as a SaaS platform providing widget applications. A widget manager with interface could provide more assistance, like widget searching, application association, service classification. All these are of great value to mobile user behavior analysis.

The purpose of service consistency framework is to make end-user apperceive “anytime, anywhere, anyway” service, with the consistency and continuity of experience.

References

[1] IResearch service. http://news.iresearch.cn/Zt/107684.shtml, 7-2010

[2] Chiu KKS, Lin RJ, Hsu MK, and Huang LH. “POWER OF BRANDING ON INTERNET SERVICE PROVIDERS”. The Journal of Computer Information Systems. Thursday, April 1 2010

[3] Wang Jing-lin. Design of Internet TV Based on Intel CE3100 and Yahoo! Widget Channel. VIDEO ENGINEERING. 2009 33(12)

[4] Miaoqing Tan, Arjona Andres, and Ylä-Jääski Antti. “Real-time service migration for voice over IP services”. 2008 The Second International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies. September 29-October 04.

[5] Liang Weihong. “Integrative Structure Mode Based on C/S and B/S”. Computer Study, 06-2003

[6] Wikipedia. http://en.wikipedia.org/wiki/Client%E2%80%93server_model, July 16 2010.

[7] Anind K. Dey, Daniel Salber, Gregory D. Abowd, and Masayasu Futakawa. “An Architecture To Support Context-Aware Applications”. Georgia Institute of Technology. 1999

[8] Ryu Eun-Seok, Hwang Jeong-Seop, and Yoo Chuck. “Widget integration framework for context-aware middleware”. Lecture Notes in Computer Science. Volume 3744/2005

[9] Srbljic S, Skvorc D, and Skrobo D. “Widget-Oriented Consumer Programming”. Journal for Control, Measurement, Electronics, Computing and Communications, Vol.50 No.3-4 December 2009.

[10] Sultan Florin, Bohra Aniruddha,Iftode, and Liviu. “Service continuations: An operating system mechanism for dynamic migration of Internet service sessions”. 22nd International Symposium on Reliable Distributed Systems (SRDS'03)

 ISBN: 978-1-61208-100-7

Design and Development of an Interoperation Framework in a Smart Space Using OSGi

Soma Bandyopadhyay
Innovation Lab, Kolkata
Tata Consultancy services
Soma.bandyopadhyay@tcs.com

Naga Kiran Guddanti
Innovation Lab, Kolkata
Tata Consultancy services
kiran.naga@tcs.com

***Abstract*—This article presents an approach towards interoperating i.e., working together with heterogeneous objects in a ubiquitous computing environment. A framework for interoperation in a ubiquitous environment as a ubiquitous service is proposed here. Smart spaces and environments like smart homes, smart healthcare system, and smart vehicular systems are the important application areas of ubiquitous computing. Here the smart home is considered as smart space and home gateway as a smart controlling device capable of performing interoperation among diverse objects inside the home. Proposed interoperation-framework service resides in this controlling device/gateway. The Open Services Gateway Initiative (OSGi) framework is used for developing this interoperation framework. Creation of service provider bundle and user bundle for the associated service interoperation-framework are depicted in detail with case study and results. The same concept can be extended in any smart environment to achieve a generic framework for interoperability to connect heterogeneous devices and to address multiple services.**

Keywords-Ubiquitous computing; Home gateway; OSGi; Smart sapce; Interoperation.

I. INTRODUCTION

Smart space comprises a number of diverse devices having different physical communication mechanism like wired, wireless, power line etc. as well as it serves multiple applications. Interoperation is an important attribute to build up any smart space. There have been a lot of researches made in building up a smart space like smart home, smart healthcare and vehicular system mainly to meet the challenges on interoperation, adaptation, addressing the context of the situation, and security aspects of the said ubiquitous environments. Among them interoperation is the most fundamental one to identify/address the heterogeneous devices and to define protocol of interoperation based on the semantics and dynamical adaptation. Therefore, there is a need for procedures/services for performing interoperation among diverse devices dynamically.

In this article, the focus has been given to design and develop an interoperation framework as OSGi services based on OSGi framework. Here this interoperation service acts as gateway to exchange data among the devices situated inside a home. Every device communicates to each other as well as gets data from Internet through this gateway device only. It is the only device where OSGi service is used to run. Participating devices need only standard applications, web interface, simple J2ME (java2 micro edition) based applications to perform interoperation. Hence this solution is easy to be deployed and cost effective. Also this interoperation solution is easy to manage since 1) inclusion of any new device seeking for interaction with any other participating device requires modification in interoperation framework inside gateway device only, 2) minimal modification may be required in the existing solution if the new device falls within the class of any existing participating device, 3) failure of any participating device is easily detectable here, 4) an inventory of active interoperating devices can be made readily available and Web based software updating among the participating devices can be done easily, and 5) any service addition or feature enhancement inside the proposed solution involves the only central device, i.e., this gateway device.

The case study presented here acts as RTP/RTSP (real time transport protocol/ real time streaming protocol) video data forwarder between two devices communicating only through this interoperation service. The devices in the above mentioned case study have different physical layer communication mechanism.

The remainder of this article is organized as follows. First, the related work in interoperation in smart space is presented, followed by an overview of the proposed system. The implementation and experimental study are then described in detail. The final section concludes this article with future scope of the current proposed model.

II. RELATED WORK

Interoperation in a smart space using OSGi (Open Services Gateway Initiative) framework is broadly studied in cases like 1) distributed peer to peer model [1] where OSGi platforms are required to run on multiple devices or participating nodes to distribute device dependent services over several devices, 2) mobile agent technology [2] with a prerequisite that every device has an OSGi based agent

 ISBN: 978-1-61208-100-7

embedded in it, and 3) using R-OSGi (Remote – OSGi) in the devices for interoperation [3]. All these cases require OSGi to run in all the participating devices, this demands every device needs JVM (java virtual machine) and operating system including Bluetooth, Zigbee based devices which have limited computation power and resources. But it is unrealistic to demand that every device has an OSGi based services implanted in it. At the same time these approaches do not specify how to manage the system.

In this paper a framework for interoperation for a smart space is proposed. This is generic in nature. It runs inside a central location, where participating devices do not require any OSGi based application. This solution is cost effective, easy to deploy, and manage. Here focus is given to the design and development of an interoperation framework consisting of multiple layers that works as a middleware stack, interconnects various heterogeneous devices, and acts as a gateway of transferring any data among them. It designates unique identifier to the participating devices. Addition of new functionality to this framework can be performed easily without modification in the interoperating devices.

III. SYSTEM OVERVIEW

The interoperation service as proposed here is developed on OSGi [4] middleware. It maintains diverse connectivity to multiple heterogeneous devices like UPnP (universal plug and play) [5], Bluetooth [6], Zigbee [7], Wi-Fi [8], and Ethernet. These devices exchange data through this proposed system, where the system acts as gateway device, and also acts as router as well as control point of the participating devices (Device1, Device 2 … Device 5 as depicted in Fig. 1), which desire interoperation between each other and thus an interoperation zone is established as depicted in Fig. 1.

The proposed system designates a unique identifier to each device during the time the devices join into the interoperation zone. It has interfaces for demanding any registered service – example streaming video, playing movie, sharing images. It has also interfaces to add new services.

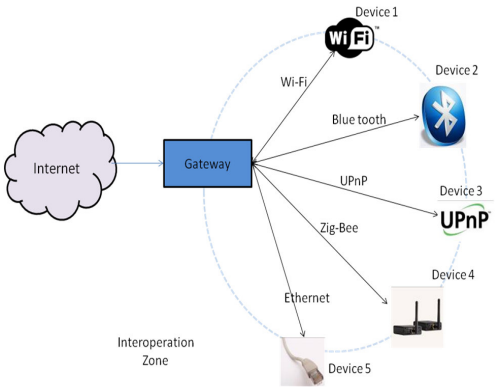


Figure 1. Interoperation network

Participating devices use readily available agents (with modification if necessary) to demand services from the proposed gateway or interoperation system, by means of the interfaces exposed by the same.

IV. SYSTEM INPLEMENTATION

This section portrays the system implementation in detail. The proposed interoperation framework is developed by using Knopflerfish [9]-OSGi framework. It runs as OSGi service. It has both service-provider and service-user bundle [10]. The overview of the proposed architecture is depicted in the Fig. 2.

The proposed system comprises the following multiple layers:

- Interface
- Core control (Threaded, interoperation control module, with multiple event handlers)
- Services

The interface layer provides the APIs (application programming interface) to interact with the interoperation framework viz. 1) joining the interoperation zone and getting registered with proposed system, 2) obtaining a unique identifier, 3) requesting various services. Interface layer uses message producer, topic publishing mechanism for handling events. Some APIs are: 'interopStart' - that starts the core control layer of the interoperation framework, 'interopStop' - that stops the proposed service.

Core control layer, the main component of interoperation framework, runs as thread. It triggers the services requested by the diverse devices after interaction with the service layer and after communicating with concerned communication-driver for interacting those devices.

Device detection and providing of a unique identifier to these devices are some of its main functionalities. Here IPv6 (Internet protocol version 6) address is assigned as unique identifier.

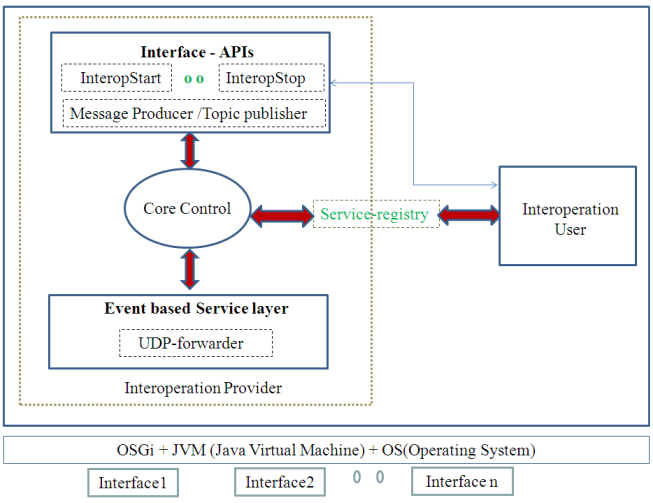


Figure 2. Functional blocks of interoperation framework

This core module registers events, associates services with these events, and performs the event handling on getting a notification of the occurrence of an event. It interacts with the service registry of the OSGi framework for interacting with other OSGi service bundles. It communicates with the multiple interfaces like Bluetooth (for communicating with mobile), Wi-Fi, Zigbee, UPnP through the operating system for exchanging data with heterogeneous devices in interoperation zone.

The service layer performs the service specific functionality as activated by the core layer. The services are furnished to the concerned participating device which invokes/requests that service by using APIs provided by the interface layer. Services act as event handler. Here, the point to be noted is that gateway on which this interoperation framework runs is the only common point among any devices for interoperation. Some of the service examples are - 1) acting as a data-forwarder among any devices, 2) video streaming from Internet, from local video server, 3) sharing non real time data, 4) getting notification of receiving SMS (short message service), email in TV (Television) while watching it.

The service user bundle of the proposed system invokes the interoperation provider and starts the interoperation framework in active mode. Following blocks summarize the steps which need to be invoked sequentially to create a service provider bundle and user bundle here.

Creation of service provider bundle:

1. Creation of service interface
    - Creation of service interface package
    - Defining the service interface class
2. Creation of service provider & implementation of service interface class
    - Creation of service provider
        - Creation of a service implementation package
        - Importing the service interface class (defined in service interface creation in step- 1)
        - Implementing the provider class
    - Implementation of service interface class
        - Use the same service implementation package
        - Importing the service interface class (defined in step-1)
        - Implementation of the interface class
3. Package of interface class to be exported to manifest.mf file

Creation of service user bundle:

1. Creation of service interface
    - Same as mentioned in service provider
2. Creation of service user class
    - Creation of service user
        - Creation of a service user implementation package
        - Importing the service interface class (defined in service interface creation)
        - Implementing the service user class

V. EXPERIMENTAL RESULTS

The complete system is implemented by using an Intel desktop-board with Linux where the interoperation is performed by using one Wi-Fi and one Ethernet interface. A local video server connects over Ethernet and a laptop connects over Wi-Fi / Ethernet with the proposed system. The software and hardware configuration details are shown in table 1 and table 2 respectively. The experimental setup is depicted in Fig. 3.

During experimentation interoperation services are executed first. The laptop registers with interoperation framework and requests RTP/RTSP video streaming service specifying the port number. Here port 3000 is used as receive port for the video service. A manual setup has been made at the local video server side with IPv6 address of the proposed system and its associated port number 2500 needed for UDP (user datagram protocol) data forwarding service. The local video server and laptop use VLC (video LAN client) media player [11], Fig. 4 and Fig. 5 represent the setup respectively. Fig. 6 depicts the interoperation framework in running condition.

 ISBN: 978-1-61208-100-7

TABLE I. SOFTWARE CONFIGURATION

The software Configuration	
Software configuration	***Description***
OSGi Framework	Knopflerfish 2.3.3
Java Virtual Machine	JDK 1.6
Operating System	Linux - Ubuntu 9.10
Java-IDE	eclipse 1.2.2

TABLE II. HARDWARE CONFIGURATION

The Hardware Configuration	
Hardware configuration	***Description***
Target platform	Intel® Desktop Board D945GCLF2
Processor	Dual-Core Intel Atom Processor 330 integrated at 1.6GHz

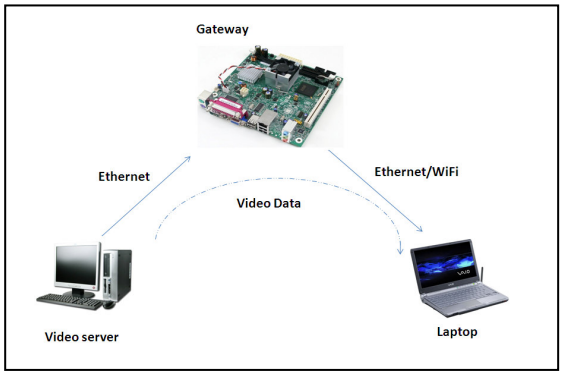


Figure 3. Experimental setup

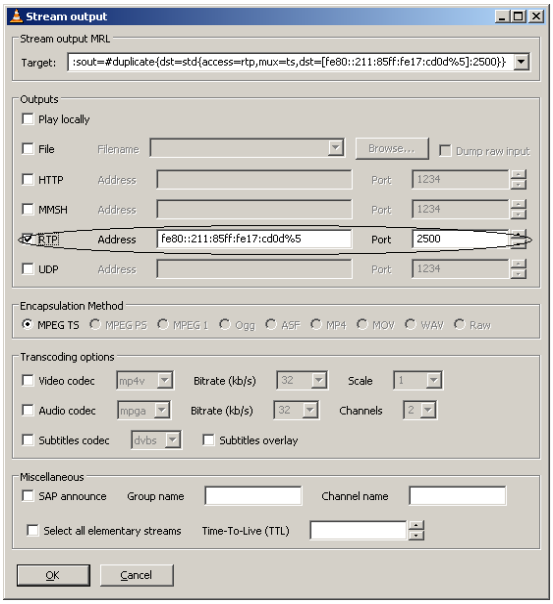


Figure 4. VLC media player setting (video server side)

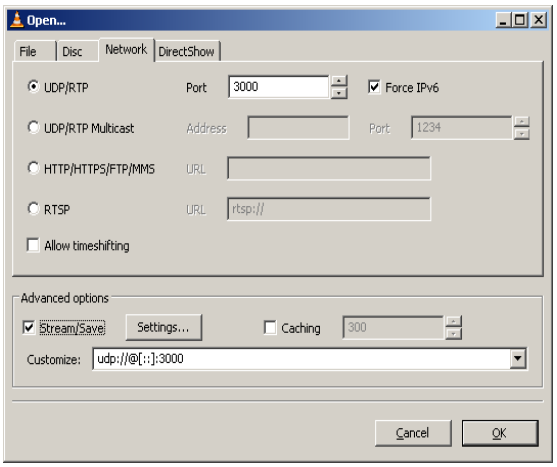


Figure 5. VLC media player setting (Laptop)

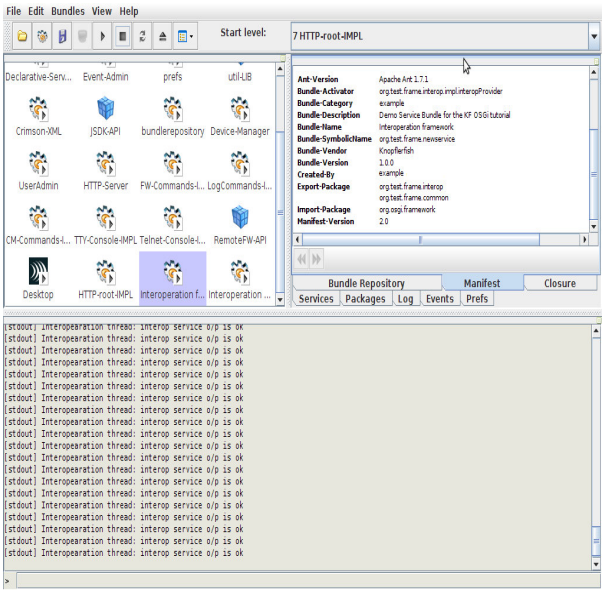


Figure 6. Interoperation with UDP forwarder service

VI. CONCLUSION AND FUTURE WORK

In this article, design and development of an interoperation framework for a smart environment like smart home consisting of heterogeneous devices with different physical communication scheme like Bluetooth, Wi-Fi, Ethernet, UPnP, Zigbee have been proposed. The framework acts as control point as well as a router among any devices. It does not require any corresponding OSGi module at the heterogeneous devices seeking for interoperation. It provides interfaces to add new services into its service layer. Therefore it is easy to manage, cost effective and easy to be deployed in real world.

Suggested system is developed based on Knopflerfish's OSGi framework and runs as OSGi service.

Video streaming based on RTP/RTSP from a local video server through the proposed interoperation service is

 ISBN: 978-1-61208-100-7

depicted as a use case. Besides, the methods of writing bundles for service provider and service user are discussed.

The limitation of the proposed system lies in its incapability of doing complete automation along with identification of context. However, the proposed system can be enhanced further by providing web based GUI (graphical user interface) for the purpose of automatic triggering of any existing service or to add new services and to make it context aware with limited use of sensors and considering mobile as the essential resource of user context data .

There is also scope for further work in making the complete system generic so that it can be used in any smart space, example - as a part of a car gateway. We are continuing to carry researches on the above mentioned future scope of work and enhancement of the proposed system.

REFERENCES

[1] Chao-Lin Wu, Chun-Feng Liao, and Li-Chen Fu, "Service-oriented smart home architecture based on OSGi and mobile agent technology", Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transaction, vol.37, issue 2, March 2007, pp. 193-197,doi: 10.1109/TSMCC.2006.886997, 22.03.2010.

[2] Seung Keun Lee, Jeong Hyun Lee, "OSGI based service mobility management for pervasive computing environments", 24th IASTED Internation Multi conferecnce on Internet and Multimedia sytems and Applications, Feb 2006,pp. 159-164, 05.04. 2010.

[3] Jan S. Rellermeyer, Michael Duller, Ken Gilmer, Damianos Maragkos, Dimitrios Papageorgiou, and Gustavo Alonso "The software fabric for Internet of Things", Lecture notes in computer science, Proceedings of the 1st international conference on The internet of things, 2008, pp. 87-104, 05.04. 2010.

[4] OSGi, http://www.osgi.org/Specifications/HomePage, 10.04.2010

[5] UPnP, http://www.upnp.org,10.04.2010

[6] Bluetooth, http://www.bluetooth.com, 17.07.2010

[7] Zigbee, http://www.zigbee.org/, 17.07.2010

[8] Wi-Fi,http://www.wi-fi.org/certified_products.php,17.07.2010

[9] www.knopflerfish.org, 10.04.2010

[10] Choonhwa Lee, David Nordstedt, and Sumi Helal, "Enabling Smart Spaces with OSGi", IEEE CS and IEEE ComSoc, PERVASIVE computing journal, vol.2, issue 3, pp. 89 -94, doi: 10.1109/MPRV.2003.1228530.computing journal, vol.2, issue 3, pp. 89 -94, doi: 10.1109/MPRV.2003.1228530, 24.03 2010.

[11] http://www.vlcmediaplayer.org, 10.04.2010

 ISBN: 978-1-61208-100-7

Particularized Cost Model for Data Mining Algorithms

Andrea Zanda
Universidad Politecnica Madrid
Facultad de Informatica
Madrid, Spain
andrea.zanda@alumnos.upm.es

Santiago Eibe
Universidad Politecnica Madrid
Facultad de Informatica
Madrid, Spain
seibe@fi.upm.es

Ernestina Menasalvas
Universidad Politecnica Madrid
Facultad de Informatica
Madrid, Spain
emenasalvas@fi.upm.es

***Abstract*—Ubiquitous devices demand autonomous and adaptive data mining. Despite some advances, the problem of calculating the cost associated to the execution of data mining algorithms is still a challenge. Thus, in this paper we provide a method for predicting the cost in terms of efficacy and efficiency associated to a mining algorithm, the resulting cost model as shown in our previous work can be exploited by a mechanism for predicting the best configuration of a mining algorithm according to context and resources. Recent work presents how a cost model not associated to any dataset can provide reliable estimations on efficiency and efficacy, here we present how we can improve the accuracy of such estimations by particularizing cost model to a predefined dataset. We provide the guidelines of the method and then we present a particularized cost model for C4.5 algorithm associated to a specific dataset (Parkinson's telemonitoring). Experimental results show how the particularized cost model achieves significant better estimations than the general cost model.**

***Keywords*-ubiquitous, data mining, cost model, algorithm.**

I. INTRODUCTION

The dissemination of ubiquitous devices has become a reality, nowadays such devices are able to execute almost any kind of application and collect considerable amount of data. To endow the devices with data mining services in order to exploit such amounts of data is a requirement. Applications in many domains require embedded intelligence to achieve their goals [7] [10], but their intelligence is not always personalized or adaptable. In [6] the authors provide a review of mobile care system which support the patient according to predefined mining models or to a server communication. An example on how to provide mobile devices with intelligence is in [1], an application of a neural network approach for the development of a system for knowledge classification in diabetes management. In the domain of intelligent transportation systems there are many situations in which an intelligent component is needed because internet connectivity in order to communicate with a server is not possible. In [8], the authors present a novel context-aware framework integrating intelligence for transportation systems. The system is able to: (1) learn patterns collisions by monitoring, (2) learn to recognize potential hazards in intersections and (3) warn particular threatened vehicles. Nevertheless, also in this domain the data mining framework has not been fully explored and developed. It is clear then how the integration of the mining technique directly into the devices can considerably increase the utility of ubiquitous applications, personalizing, assuring privacy and adapting to the changing world.

Data mining in ubiquitous devices has at least two requirements, to lead the process in an autonomous way and to adapt the process to the changing world. Some works in literature [3] and [2] provide approaches to adapt stream mining algorithms according to context information and available resources, but solutions for adaptable algorithms for the static case are still lacking. Further the methods applied for stream scenarios cannot be applied to batch scenarios, in fact in batch algorithms it is not possible to control the execution while the process is running, the initial algorithm configuration cannot be modified. In [4], some works providing methods for seeking the optimum neural networks algorithm configuration are presented. The main drawbacks concern the resource consumption of the methods to find the optimum and the fact they do not take into consideration external factors as context and internal resources, but only the dataset to be mined.

In [13], a mechanism able to select the best configuration for a C4.5 algorithm according to resources and context was presented. The mechanism is based on the EE-Model, which is able to estimate the efficiency and the efficacy of the C4.5 algorithm in terms of memory, CPU cycles, battery and accuracy, given the metadata of the dataset to be mined and the algorithm configuration. The model is calculated on the past behavior of the algorithm, which has been executed with different configurations and datasets. The main advantage of the model presented is the generality as it can be used to predict the efficiency and efficacy of that algorithm in any circumstances and with any dataset, but this is also its main drawback as it is not particularized. Normally in a particular device the dataset features will not vary and this is our main motivation to present a particularization of the EE-Model for a given dataset. As experiments will show, the particularization of the EE-Model for C4.5 algorithm makes it possible to get more accurate prediction on memory and CPU cycles.

The rest of the paper has been organized as follows: in

 ISBN: 978-1-61208-100-7

Section 2 we focus the paper on the requirements of a cost model associated to a data mining algorithm, Section 3 sets the problems for calculating the cost model. In Section 4 we describe the guidelines in order to build a P-EE-Model and then we present a P-EE-Model for C4.5 associated to a Parkinson's tele-monitoring dataset. In Section 5 we show experimental results on the customized cost model. The Section 6 presents the conclusions and the future research.

II. Preliminaries

In [13], a mechanism to select the best algorithm configuration to execute a mining algorithm, taking into account information regarding the situation is presented. What they call situation is defined by the external factors, and it is divided into two main groups:

- Factors describing resources: memory, battery, CPU;
- Factors describing context information: information that can be sensed from sensors (location, temperature, time, etc).

Consequently, the authors divide the issue to decide the best configuration of the mining algorithm into two subproblems, on one hand how the external factors influence the requirements of the mining process in terms of efficacy, efficiency and semantics (meaning of the results), and on the other hand how the algorithm behaves when altering input data and input parameters. The main assumption under the division into two subproblem is that no matter the external factors, the algorithm inputs determine the quality of the model and performance that can be obtained. By efficiency they understand the resource consumption of the execution. On the other hand, the efficacy in a classification algorithm can be defined as accuracy (i.e., percentage of corrected classified items).

The method behind the cost model (EE-Model) presented in [13] relies on historical analysis of past execution of a particular algorithm to calculate the influence of inputs on the cost and results of the model. This is to say, information on the cost of past executions of the algorithm on different configurations and with different datasets are analyzed and knowledge discovery process is applied to extract rules that can be used to predict the bahaviour of the algorithm in new cases. As the experimental results show, the model presented there (EE-Model) provides estimations which are closer to the real efficacy and efficiency, nevertheless it presents the following drawback: it has been defined for general datasets, this is to say, historical executions analyzed consider different datasets. The features of a particular dataset can influence the behavior of the algorithm differently and this has motivated the present research in which we propose to particularize the cost-model depending on the features of a particular dataset.

Consequently it would be good to have a particularized cost model built on a determined dataset that could lead to more accurate prediction of the behavior of the algorithm both in terms of efficiency and efficacy. The underlying drawback behind is the lack of flexibility as the EE-Model customized this way would only be valid for that particular dataset. Nevertheless in real cases the dataset of a particular domain or application will change only in terms of number of records, size and distribution, all factors which our particularized cost model can adapt to. As experiments will show the particularized model can provide significantly more accurate estimations. In what follows we first present the problem and later we present the particularized EE-model.

III. Setting the problem

The same mining algorithm can lead to different resource consumption and different accuracy of the model depending on many factors, but which are the factors altering such behavior? And how do they alter such behavior? As it is depicted in Figure 1, the mining algorithm efficiency and efficacy depends on:

1) The input data;
2) The configuration.

We describe in depth these features in the next subsections to see how they affect the algorithm behavior to take advantage of these features to better predict algorithm resource consumption and performance. Note that the analysis of the semantics is out of scope in this paper.

Figure 1. Mining algorithm inputs and outputs

A. Input dataset

The input dataset is the main input to the mining process. Depending on the data quality so there will be the results. Consequently we analyze in what follows how the data quality can impact the process. The quality of the data is related to:

- The number of records;
- The number of attributes;
- The type of each attribute;
- The values and distribution.

The dataset features will influence both the efficiency and the efficacy of the algorithm, in this sense for example a bad quality dataset in terms of data distribution can lead to a not precise model. Note that for example the number of columns could affect the efficiency, although it could also affect the quality of the model, increasing the number of columns leads to high dimensional problems that can

 ISBN: 978-1-61208-100-7

be a significant obstacle to achieve high quality models. Also increasing the size of the dataset will probably result in a lower efficiency, but then the efficacy has to be explored.

B. Algorithm configuration

Setting the configuration of the algorithm means to assign values to the algorithm parameters for an execution. The configuration also determines the resource consumption and the accuracy of the results. In [5], a number of algorithms are tested with the same dataset in order to analyze their performance, the authors show the resource consumption and the accuracy achieved by testing them with different configurations. Such relations between configuration and result have to be known. Binary split option of a C4.5 algorithm for example can increase the efficiency of the algorithm because building a more branched tree, nevertheless the option can be suitable for certain types of dataset and increase the efficacy.

IV. Approach

In [13], a mechanism able to select the best configuration to execute the C4.5 algorithm according to external factors is presented. Figure 2 shows how the various phases of the process, the mechanism has a central role, it can access dataset information, configuration metadata and external factors, and it gives as result the best configuration for the mining algorithm. The EE-Model supports the mechanism by providing estimations on efficacy and efficiency of the mining algorithm execution. This solution has many advantages, first of all the system can have an estimation of the resources needed for the execution, in some cases the system can avoid executions that cannot be terminated (for battery low for example). It is also possible to avoid out of memory problems and CPU bottlenecks. In fact, for ubiquitous devices, resource aware is an important process requirement. In this paper our goal now is to: check that the EE-Model can get better results when built for a particular dataset. In Section IV-A we present how to obtain the particularized EE-Model.

A. The particularized EE-Model

The dataset features of a particular dataset can influence the behavior of the algorithm, consequently it would be good to have a particularized model for certain datasets in those domains or applications where we know the dataset features will not dramatically change. The underlying drawback behind is the lack of flexibility as the EE-Model cannot be valid for any type of dataset, but on the other hand a cost model suitable for certain purposes would improve the accuracy of the estimations. Here we will focus on the guidelines for the particularization of the EE-Model predicting C4.5 classification algorithm.

In order to build the EE-Model the steps are the following:

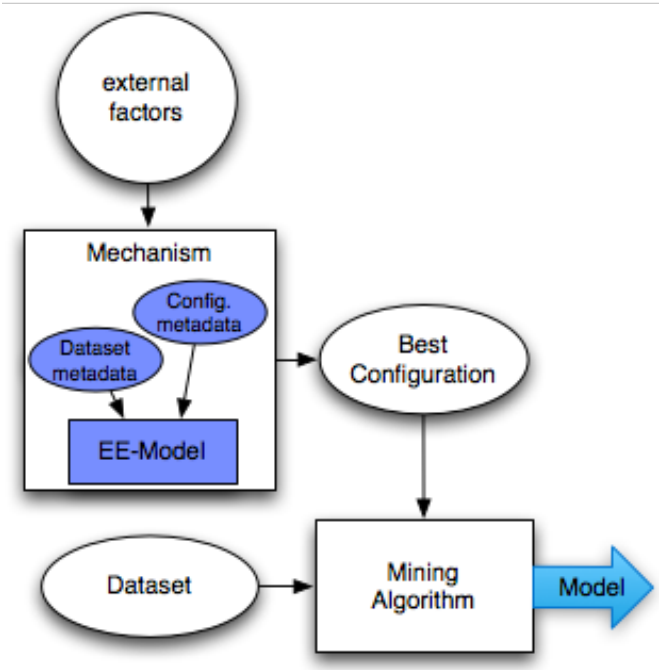


Figure 2. Approach

1) Define the set of variables to describe the executions of the algorithm:
   - Define the *condition variables* that describe the algorithm inputs as parameter settings (i.e. type of pruning) and dataset metadata. The dataset metadata in [13] is defined as number of attributes, type of attributes and so on, are not needed as they are not changing for a particular dataset, nevertheless we include richer and ad hoc information that describe that dataset. In fact the metadata has to include information such as of number of records and dataset size, class distribution and attributes distribution, in general any feature of the dataset that might change over time.
   - Define the *decision variables* that describe the set of executions information as memory and battery to name a few, or any measure is needed to predict with the model.
2) Execute a representative number of times the algorithm altering the condition variables;
3) Apply knowledge discovery process to the collected data relative to the executions in order to build one model for each decision variable. Most of times the decision variables will be numeric. According to our experience, we suggest to apply techniques as linear regression or regression tree.

In order to build the EE-Model the steps are the following:

B. Particularized EE-Model for parkinson's tele-monitoring

After an outline on how to build the model, here we present the EE-Model we built for a dataset relative to Parkinson's tele-monitoring [11]. The dataset is composed of a range of biomedical voice measurements with early-stage Parkinson's disease recruited for remote symptom progression monitoring, the description of the dataset attributes is given in Table I.

Table I
DATASET ATTRIBUTES DESCRIPTION

subject	Integer that uniquely identifies each subject
age	Subject age
sex	Subject gender '0' - male, '1' - female
test-ime	Time since recruitment into the trial
Jitter	Several measures of variation in fundamental frequency
Shimmer	Several measures of variation in amplitude
NHR,HNR	Measures of ratio of noise to tonal components in the voice
RPDE	A nonlinear dynamical complexity measure
DFA	Signal fractal scaling exponent
PPE	A nonlinear measure of fundamental frequency variation
total-UPDRS (CLASS)	Clinician's total UPDRS score (discretized)

Following the guidelines of Section IV-A:

1) We define the decision variables as in II, there are two measures on the efficiency and one on the efficacy of the algorithm. The accuracy is obtained with an evaluation of the model with a test set. Then we defined the condition variables as in Table III. We can notice that the dataset metadata contains on one hand information on the size of the dataset, on the attributes type and in general on the distribution of the attributes values, on the other hand the metadata associated to the algorithm parameters (in order to represent all the possible different configurations).

Table II
ALGORITHM EXECUTION INFORMATION

Memory	Memory used	Integer
CPU	Number of CPU cycles	Integer
Accuracy	Accuracy of the obtained mining model	Real

2) We obtain a dataset of historical data of execution of the algorithm in a system with 2.16 GHz Core 2 processor and 2.5GB 667 MHz DDR2 SDRAM memory. The number of execution is an important point to obtain a dataset able to represent the domain, we generated a number of 30023 covering all parameter configurations (increment of 0.10 for continuous parameters) related to the same dataset, but with different number of records and so with different class and attributes distributions.
3) This step concerns the application of data mining techniques in order to discover the relations between condition variables and decision variables. The algorithm used for memory and CPU cycles is REPTree [12], for accuracy linear regression [9]. The model for accuracy has not improved the previous results achieved, while the results for memory and CPU cycles are shown in Section V.

Table III
INPUT INFORMATION

Attribute number	Number of attributes	Integer
NInstances	Number of instances	Integer
Size	Dataset size in KB	Integer
Attribute distinct	distinct values of the column X (i.e. Class)	Integer
Attribute StdDEV	Standard deviation of the column X (i.e. Class)	Real
Attribute type	Number of columns of type Y (i.e. real)	Integer
Pruning	Whether pruning is performed. ('0' → no pruning, '1' → pruning, '2' → Reduced error pruning)	Nominal
Binary	Whether to use binary splits on nominal attributes when building the trees	Boolean
Laplace	Whether counts at leaves are smoothed based on Laplace	Boolean
CF	The confidence factor used for pruning (smaller values incur more pruning)	Real
Sub	Whether to consider the subtree raising operation when pruning	Boolean
MinNumObj	The minimum number of instances per leaf	Integer
NumFolds	Determines the amount of data used for reduced-error pruning. One fold is used for pruning, the rest for growing the tree.	Integer
Seed	The seed used for randomizing the data when reduced-error pruning is used.	Integer

V. EXPERIMENTATION

The experimentation is carried out evaluating the efficiency prediction of the presented particularized EE-Model (P-EE-Model) for Parkinson's tele-monitoring in comparison with the general EE-Model in [13]. The estimations of the two models are compared with the values of the real executions. Given a configuration and dataset metadata the EE-Model is able to estimate efficiency and efficacy, so in order to describe the experiment we first define the configurations and the dataset metadata we consider for our analysis. We define the three configurations presented in Table V, they mainly differ in the type of pruning applied in the execution. The reason is related to the fact that the pruning is the main parameter setting altering the efficiency of the algorithm. We evaluate the P-EE-Model with the Parkinson's tele-monitoring dataset we used for building the model, its description is in Table I, but here we take a sample of the original dataset having the number of records equal to 2543, the size to 424KB and different attributes and class distribution.

Then we evaluate the P-EE-Model comparing the estimations for the three configurations to the real values. Table V shows the results of the evaluation, the models provide more reliable for average memory, in fact a mean is supposed to be more stable.

 ISBN: 978-1-61208-100-7

Table IV
ALGORITHM CONFIGURATIONS

Config.	Prun	Bin	Lap	CF	Sub	MinOb	#Folds	S
1	1	Yes	Yes	0.25	Yes	3	–	–
2	0	Yes	No	–	0	2	–	–
3	2	No	No	–	Yes	5	5	5

Table V
EE-MODEL EVALUATION

	Average Memory	CPU cycles
Correlation Coeff.	0.95	0.99
Relative absolute error	8.0%	20.0%
Root relative squared error	9.9%	15.3%

Now we compare the performance of the P-EE-Model with the general EE-Model (G-EE-Model) in [13]. Figure 3 shows the absolute squared error obtained for the three configurations while considering first the P-EE-Model (sky blue) and then the G-EE-Model (red). The comparison is relative to the average memory and denotes a significant improvement on the accuracy of P-EE-Model estimations.

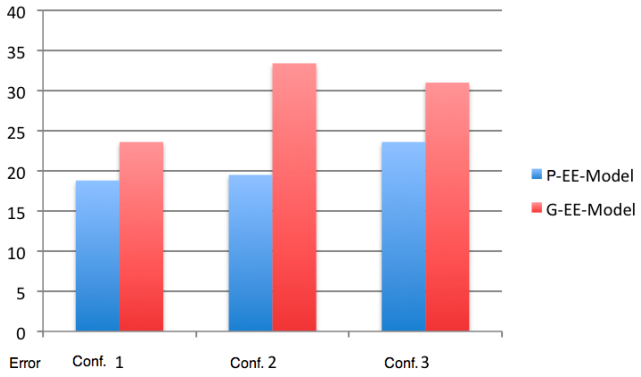


Figure 3. Comparing average memory

Figure 4 shows the comparison of the absolute squared error relative to CPU cycles, even in this case the estimations of the P-EE-Model overcome the general one. In this paper we argued the hypothesis that an EE-Model build for a particular dataset could achieve better estimations than a general one built on many datasets. According to the results above the hypothesis is verified and the estimations overcome the general model significantly.

Nevertheless, the drawback of the P-EE-Model concerns the lack of flexibility, in fact the model is only usable for the dataset on which it is built. To conclude we carried out an evaluation of the P-EE-Model which has been built on the dataset in [11], with another synthetic dataset. As we argued the estimations of such model are worse than the one provided by the general EE-Model.

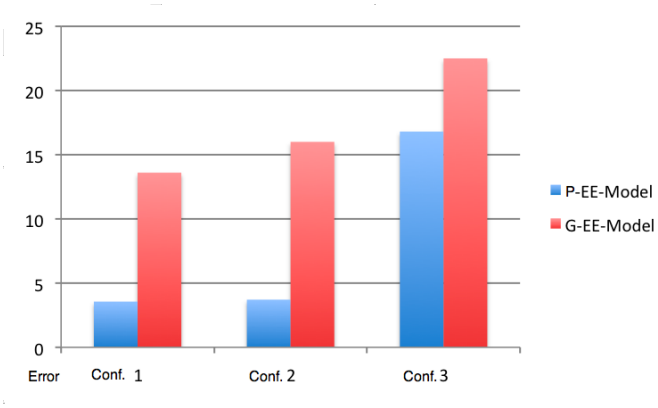


Figure 4. Comparing CPU cycles

VI. CONCLUSION

In this paper, we have presented a cost model to predict the behavior of a data mining algorithm with a specific dataset in terms of efficacy and efficiency that overcomes in accuracy the previous general cost models predicting the algorithm behavior with any kind of dataset. After describing the guidelines in order to build a particularized cost model (P-EE-Model), we present a P-EE-Model for C4.5 algorithm specific for a Parkinson's tele-monitoring dataset. According to our experimental results the E-PP-Model significantly improve the estimations of CPU cycles and average memory. Nevertheless the drawback of the P-EE-Model: less flexibility as it is associated to a specific dataset, has not to be ignored in domains of applications where the dataset can change dramatically.

REFERENCES

[1] G. Gogou, N. Maglaveras, B. V. Ambrosiadou, D. Goulis, and C. Pappas. A neural network approach in diabetes management by insulin administration. *J. Med. Syst.*, 25(2):119–131, 2001.

[2] P. D. Haghighi, M. M. Gaber, S. Krishnaswamy, and S. Loke. An architecture for context-aware adaptive data stream mining.

[3] P. D. Haghighi, A. B. Zaslavsky, S. Krishnaswamy, and M. M. Gaber. Mobile data mining for intelligent healthcare support. In *HICSS '09: Proceedings of the 42nd Hawaii International Conference on System Sciences*, pages 1–10, Washington, DC, USA, 2009. IEEE Computer Society.

[4] A. Kuzmenko and N. Zagoruyko. Structure relaxation method for self-organizing neural networks. In *ICPR*, pages IV: 589–592, 2004.

[5] T.-S. Lim, W.-Y. Loh, and Y.-S. Shih. A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. *Mach. Learn.*, 40(3):203–228, 2000.

 ISBN: 978-1-61208-100-7

[6] Y.-C. Lu, Y. Xiao, A. Sears, and J. A. Jacko. A review and a framework of handheld computer adoption in healthcare. *I. J. Medical Informatics*, 74(5):409–422, 2005.

[7] C. Marx, W. Gwinner, J. Krückeberg, U. von Jan, B. Engelke, and H. K. Matthies. Mobile learning applications for education in medicine and dentistry. *Adv. Technol. Learn.*, 4(2):92–98, 2007.

[8] D. Preuveneers and Y. Berbers. Mobile phones assisting with health self-care: a diabetes case study. In *MobileHCI '08: Proceedings of the 10th international conference on Human computer interaction with mobile devices and services*, pages 177–186, New York, NY, USA, 2008. ACM.

[9] Radhakrishna. *Linear Statistical Inference and Its Applications*. John Wiley & Sons Inc, November 1973.

[10] D. Siewiorek, A. Smailagic, J. Furukawa, A. Krause, N. Moraveji, K. Reiger, J. Shaffer, and F. L. Wong. Sensay: A context-aware mobile phone. In *ISWC '03: Proceedings of the 7th IEEE International Symposium on Wearable Computers*, page 248, Washington, DC, USA, 2003. IEEE Computer Society.

[11] A. Tsanas, M. A. Little, P. McSharry, and L. Ramig. Accurate telemonitoring of parkinsons disease progression by non-invasive speech tests. *IEEE Transactions on Biomedical Engineering*, 2009.

[12] I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 2005.

[13] A. Zanda, S. Eibe, and E. Menasalvas. Adapting batch learning algorithms execution in ubiquitous devices. In *MDM '10: Proceedings of the 2010 Tenth International Conference on Mobile Data Management: Systems, Services and Middleware*, Kansas city, USA, 2010. IEEE Computer Society.

 ISBN: 978-1-61208-100-7

Bluetooth and Filesystem to Manage an Ubiquitous Mesh Network

Nicola Corriero - Emanuele Covino - Giovanni Pani
University of Bari
Departiment of Computer Science
Italy, Bari, Via Orabona 4, 70125
Email: {ncorriero,covino,pani}@di.uniba.it

Eustrat Zhupa
University of Vlora
Sheshi Pavaresia, Skele,
Vlora, Albania
Email: ezhupa@univlora.edu.al

***Abstract*—Hixosfs is a filesystem that lets you manage a network of Bluetooth devices. We have analyzed this filesystem in contexts of great movement and file sharing systems which requires high performance time. We study the system both for the management and to extract statistics.**

***Keywords*-Ubiquitous mesh; ad hoc networks; sensor networks; Bluetooth; Tracking.**

I. INTRODUCTION

Ubiquitous multimedia computing will change the way we operate and interact with the world with the development of numerous interesting ubiquitous multimedia applications.
Our purpose is to create a layer among operating systems, middleware and user devices to facilitate the modeling of the system and to make possible the efficient use of the embedded devices. The main idea is the use of bluetooth technology to create an ad-hoc indoor network. The system has been implemented using embedded systems linked to each other using bluetooth connection or an ad-hoc wireless network. In each of these, a hixosfs filesystem based database has been used to improve the performance.

For such reasons the choice is to try an innovative solution: hixosfs.
Hixosfs is a Linux filesystem that can be included in the standard Linux kernel. Differences respect to the widely used ext2 filesystem are additional features for storing and efficiently retrieving data.

We introduce scenarios, the critical points of these scenarios, other approaches and our solution. We explain the architecture of our system and all of its components. We explain how it's possible to build an alternative simple approach to this problem with a common filesystem. Finally some comparitions with other approaches.

II. SCENARIO AND CRITICAL POINTS

In this work we present an update of the scenario presented in [12]. Two scenario: bluetooth markentig, file sharing in a mesh network. Bluetooth marketing. In this context, there are mobile devices that stop or pass near a place of interest (library, cinema, stadium). Some device has bluetooth and is available to receive data from our system. Our device must periodically scan the environment looking for some device bluetooth and try to establish an exchange of data used to send content usually advertising media.
The second scenario is a laboratory or a library where people want to share information such as music files. When each user enters within range Action Bluetooth antenna automatically start the sharing of files in a profile stored on their devices (Openmoko).
In the context described above it's possible to identify some critical points of the system. In such points we see the differences with the other solutions offered by the market.
In changing contexts is necessary to have softwares which react in short time. The case of bluetooth devices is lampant. Each device stays in the antenna covered area for few seconds. During this short period the antenna should verify the presence of device services (so verifying the presence of Push) and establish a connection with it if it has not been contacted before. Checking if the device is already in the database of the contacted devices can be time consuming. Contexts on move suggests to use intelligent systems that can decide in a small time.
Client-server systems are not recommended to manage contexts where we need speed of reaction. Embedded system that can decide for themselves are advised.
Despite everything you need to use servers to handle large informations about the contexts.
Send sms bluetooth to someone on the move or turn on the light in an environment requiring considerable reaction. Embedded systems are designed for a permanent use and to carry out little well-defined tasks.

III. OTHER APPROACHES

Normally these situations are handled using PCs with large primary and secondary memories that make possible the use of every operating system and every mean for saving and managing information.
Other approaches of the system require the use of complex databases over servers and/or embedded databases over embedded machines.

In the embedded systems we have evident problems of memory that during the time have solicited light-weight and high-performance ad-hoc solutions. Sqlite [8] is an application that implements all the functionalities of a database

 ISBN: 978-1-61208-100-7

using simple text files. This enlighten the execution load of the system and facilitates the integration of the system inside an embedded system. However, the system installation produces a certain load to the mass memory.
Currently Linux fs as ext2 [7], ext3, reiserfs allows to manage with metainformation related to a file with *xattr* feature. Patching the kernel with xattr you have a way to extend inode attributes that doesn't physically modify the inode struct. This is possible since in *xattr* the attributes are stored as a couple attribute-value out of the inode as a variable length string. Generally the basic command used to deal with extended attributes in *Xattr* is `attr` that allows to specifies different options to set and get attribute values, to remove attributes to list all of them and then to read or writes these values to standard output. The programs we implemented in our testing scenario are based on this user space tool.

The last approaches we have compared is to avoid to use a particular filesystem to manage these scenario. We have compared our solution with an ext2 filesystem and some script in bash.

IV. Our Solution

An ad-hoc filesystem created inside the kernel will be used to handle the data provided by the various sensors of the network. The communication of the devices will be realized using an ad-hoc network implemented with aodv. Finally, the devices will have a linux distribution created ad-hoc with the necessary programs only. Python has been choosed as a programming language for the scripts to make the system independent from the final device. For this reason it has been necessary to install only a python compiler in the devices, saving a lot of space.

A high performance software infrastructure is needed in dynamic systems if we consider the time factor. Nowadays the embedded systems are spreading very rapidly. Little computers designed to spend limited energetical and economical resources, but with precise and well defined duties. On the other side, modern operating systems are designed to use all the resources. Our suggestion is to use the embedded devices like bluetooth antennas, placed by the entrances. The limited dimensions of the embedded devices make possible an adequate use of the space.

V. System architecture

A. Hixosfs

Hixosfs [10] is as an ext2 Linux filesystem (fs) extension able to classify and to manage metadata files collections in a fast and high performant way. This is allowed since high priority is given to the storing and retrieving of metadata respect tags because they are managed at fs level in kernel mode.
The *hixosfs* core idea is that information regarding the content of a metadata of a file belong to the fs structure itself.

Hixosfs has been used to tag multimedia files and bluetooth devices as well as user profiles. In this way all the load has been transfered to the kernel that handles and organizes the hixosfs files as occurrs. The servers and the clients contain partitions that can read and set the hixosfs tags so to manage the database.

In the case of bluetooth file, hixosfs extends the inode definition with a struct tag:

```
struct tag {
#ifdef CONFIG_HIXOSFS_BLUEZ
char macbyte1[3];
char macbyte2[3];
char macbyte3[3];
char macbyte4[3];
char macbyte5[3];
char macbyte6[3];
char devicename[40];
char scanningdate[9];
unsigned int tag_valid;
#endif
}
```

The struct tag has four fields for a total of about 100 byte of stored information, theoretically an inode can be extended until 4 kb then it's possible to customize it with many tags for your purpose. It's convenient to choose tags that are most of the time used in the file search to discriminate the files depending their content. We choose here what was able to maximize the time of search musical files by most commonly used criteria as album or author name and so on.

In our experiment we understood the limits of the originary idea of hixosfs that suggests the compiling of an ad-hoc filesystem for each type of tag. In our system was necessary to compile 3 different filesystems with 3 different types of tags (music, user profile, bluetooth). For such a reason we decided to use a generic version of hixosfs with a generic structure in which is possible to insert file representative tags case by case.

```
struct tag {
#ifdef CONFIG_HIXOSFS_TAG
char tag1[30];
char tag2[30];
char tag3[30];
char tag4[30];
unsigned int tag_valid;
#endif
}
```

In our partition there was only a hixosfs filesystem mounted in RAM with three folders containing files provided by three running processes: bluetooth devices detecting, musical files manager and user profiles.

An example of file handling.
`chbluez -m` sets the mac address of the device.
`chbluez -n` device name (when detected).
`chbluez -d` date of the first detection of the device in the system.

 ISBN: 978-1-61208-100-7

B. Bluetooth

There are two types of services used for file transfer:

- Object File Transfer
- Object Push

Most of the bluetooth devices implement "Object File Transfer" service, in which for file transfer is necessary an authentication using a unique PIN number on both client and server.

Around 40% of the devices implement the "Object Push" service in which is possible to send files without using any PIN number. However, to receive a file is necessary to accept a connection request. This technique facilitates the interaction cause you don't need any PIN number, even if in the majority of the devices there is a hardware lock in case of repeated connection requests via the Push protocol.

C. AODV

We are choosing to use the routing AODV (Ad hoc On demand Distance Vector) protocol, because it's suited for route discovery and routes maintenance within ad hoc network.

To this purpose a Linux kernel and a mini distribution have been compiled and analyzed in the case of an handhelds HTC blue-angel and for Openmoko. Then it has been extended including the Manet support, specially compiling the aodv-uu, the AODV protocol implementation realized by the Uppsala University[4]. At this point we are ready to generate a mesh network. Etherogeneous nodes like handhelds, work station, notebook, mobile phone, router, having wifi connectivity with Linux operating system and aodv-uu module installed, can contribute to the creation of a decentralized newtork, working in a mobile context.

In our system we have cross-compiled code that implements the AODV protocol for ARM architecture [1]. The implementation used provides a daemon that constantly sends " hello"message searching for other nodes. This way when the user comes within range of another device to authenticate.

D. Openmoko

To test hixosfs filesystem for music files we have choose Openmoko Freerunner[5], an open source project with the aim to port Linux on mobile device. Inside Freerunner Linux os already works, we had just to patch the kernel with our modifications and we choose to create hixosfs partition inside a minisd with 2 gb of stored music file.

VI. Evaluation of system's performance

The testing scenario was implemented in a shop and in a music laboratory. Every user was provided a Openmoko[5], a bluetooth mobile, gps and wireless with Linux.

Inside the buildings there were embedded systems with bluetooth antennas for identifying the users and send bluetooth sms.

A. Cinema

The aim of the experiment was sending advertisement sms via bluetooth. The text of the sms was of multimedial type and was addressed to promote objects of the shop itself. Each user was invited to switch on the bluetooth device of his mobile.

The antennas were in communication with the server for sharing of the contents to send and at the same time to check the mac addresses in order to avoid sending the same content for more than once to a given mac address.

The system was handled in remote with a little web based application for managing the synchronization of the bluetooth advertisement campaign with respect to the contents and the hours. Every antenna was sending daily reports to the server and all the data were handled in a hixosfs filesystem. The reports were compressed in files of a hixosfs partition in which for every campaign were registered: message sending, message not accepted, disinterest of the user for the system.

The data of the experiment show that the system must be promoted better. Only 10% of the users were aware of the connection request and only 80% of those accepted the message. The main problem (90%) that came out with a observation of the log files is the fact that people didn't knew the system so not knowing what kind of messages they were receiving, they choose to ignore the request or just not accept.

B. Laboratory

The motivation of the experiment was to share some multimedia files inside the laboratory based on the profile of each user.

With every user there is a Rfid tag associated, an Openmoko mobile with a bluetooth device. In the server there have been loaded the musical preferences of each user. The entrance of the lab is managed by an "intelligent" system that performs an authentication via a Rfid. Inside the lab there is an ad-hoc wireless connection implemented with aodv protocol. When the user is detected by the system, he is profiled and some folders are shared containing musical tracks coherent with the user profile. At the same time every user shares musical tracks with the system. All the musical tracks of the users are not on the server but in the device of each user. Tracks are saved in hixosfs partition of the flash memory where each file has been tagged. The music files were tagged for the multimedia contents and for the profile. Each file shared by the user A with profile P_A has been tagged with the elements of the profile P_A.

The user B, with profile P_B that has at least one element same as in the profile P_A, has the file sharing from the user A. With such approach it was possible the to implement the probably interesting file sharing between users with similar profiles and avoiding probably not interesting file sharing between users with different profiles.

 ISBN: 978-1-61208-100-7

VII. SCRIPTING

The final Linux distribution will contain only the python interpreter and the essential commands without adding programs for database management. The system therefore includes a Linux kernel (2.6.23) patched with hixosfs and a shell in Python through which you can either start the daemons either use hixosfs. Pybox is a shell created to enlighten the work load on the embedded systems. Pybox has been written completely in Python and it occupies less space in the device having interesting performances. The Pybox project implements the rewriting in Python of the basic commands of the system. To use Pybox in our embedded systems we rewrote the commands interfaced with hixosfs.

The system is thus platform independent. Using Python has facilitated the porting of the system on ARM architecture (Openmoko). System startup will mount a partition of the flash memory is partitioned with hixosfs and through openvpn also a remote folder (mounted via NFS) for sharing the system log. Periodically, in fact, the system performs a backup on the remote server.

A. Directory tree

The user process interface has been extended with the two programs *stattag* and *chtag* to access or modify the new inode information for one file. Ad hoc user mode tools have been implemented to extract metadata and populate the whole fs in one step. One example is the command *addbluez* that extract mac address of bluetooth device to fill the inode.

To create a directory tree with tagged files, there is the command *orderby*, with syntax:

```
$ orderby [-mac | -names | -data]
```

The command *scan -bluez* can find inside fs files with a specific content of tags and has the syntax.

```
$ scan -bluez -option [ value | ALL ]
Options are:
-datestart DATE -dateend DATE
-date DATE
-mac MAC
-name NAME
```

For example: where *00, 19, 2D, 14, C9, 93, ...* are folders.

In the foils of the tree (representing each device) have been created files to save the events:

- `to send` - to lock the device for other sending;
- `sent` - to identify the interaction (with positive or negative result) with the device;
- `ping` - to save all the times in which the device is in the scope of the bluetooth antennas;
- `problem` - to save all the problems of each mac address.

```
$ scan -bluez -date all -mac AA:BB:CC:DD:EE:FF
```

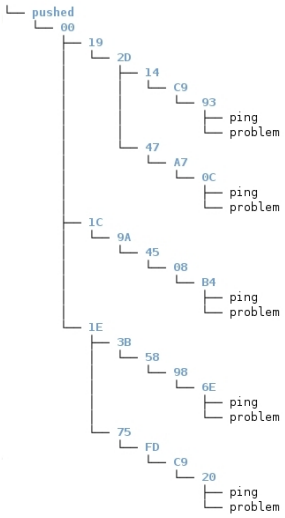


Figure 1. *orderby -mac*

is the command to find all activity of mac AA:BB:CC:DD:EE:FF.

B. Bluetooth

Lightblue [6] is a layer in python for handling bluetooth devices. This layer offers high level functionalities .

For each functionality there is an appropriate error handling. For example in case of denied connection the system returns an error code.

Script for sending a bluetooth sms.
Input mac address, Push channel, path of the file in th system, filename to display in the ending device.

```
import lightblue
def sendsms ( mac , channel , pathfile , nomeoutput ):
        client = lightblue.obex.OBEXClient(mac, channel)
        client.connect()
        putresponse = client.put({"name": nomeoutput},
        file(pathfile, 'rb'))
        client.disconnect()
        return;
```

Script for handling errors caused during bluetooth sms sending. In case of error a file is created with all the information regarding mac address, error and date of the error.

VIII. EXT2 SOLUTION

We have created the same result as hixosfs folder tree inside a simple ext2 fs. We have writen some simple bash scripting to create our data log with a tree of folders and we have created simple bash scripting to select data from this tree.

A. Bash scripting

The particular organization of data in folders allow easy reference data using sample scripts in bash. The following

 ISBN: 978-1-61208-100-7

script allows us to scan the " database" the number of unique devices discovered on a certain day every hour.

The system has a policy to store the log output of this script every day to provide simple statistics for the owner of the shop.

devicexhour.sh

```
echo -e "hour\t device"; data=$1
for hour in {00..24};
do
  echo -en "$hour\t "
  count=$(grep $data-$hour pushed/*/*/*/*/*/*/ping |
      cut -c 1-29 | uniq | wc -l);
  echo $count
done
```

The following script allows us to count how many minutes each device is near of our antenna. With this information you can determine the periods of day when most people stop near the antenna and then inside the shop.

With the next script you can count the number of unique devices to date input. With this script you can establish a trend throughout the year the inflows of people inside the shop.

devicexday.sh

```
count=0; data=$1;
for i in `ls pushed/*/*/*/*/*/*/ping`;
 do
    grep $data  $i> /dev/null;
    if [ $? -eq 0 ]; then let count=count+1;
    fi
done
echo $data = $count
```

This is the most important script inside our antenna-device. This script runs and creates a tree of folders as in Figure 1 like hixosfs without hixosfs.

pushed_no_message.sh

```
while [ true ]
do
contatore=1
numrighe=0
sdptool search OPUSH > ./opush
fallito=`cat opush | tail -1 |
    grep 'Inquiry failed' | wc -l`
if [ $fallito -ne 1 ]; then
{
   cat ./opush >> log/logopush
   date >> log/logopush
   cat opush | grep -B 1 Service | grep Searchin |
     awk '{print $5}' | sort -u > pushmac
  for i in $(cat pushmac); do
    echo -n $i; cat opush |
    grep -A 9 $i | grep Channel; done
    | grep :..: > macechannel
   numrighe=`wc -l macechannel | cut -c 1-2`
}
...
if [ $numrighe -ne 0 ]; then
{
sed -i s/Channel:/\\n/ macechannel
..
for i in $(seq $numrighe)
do
...
mac=${array[$j]}
channel=${array[$i]}
temp_dir=$(echo $mac | sed -e 's/:/\//g')
...
  mkdir -p pushed/$temp_dir
  touch pushed/$temp_dir/ping
  DATA=`date +%d-%m-%y-%H-%M`
guardadata=`grep $DATA pushed/$temp_dir/ping |
  wc -l`
if [ $guardadata -eq 0 ]; then
{
echo $DATA >> pushed/$temp_dir/ping
...
```

IX. Performance measures

The presented testing scenario is similar to the system proposed in [12]. The differences are the extra time to test the product and more detailed statistics.

In this section we present performance measurements made on *hixosfs*. The monitored operation is reading tags from musical file o bluetooth file.

Here we show that the disk performs better for indexing and retrieving music over standard file systems but we didn't study so far the loss of performance respect other uses of the fs. The idea in fact is that hixosfs will be used only for a disk partition containing a musical data or bluetooth device file collection to better organize and query such data and not for the whole disk.

We measured in fact the time required to perform this operations by *hixosfs* and we compared it with the time needed in the case the data are stored in a common fs and accessed by a Sqlite db.

Finally we have compared hixosfs solution with a simple ext2 filesystem managed by simple bash scripting.

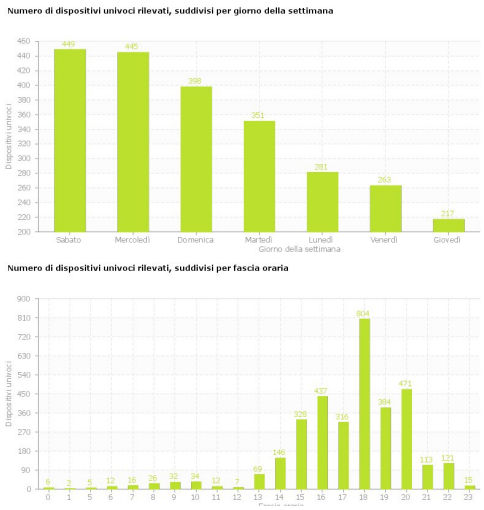


Figure 2. *Unique devices per day and hour*

A. Sqlite vs hixosfs vs ext2

After three months of testing in a cinema, our filesystem contained about 20,000 unique mac address scanned by the system. We have recreated the conditions both for the sqlite database and in an ext2 filesystem to compare systems.

First we compared the search for single mac address in the db.

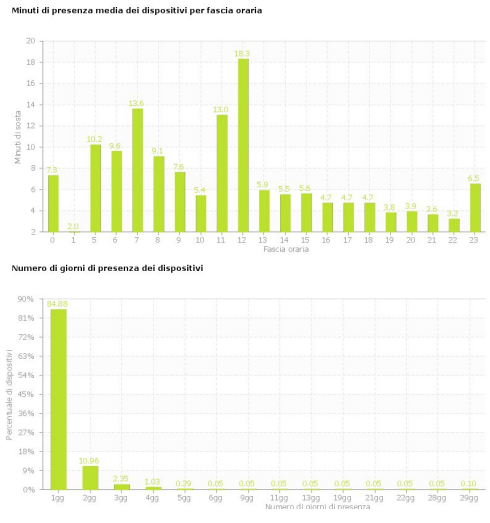


Figure 3. *Average waiting time near antenna*

A simple table `mac` have been created with:

- *id smallint*
- *mac varchar(20)*

We have insert into table 20.000 mac address generated by a casual ad-hoc algorithm. We have created 20.000 files inside a hixosfs partition and we have ordinated (by `orderby` command) a tree of directory to store all informations about every device. We have compared the time to execute "scanbluez" that works with hixosfs and "bashbluez" a simple bash scripting that works with ext2 filesystem.

We have compared time to execute a simple search of one mac address (case A in Table I) .

During testing we had the need to see statistics on the state of the system eg

- the number of devices that have agreed to receive SMS advertising
- the device number that comes back regularly and how often
- problems related to non-submission of the SMS advertisement

Our system provided the files inside the folder which contains for each device problems, and pings sent. We created the database sqlite a second table "sent"

- *id smallint*
- *data varchar(20)*
- *problem varchar(20)*
- *flag smallint;*

We have populated this table with 150,000 records of problems and on random times and we have compared the statistics to extrapolate from the system (Figure 2). (case B in Table I) .

Case\System	sqlite	scanbluez	bashbluez
A	0m0.093s	0m0.004s	0m.010s
B	0m.183s	0m,012s	0m.098s

Table I

X. Conclusion and Future Work

We presented hixosfs as a new solution to manage bluetooth marketing context. We tested this idea in a couple of real context. We improved the tests in [12] applying to cases where there are dependencies between data, when access to data in databases is perhaps the most performant.The system presented use also a platform-indipented code in Python to manage bluetooth connectivity and hixosfs user space program. We compared the idea with the standard approach and we noticed how the approach is more performant. In the future we need to improve the file system, adapting it to a file system lighter than ext2.

References

[1] Corriero, Cozza, Pistillo, and Zhupa. *Wifi Mesh for HandHelds in Linux*, 978-989-8111-60-9, pag 380-383. ICWN 2008.

[2] Corriero and Cozza. *The hixosfs music approach vs common musical file management solutions*, ISBN: 978-989-674-007-8, pag. 189-193. SIGMAP 2009.

[3] Rubini. *The "virtual filesystem" in Linux* . Kernel Korner. http://www.linux.it/∼rubini/docs/vfs/vfs.html.

[4] AODV-uu homepage - Uppsala University, http://www.it.uu.se/research/group/coregroup, 12.2008

[5] Openmoko. http://www.openmoko.org, 03.2010. Home Page.

[6] Bea Lam. LightBlue: a cross-platform Python Bluetooth API. 03.2010. http://lightblue.sourceforge.net/

[7] Card, Ts'o, and Tweedie. *Design and Implementation of the Second Extended Filesystem*, http://e2fsprogs.sourceforge.net/ext2intro.html.

[8] Sqlite. http://www.sqlite.org/. Home Page. 03.2010

[9] Kernel. Torvalds. www.kernel.org. Home Page. 03.2010

[10] Hixosfs. Hixos. www.di.uniba.it/ hixos/hixosfs. Home Page. 03.2009

[11] Corriero and Zhupa, *An embedded filesystem for mobile and ubiquitous multimedia*, MMEDIA2010.

[12] Corriero and Zhupa, *Hixosfs for ubiquitous commerce through bluetooth*, FutureTech2010.

 ISBN: 978-1-61208-100-7

A Mobile Knowledge-based System for On-Board Diagnostics and Car Driving Assistance

Michele Ruta, Floriano Scioscia, Filippo Gramegna, Eugenio Di Sciascio
Politecnico di Bari
via Re David 200
I-70125 Bari,Italy
m.ruta@poliba.it, f.scioscia@poliba.it, gramegna@deemail.poliba.it, disciascio@poliba.it

***Abstract*—*In-vehicle* electronic equipment aims to increase safety, by detecting risk factors and taking/suggesting corrective actions. This paper presents a knowledge-based framework for assisting a driver via her PDA. Car data extracted under On Board Diagnostics (OBD-II) protocol, data acquired from PDA embedded micro-devices and information retrieved from the Web are properly combined: a simple data fusion algorithm has been devised to collect and semantically annotate relevant safety events. Finally, a logic-based matchmaking allows to infer potential risk factors, enabling the system to issue accurate and timely warnings. The proposed approach has been implemented in a prototypical application for the Apple iPhone platform, in order to provide experimental evaluation in real-world test drives for corroborating the approach.**

***Keywords*-Semantic Web; On Board Diagnostics; Ubiquitous Computing; Data Fusion; Intelligent Transportation Systems**

I. INTRODUCTION

The social and economic costs of road accidents are widely acknowledged. Three main factors able to influence their incidence have been identified: to educate drivers to a more careful behavior; to improve road conditions; to enhance features and capabilities of protection devices on vehicles. Evidence shows that investing resources in any of these fields can lead to a decrease in the frequency and severity of car crashes [1].

Modern vehicles are equipped with several Electronic Control Units (ECUs) coordinating and monitoring internal components and devices, communicating over one or more car network buses, such as for example *CAN-Bus* [2]. International standards require new vehicles support the *On Board Diagnostics, version 2* (OBD-II) protocol (http://www.arb.ca.gov/msprog/obdprog/obdprog.htm - last accessed on July 19th, 2010) and be equipped with an OBD-compliant interface to provide direct access to data in the vehicle network. The OBD-II port can be found under the dashboard in the majority of current automobiles. It provides real-time access to a large number of vehicle status parameters. Furthermore, in case of malfunctions, Diagnostic Trouble Code (DTC) values are stored in the car ECU and can be later retrieved by maintenance technicians using proper hardware and software kits. In latest years, access has been granted also to the general public of car enthusiasts by developing OBD-II *Scan Tools*, *i.e.*, cheap electronic devices that bridge the OBD-II port with standard wired (RS-232, USB) or wireless (Bluetooth, IEEE 802.11) computer communication interfaces.

This paper presents a knowledge-based framework for assisting drivers, able to monitor vehicle data extracted via OBD-II and integrating environmental information gathered from external sources in order to detect potential risk factors and to provide warnings and suggestions in real-time. The mobile system we propose allows to process:

- vehicle status data collected from an OBD-II Scan Tool;
- data acquired from embedded smartphone micro-devices, such as GPS (Global Positioning System) and accelerometer;
- optionally, information retrieved from external Web-based data sources, *e.g.*, weather conditions.

Data are collected within short observation intervals and, by means of proper processing and fusion algorithms, the system is able to identify specific high-level events and conditions, based on low-level data streams. Detectable conditions include: vehicle health and safety equipments status; environmental factors (road surface, traffic); driving style. Furthermore, exploiting common Semantic Web techniques and technologies, got events are semantically annotated w.r.t. an ontology modeling factors influencing driving safety. Annotated descriptions undergo a matchmaking process – exploiting non-standard reasoning services [3]– which is able to discover all possible risks referred to current state of the "driver+vehicle+environment" system. The matchmaking outcome is used to suggest the driver actions and behaviors she can adopt in order to minimize perils.

The proposed framework has been implemented in a prototypical mobile software system, using the Apple iPhone smartphone (iPhone Specifications, http://www.apple.com/iphone/specs.html - last accessed on July 19th, 2010) as reference platform. The experimental evaluation has been carried out taking into account several real-world test drives under different conditions. Obtained results prove both feasibility and usefulness of the presented approach.

In the remaining of the paper, after a survey on most

 ISBN: 978-1-61208-100-7

relevant related work in Section II, the proposed framework is described in detail in Section III. Experiments corroborating the approach are presented in Section IV, and finally, conclusion and future work close the paper.

II. Related Work

The basic design scheme for systems using OBD for automobile fault diagnostics is reported in [4]. It consists of three main elements: (i) on-board sensors and fault indicators, built in the vehicle and communicating with the ECU through a bus; (ii) VCI (Vehicle Communication Interface) that bridges the ECU and the computer diagnosis system through a wired or wireless interface leveraging either OBD or CAN-Bus protocols; (iii) diagnostic software, which provides both user interface and connection capabilities toward a remote maintenance center.

Available literature about OBD-based systems for real-time vehicle monitoring and signaling refers to *remote* and *on-board* solutions, respectively. The former follow the basic architectural model introduced in [5], where a system for both on-line vehicle diagnosis and real-time early warning is presented. It acquires GPS coordinates and vehicle OBD DTCs sending them to a Maintenance Center server via GPRS for immediate actions. All the collected data are stored into a database which is scanned by a diagnostics expert system that classifies vehicle status into either *critical* or *non-critical* and generates a rough suggestion advising the maintenance engineer for taking next action. A similar approach can be found in [6].

Our proposal differs from such works because it does not require expert technicians to understand system outputs. Furthermore, in our solution all processing happens in a smartphone application and then it better reflects an on-board framework.

Consider that, though useful for managing vehicle fleets, remote monitoring do not allow a direct driver assistance. To this aim, on-board monitoring prototypes and reporting systems have been developed [7]. They allow the car driver to be informed about relevant vehicle status conditions during trip. Such systems use custom circuitry for the OBD-II-to-computer interface and include several independent devices, communicating through both wired and wireless technologies. Nowadays freeware and commercial software packages are available, allowing to monitor OBD-II vehicle data by using just a smartphone and off-the-shelf Scan Tools. Nevertheless, all existing on-board monitoring systems directly display the acquired low-level data, and they do not provide more user-friendly information. Particularly, no solutions exploiting logic-based techniques for on-line monitoring of driving risks able to meaningfully assist drivers have been yet presented, to the best of our knowledge.

More recently, researchers acknowledged the possibility to exploit the wealth of real-time vehicle data available trough OBD in order to analyze driver behavior [8]. Current efforts aim at using multi-source information fusion to better understand the relationships between driving habits and vehicle performance, as well as to detect risk situations. Nevertheless, in current approaches, the analysis is performed off-line after data gathering, so they are not able to provide driver support in real time.

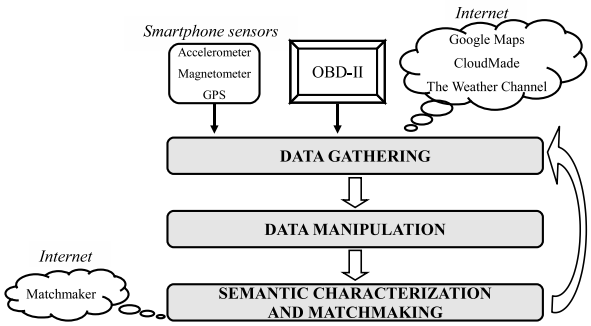


Figure 1. Workflow of the proposed framework

III. Framework

The framework we present includes both architecture and algorithms of a knowledge-based system leveraging the OBD-II car diagnosis and Apple iPhone to monitor environmental conditions, vehicle features, enabled protection equipments and driving style. Through a semantic-based matchmaking the system will be able to evaluate the driving risk level and to suggest how to reduce or even eliminate danger.

A *Kiwi Wifi* PLX wireless adapter (PLX Devices, Kiwi Wifi, http://www.plxkiwi.com/kiwiwifi/hardware.html - last accessed on July 19th, 2010) is exploited for interacting with OBD-II. When turned-on, it builds an ad-hoc network exposing a static IP address allowing an application to communicate with the OBD interface via socket in read/write mode.

As sketched in Figure 1, the proposed approach works along three subsequent stages: (i) data gathering; (ii) data manipulation; (iii) semantic characterization and matchmaking. They are repeatedly executed, within a fixed observation interval (in our case study, a period of 60 seconds was selected). At the end of each data gathering cycle, the data manipulation processing and semantic matchmaking steps are executed and outcomes are displayed to the user on the iPhone screen. In what follows framework details are reported.

A. Data Gathering

At this stage low level data are collected, such as kinetic and vehicle parameters useful to determine driving style, status of safety car equipments, weather conditions, road and traffic information.

The OBD-II interface is used to get data about vehicle performance. OBD-II specifications only comprise the *Physical*

 ISBN: 978-1-61208-100-7

HEADERS			DATA							CRC
H1 TYPE	H2 TARGET	H3 SOURCE	D1 MODE	D2 PID	D3	D4	D5	D6	D7	CHECKSUM

Figure 2. OBD frame common structure

Signal Layer (PSL) and *OBD-II Data Communication Layer* (DCL) w.r.t. the ISO/OSI model. Particularly, PSL outlines hardware characteristics, standard connector (SAE J1962 [9]) conformation and exploited protocols. DCL defines the structure of diagnosis messages exchanged with the ECU, as described in SAE J1979 standard [10]. Request and reply messages have the same conformation, reported in Figure 2. Header bytes H1, H2 and H3 denote priority or message type, destination and sender addresses, respectively. The first data byte D1, namely *mode byte*, indicates the modality to access OBD information. The standard supports 10 modes for diagnostic requests. In particular, *mode 1* is used to obtain current diagnosis data, and it is arguably the most useful mode for our purposes. The second data byte comprises the so-called PID (Parameter IDentification): a value indicating what data is required. The PID also fills the second byte in the corresponding reply packet coming from the vehicle. The remaining data bytes, when used, are reserved for further specification about required data; in a reply message, they are the actual data returned from the vehicle ECU. The last byte is exploited for message error control.

Though the proposed system is able to retrieve all possible vehicle parameters via the OBD-II interface, our case study focuses on vehicle speed (PID $0D_h$) and RPM (PID $0C_h$), which contribute to characterize driving style as well as road traffic. The Apple iPhone (like many currently available high-end smartphones) integrates several micro-devices and offers wireless Internet connectivity through the cellular network. Such capabilities are exploited to collect information about the environment. The GPS receiver provides latitude and longitude coordinates of vehicle current position, which are used for a *reverse geocoding* query using the Google Maps API (Google Maps API Family, http://code.google.com/intl/it-IT/apis/maps/ - last accessed on July 19th, 2010) to get the corresponding location address. Location is further exploited to get weather conditions, using a free service offered by TWC (The Weather Channel: weather XML Data Feed, http://www.weather.com/services/xmloap.html - last accessed on July 19th, 2010) website. For our purposes data concerning weather description (rain, snow, fog, cloudy) and wind speed are exploited. Finally, in order to access roads information, an additional reverse-geocoding operation is implemented using the CloudMade service (http://cloudmade.com/ - last accessed on July 19th, 2010). The reply message, in JSON (JavaScript Object Notation data interchange format, http://www.json.org/ - last accessed on July 19th, 2010) format, contains an indication of road type and speed limit. For what concerns the safety equipments available on the vehicle, the user can enable/disable/check their status exploiting the *Settings* view of the implemented application leveraging the interaction with the car via OBD.

B. Data Manipulation

The final aim of this application phase is to process previously collected data in order to identify conditions and events, which can be annotated w.r.t. a reference ontology (described in Section III-C). Several statistical procedures and data fusion algorithms were devised and tested in order to build a set of (binary or multiple) classifiers for mapping data series to events. Solutions outlined hereafter were selected because they provide adequate sensitivity also maintaining moderate computational and memory requirements.

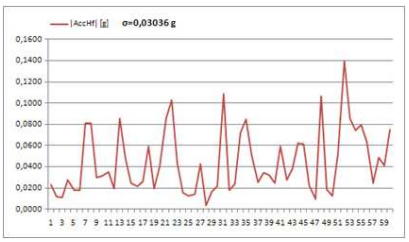
Road conditions. It is possible to distinguish between even and uneven road surface by computing the standard deviation of acceleration values produced by vehicle oscillations. In the previous data gathering step, the three-dimensional acceleration vector is sampled from the iPhone accelerometer at a 2 Hz frequency. Then high-pass filtering is applied, in order to discard components due to gravity and normal vehicle acceleration/deceleration: 15 Hz was found as the optimal cutoff frequency. As shown in Figure 3(a) and 3(d), acceleration values variability is significantly higher on an uneven road surface w.r.t. to an even one. Experimental tests proved an optimal threshold value of $\sigma = 0.020g$ ($g = 9.80665\, m/s^2$) for the standard deviation of acceleration to classify road surface.

Driving style. In order to distinguish between an *imprudent* and a *regular* drive, abrupt speed and direction changes should be detected. The standard deviation of RPM (Revolutions Per Minute) of the vehicle engine –retrieved from OBD-II interface at 1 Hz frequency in the observation period– was selected as discriminatory parameter. An imprudent driving style can be distinguished from an even pace by observing the variability of RPM, as depicted in Figure 3(b) and 3(e). Our experiments proved a threshold value of $\sigma = 400$ RPM provide good reliability in distinguishing the two driving styles.

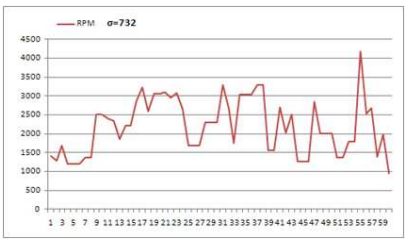
Speed. To characterize vehicle speed, it is sufficient to compute the average speed value for the observation interval. W.r.t. Italian urban speed limits, the threshold value to distinguish a high-speed driving from a low-speed one was set to 40 km/h.

Traffic conditions. OBD parameters are also useful to characterize traffic conditions. In congested traffic situations, a driver usually alternates frequent and fast speedups/stops and downtimes. In terms of RPM, this behavior produces a typical sawtooth waveform –with sharper upward and downward slopes– that alternate with stages at a minimum

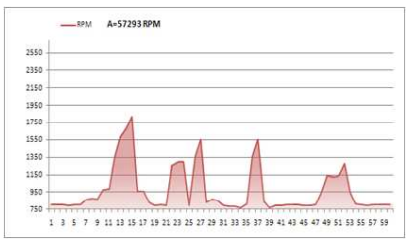
 ISBN: 978-1-61208-100-7



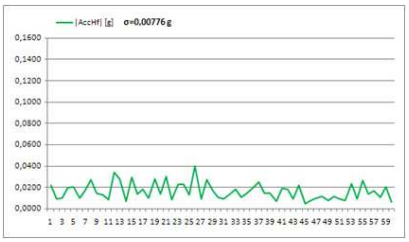
(a) Acceleration values for an uneven road



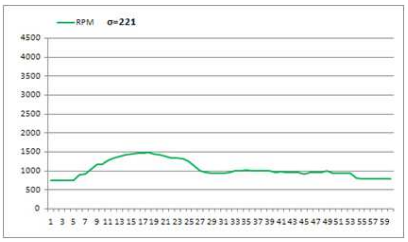
(b) RPM values for an aggressive driving style



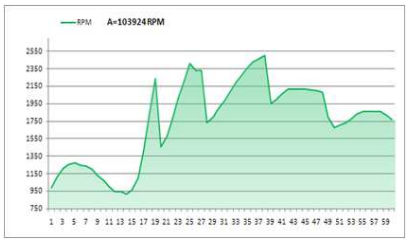
(c) RPM integral area for high density traffic



(d) Acceleration values for an even road



(e) RPM values for an even pace driving style



(f) RPM integral area for low density traffic

Figure 3. Acceleration and RPM values acquired in data gathering phase

value. As depicted in Figure 3(c) and 3(f), the integral area computed in case of traffic congestion is clearly lower than in case of lack of traffic. A threshold value of 65000 (with a data gathering phase of 60 seconds) allows to discern between these two traffic conditions.

Wind. Wind speed is a relevant driving risk factor when it exceeds a specific value. In accordance with commonly exploited *Beaufort scale*, in our framework the threshold value is set to 40 km/h.

C. Semantic Characterization and Matchmaking

The semantic annotation of environmental and driving events closes the context extraction and it prepares the subsequent matchmaking phase. A prototypical ontology modeling the domain of interest has been implemented, using OWL-DL [11] formal language, grounded on Description Logics (DL) semantics. It specifies classes and properties (a.k.a. concepts and roles, respectively) needed to characterize all the events and situations that can be detected by the data gathering and manipulation steps. As the framework will be augmented with new data sources and algorithms to detect more situations, it will be possible to extend the domain ontology accordingly. Consistency checks are performed at each stage of ontology evolution, in order to ensure that new knowledge does not conflict with previously modeled one. In greater detail, the following classes and properties have been defined.

- *Weather* describes weather conditions. It has five subclasses: *Fog*, *Snow*, *Cloudy*, *Rain*, *Clear*.
- *Wind* refers to wind strength. The corresponding subclasses are *Weak_Wind* and *Strong_Wind*.
- *Road_Surface* represents road conditions. Two different kinds were modeled, through subclasses *Uneven_Road* and *Even_Road*.
- *Road_Condition* models the kind of road. The corresponding subclasses are *High_Speed_Road* and *Low_Speed_Road*.
- *Traffic* is related to traffic conditions. The subclasses are *High_Density_Traffic* and *Low_Density_Traffic*.
- *Driving_Style* refers to user driving style, with subclasses *Even_Pace_Style* and *Imprudent_Style*.
- *Vehicle_Speed* describes the vehicle speed, whose subclasses are *High_Speed* and *Low_Speed*.
- *Safety_Equipment* represents protection devices that may be available on a vehicle. In particular, the following subclasses were modeled: *Fog_Lamp*, *ABS*, *ESP* and *Snow_Chains*.
- *Vehicle* is related to the car. It is involved in the following property relations: *hasDriving_Style* with *Driving_Style* class, *hasSpeed* with *Vehicle_Speed* class and *hasSafety_Equipment* with class *Safety_Equipment*.

The first five classes model the environment, whereas the remaining ones are used to describe both vehicle and user driving style. It is important to note that the ontology is not just a taxonomy, since definition and inclusion axioms are used. Particularly, the semantic description of environmental conditions focuses on potential risks they can cause to driver, on the required vehicle equipment and driving style able to minimize the risk. For example, let us consider the following semantic annotation in DL formalism: $Fog \equiv Weather \sqcap \forall\, hasSafety\_Equipment.(Fog\_Lamp \sqcap ABS) \sqcap \forall\, hasSpeed.Low\_Speed \sqcap \forall\, hasDriving\_Style.Even\_Pace\_Style.$ It means that, in order to avoid risks produced by *Fog*, the vehicle must be equipped with fog lamp and ABS and the user must adopt an even pace driving style with low speed.

Semantic matchmaking process exploits MaMaS-TNG (MatchMaking Service-The Next Generation, available as

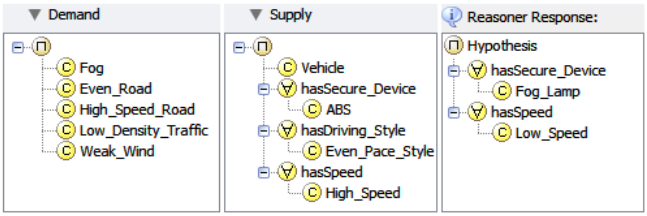


Figure 4. Abduction query example

an HTTP service at: http://dee227.poliba.it:8080/MAMAS-tng/DIG) matchmaker to infer possible risks for the user, warning her to avoid danger, given the context. The *Concept Abduction* non-standard inference service was selected to perform this task. Due to space limitations, the reader is referred to [3] for a thorough presentation of the Concept Abduction Problem (CAP). In a nutshell, given a request R and an available resource/service S, described w.r.t. a common ontology, Concept Abduction can be used to identify what is missing in S in order to completely satisfy R. In our framework, the context semantic annotation represents the request (*i.e.*, what requirements are needed to travel safely), while the semantic description of vehicle and user driving style model resources availability (*i.e.*, what is provided by the "vehicle+driver" system). In this way, the abduction process will infer the safety requirements that are not explicitly satisfied by current vehicle configuration and driver behavior, thus providing proper suggestions to the driver so that every kind of danger is prevented.

Figure 4 shows an example of abduction query and results with reference to the previous very small example. As the reasoner reply underlines, the presence of *Fog* concept in the request implies the need of both fog lamps and ABS, as well as an even pace driving style with low speed. On the other hand, as the vehicle only offers ABS and the driver has an even pace driving style with high speed, the abduction outcome suggests the driver to activate fog lamps and adopt low speed to attenuate risks.

IV. Tests and Experimental Results

Effectiveness of the devised framework and usability of the mobile application were tested in three different environmental conditions described in Table I, using a Mercedes C220 CDI registered in 2003. For each scenario, two test drives were performed with different speed, driving style and safety equipments settings, as described in Table II. Video recordings of tests are available at http://sisinflab.poliba.it/idrivesafe/. It is possible to see that, under good cellular network coverage, system performance is adequate to grant a satisfactory user experience.

The first example scenario is featured by an uneven and low speed road. In Figure 5 a screenshot of system outcome is shown. Part (a) shows information about weather, traffic and road description. Information about vehicle speed and driving style are reported in part (b). Finally, part (c) will

Table I
Context Description

	Test 1	Test 2	Test 3
Location	Grumo Appula(BA)	Bari	Toritto(BA)
Road	SP. 71	C.so V. Emanuele II	SP. 1
Weather	Cloudy	Clear	Rain
Wind	Weak	Weak	Strong
Road type	Low Speed	Low Speed	Low Speed
Road surface	Uneven	Even	Uneven
Traffic	Low density	High Density	Low Density

Table II
Test Settings

	Speed	Driving Style	Safety Equipments
Test 1, Setting 1	Low Speed	Even Pace	ABS, ESP, Fog Lamp
Test 1, Setting 2	High Speed	Imprudent	None
Test 2, Setting 1	Low Speed	Even Pace	ABS, ESP, Fog Lamp
Test 2, Setting 2	Low Speed	Imprudent	None
Test 3, Setting 1	Low Speed	Even Pace	ABS, ESP, Fog Lamp
Test 3, Setting 2	High Speed	Even Pace	None

contain system suggestions to the user. In the first case, the system detects a risk-free situation. Although the road is uneven and imposes a low speed, the driver is adopting a driving style suitable for these conditions and the car is equipped with the needed safety equipments. In the second case, the system detects a dangerous situation, due to an imprudent driving style, a high speed (inappropriate for given road conditions) and the lack of ABS and ESP (strictly needed on an uneven road). To reduce such risk factors the system suggests the driver to moderate her driving style, to reduce speed and to activate the required safety devices.

The second test was performed in high-density traffic conditions on a low-speed road. As depicted in Figure 6, the first configuration is not dangerous for the user. Although the traffic is intense, the driver adopts an even pace and the car is featured by the needed safety equipment. In the second case, instead, the system detects risk factors, due to an imprudent driving style (absolutely not suitable in high-density traffic) and lack of ABS. Hence, the system suggests to drive with caution and to activate ABS (if possible).

The last test refers to adverse climatic conditions with rain and high wind. Figure 7 shows system outcomes. In the first case no risk for the user is detected. Notwithstanding adverse weather conditions, the driving style is proper and

Figure 5. System outcome in Test 1

Figure 6. System outcome in Test 2

Figure 7. System outcome in Test 3

the car is safe (ABS and ESP are turned on). In the second configuration, on the contrary, the system reveals a danger, deriving from high speed (very risky in case of rain, high wind and narrow roads) and from the lack of ABS and ESP protection. So the system suggests the user to activate ABS, ESP and to reduce the speed.

V. Conclusion and Future Work

We have presented a knowledge-based framework and a prototypical system for real-time driving assistance. They refer to every OBD-based vehicle and comply with several driving context without the need of learning stages. By means of information extracted trough the accelerometer and GPS embedded in an Apple iPhone PDA and exploiting Web-based available services, a context annotation is performed. It enables semantic-based inferences which finally provide useful recommendations for driving safely. Experimental evaluations evidenced that the system is able to detect a variety of road and traffic conditions, as well as driving behavior, issuing accurate suggestions to minimize risk factors.

Future work includes enhancements to the mobile prototype, such as voice alerts and the local integration of the inference engine. The mobile matchmaker devised in [12] will be ported to the target smartphone platform to remove the dependency on centralized reasoners and to reduce bandwidth usage. An extensive experimental campaign is also under planning, to evaluate system performance in each workflow stage. As far as research is concerned, more OBD parameters and smartphone peripherals (*e.g.*, camera, microphone) could be used, in order to detect and feature a larger array of contexts.

Acknowledgments

The authors acknowledge partial support of Apulia Region Strategic Project PS_121 - Telecommunication Facilities and Wireless Sensor Networks in Emergency Management.

References

[1] L. Evans, *Traffic safety and the driver*. Van Nostrand Reinhold, 1991.

[2] BOSCH, *CAN Specification*, 2nd ed., ROBERT BOSCH GmbH, 1991.

[3] S. Colucci, T. Di Noia, A. Pinto, A. Ragone, M. Ruta, and E. Tinelli, "A non-monotonic approach to semantic matchmaking and request refinement in e-marketplaces," *International Journal of Electronic Commerce*, vol. 12, no. 2, pp. 127–154, 2007.

[4] J. Hu, F. Yan, J. Tian, P. Wang, and K. Cao, "Developing PC-Based Automobile Diagnostic System Based on OBD System," in *Power and Energy Engineering Conference (APPEEC), 2010 Asia-Pacific*, March 2010, pp. 1 –5.

[5] C. Lin, C. C. Li, S. H. Yang, S. H. Lin, and C. Y. Lin, "Development of On-Line Diagnostics and Real Time Early Warning System for Vehicles," in *Sensors for Industry Conference, 2005*, 8-10 2005, pp. 45 –51.

[6] J. Lin, S. C. Chen, Y. T. Shin, and S. H. Chen, "A Study on Remote On-Line Diagnostic System for Vehicles by Integrating the Technology of OBD, GPS, and 3G," in *World Academy of Science, Engineering and Technology, 2009*, aug. 2009, pp. 435–441.

[7] Y. Chen, Z. Xiang, W. Jian, and W. Jiang, "Design and implementation of multi-source vehicular information monitoring system in real time," in *Automation and Logistics, 2009. ICAL '09. IEEE International Conference on*, aug. 2009, pp. 1771 –1775.

[8] S. Choi, J. Kim, D. Kwak, P. Angkititrakul, and J. Hansen, "Analysis and Classification of Driver Behavior using In-Vehicle CAN-Bus Information," in *Biennial Workshop on DSP for In-Vehicle and Mobile Systems*, June 2007.

[9] SAE Standard J1962, "Diagnostic Connector - Equivalent to ISO/DIS 15031-3," 2002.

[10] SAE Standard J1979, "E/E Diagnostic Test Modes - Equivalent to ISO/DIS 15031-5," 2002.

[11] W3C Recommendation, "OWL Web Ontology Language," 2004, http://www.w3.org/TR/owl-features/ last accessed on July 19th, 2010.

[12] M. Ruta, F. Scioscia, and E. Di Sciascio, "Mobile Semantic-Based Matchmaking: A Fuzzy DL Approach," in *7th Extended Semantic Web Conference (ESWC2010)*, ser. Lecture Notes in Computer Science, vol. 6088. Springer, 2010, pp. 16–30.

 ISBN: 978-1-61208-100-7

Wireless service developing for ubiquitous computing environments using J2ME technologies

José Miguel Rubio
Escuela de Ingeniería Informática
Facultad de Ingeniería, PUCV
Valparaíso, Chile
jose.rubio.l@ucv.cl

Claudio Cubillos
Escuela de Ingeniería Informática
Facultad de Ingeniería, PUCV
Valparaíso, Chile
claudio.cubillos@ucv.cl

***Abstract*—From some years ago, technologies are ubiquitous, omnipresent and integrated seamlessly in our common activities. There is a need for developing services and applications that use all the capabilities of mobile devices available, for connectivity and data processing. Java 2 Micro Edition (J2ME) technology develops these services for use in a wide range of devices, with portable code, and compatible with several classes of hardware and software. Wireless communications technologies (WiFi, Bluetooth) allow devices to communicate with their environment. Bluetooth technologies are becoming massive, and are present in the majority of these devices. This kind of connectivity is possible using the J2ME Application Programming Interfaces (APIs), but it requires some expertise and a wide knowledge in order to implement these components. This document proposes a framework focused towards the service development using bluetooth technologies with more simplicity. Also, it presents the design and implementation of a mobile information service that uses this framework in a ubiquitous environment.**

***Keywords*-ubiquitous computing; wireless services; service discovering; bluetooth; mobility.**

I. INTRODUCTION

Using technology in daily tasks today is a very common aspect of our lives, leaving behind the fact that we interact with several classes of devices in every activity we do. This was the vision of Mark Weiser [1], 18 years ago. He wrote about environments that integrate computing with common tasks, allowing the user to interact in natural ways making computers nearly imperceptible. He designated this integration as Ubiquitous Computing, talking about the ease of access and availability of these technologies. This vision is materialized in the proliferation of a wide range of devices (mobile phones, Personal Digital Assistants: PDAs) and different connectivity options, (Short Message Service: SMS, Bluetooth, GPRS: General Packet Radio Service, EDGE: Enhanced Data rates for GSM of Evolution), giving multiple interaction methods. In order to integrate these devices with our common tasks in a more profitable way, there are several technologies centralized in building applications and services for this class of devices. One of them is the J2ME technology.

The next sections describe a simplified approach for using J2ME technologies available for mobile device development. A framework is proposed with components and development directions for this class of devices. Different practical applications are proposed. Finally, a development of a mobile information service is exposed, based on this framework, mixing wireless and web technologies to access information in a localized context.

II. THE J2ME PLATFORM

J2ME is a technology developed by Sun Microsystems to build applications for mobile devices. It is different from other major editions of Java (Java 2 Standard Edition (J2SE), Java 2 Enterprise Edition (J2EE)); basically in a smaller group of functionality that fits the computing, memory and data storing capabilities of limited devices.

A. Configurations and Profiles

To make a difference between different levels of capacities, there are some configurations (minimal technological requirements) and profiles (basic functionality and services) to differentiate between device classes. A common configuration is CLDC (Connected Limited Device Configuration), with minimal requirements of memory, power consumption, and some kind of wireless connectivity, generally HTTP (Hypertext Transfer Protocol) and SMS [2]. The common profile adopted by mobile devices is MIDP (Mobile Information Device Profile), that defines some required functionality about user interface, data persistency, application lifecycle and multimedia control [3].

B. User Interface

The user interface in J2ME offers a limited group of components, based on limitations of screen size and reduced keyboard functions of mobile devices. The user interface is limited to selection lists, basic forms, text input dialogs, and low-level screen painting (Canvas). The user inputs are command buttons and an alphanumerical keyboard.

 ISBN: 978-1-61208-100-7

C. Connectivity

Different connectivity options have a centralized access through Generic Connection Framework (GCF) APIs. GCF allows connecting with any connection available (HTTP, SMS, Bluetooth, local file system) using a unified API. Each connection implemented is based on a generic connection class, and extended to support specific functions related to each particular connection implementation.

D. Optional Packages

To provide additional functions over the basic configurations and profiles, there are some optional packages that access extended functionality, standardized in different Java Specification Request (JSR). Table I lists typical optional packages for mobile devices.

Table I
SOME OPTIONAL PACKAGES FOR J2ME

Specification	Description
JSR-75	File system access and PIM
JSR-82	Bluetooth
JSR-120	Wireless Messaging (SMS)
JSR-135	Multimedia API
JSR-172	Web Services

E. J2ME-Enabled Devices

J2ME is the most available platform in the market. There are several mobile devices today that allow executing J2ME applications, basically mobile phones, smartphones, and PDAs. In Windows Mobile (Pocket PC) systems a commercial J2ME suite must be installed.

F. Application and Service Development in J2ME

There is a variety of tools and programming environments that support the complete mobile development cycle: design, coding, compilation, preverification, packaging, tests, and deployment. The main suite for J2ME programming is the Wireless Toolkit from Sun Microsystems. Programming environments like Netbeans and Eclipse, integrate this toolkit to allow developing J2ME applications. Emulation utilities allow verifying applications for different device models. Deploying applications to mobile devices are performed through direct transfer (infrared, bluetooth) or a web download. Usually, a JAR file (Java ARchive) is provided with all the application components and resources. In some cases, a JAD file (Java Application Descriptor) is required to simplify the install process.

III. J2ME CONNECTIVITY

Mobile devices have a lot of connectivity options, even more than some major device classes. This is one of the main reasons for developing applications for J2ME-enabled devices. A typical mobile phone can be converted into a client's email, web browser, game console, multimedia player and even run business applications and perform online commercial transactions.

A. HTTP/HTTPS Connectivity

In a typical J2ME device, at least HTTP connectivity is required. The majority of online services are web-based. The current increase of web technologies allows access to a variety of contents (information, images, audio/video). Some mobile devices cannot support all of them due to it's limitations in processing, memory or screen size. Many online services (email, social networks) have a "mobile" version or supply a J2ME client application.

B. Bluetooth Connectivity

Bluetooth is a wireless communication technology oriented to connect different classes of device in a reduced area (10-100mts), using low power consumption and a minor complexity compared with other wireless technologies like WiFi. Specified initially for the Bluetooth Special Interest Group (SIG), composed by several industry leaders. The current specification adopted its version 2.1 [4]. From their first release, it has become a standard industry for wireless transfer and personal area networks (PAN). The majority of new mobile devices have included this technology, allowing data transfers, network connections, use of headsets and other wireless peripherals [6].

The bluetooth specification allows inter-operating between independent systems, defining messages for each layer of the protocol stack (high and low level), and between internal subsystems (controller and host). To ensure interoperability between applications some bluetooth profiles are defined to standardize mechanisms and protocols for different applications like file transfer, networking, printing, or synchronization. Each profile defines messages, procedures and rules for the most common services. The Generic Access Profile (GAP) is the basis for all other profiles. Figure 1 shows the hierarchy of basic bluetooth profiles [5].

Several mobile phones and PDAs offer standard bluetooth services and integrate these services with personal information management (PIM) utilities like contact management, tasks and calendar events. There are many potential uses for taking advantage of those functions, developing custom services to allow users to interact with their environment using these devices.

C. JSR-82: Java APIs for Bluetooth Wireless Technology (JAWBT)

J2ME platform offers an additional API for accessing the bluetooth technology: the JSR-82 optional package. Motorola defined the original specification in 2002. Based on the core bluetooth specification, their objective is *to define a standard set of APIs that will enable an open, third-party*

 ISBN: 978-1-61208-100-7

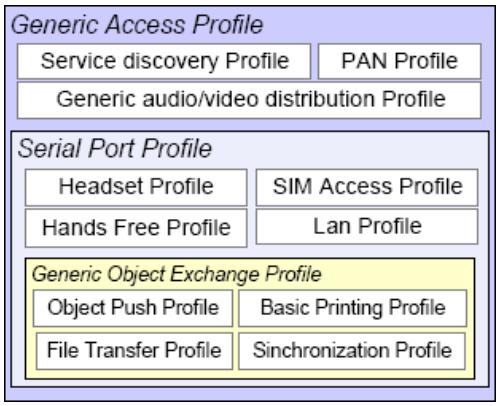


Figure 1. Main Hierarchy of Bluetooth Profiles

application development environment for Bluetooth wireless technology [8].

JAWBT defines a set of components to represent two aspects of the bluetooth specification: core elements and the Generic Object Exchange Profile (GOEP, also known as OBEX). The javax.bluetooth package contains interfaces and classes that represent the core specification: devices, data structures, service discovering, including the L2CAP and SPP protocol. The javax.obex package is another independent set of elements that meets the specification for OBEX connections: header sets, authentification, sessions and operations for client/server connections [5].

IV. Framework for Bluetooth-Based Service Development

The main motivation behind this framework is the high potential of bluetooth technology as a mean for offering extensive functionality in ubiquitous environments, using common devices, standard protocols and data formats, ensuring high interoperability in this class of environments. This section shows the proposed framework for development of bluetooth-based services. The main goal is to provide ready-to-use components in applications that need bluetooth connectivity.

A. Bluetooth Service Development

The JSR-82 API offer the core components for accessing bluetooth technology, nevertheless, this does not provide any direct implementation of common bluetooth profiles. Each developer must implement standard or custom services based on these components. To develop bluetooth based services it is required some knowledge of the technology, and API specifications. There are a few public projects that offer these kind of components, but examples for coding bluetooth services are basic and limited in functionality, and do not provide full implementation of some basic and useful services, leaving the implementation of these class of services out of scope for some projects that would need it.

The goal is to define a common framework for direct access of common bluetooth services, to quickly develop any kind of bluetooth applications. The initial approach is to get a set of components that implement basic bluetooth profiles (Serial Port Profile, Object Push Profile), giving the chance to extend and create new custom services based on the core services. Also, it is defined for some basic directions using patterns and rules to make stable and optimal services. The final goal is to centralize the development efforts in the main application, abstracting bluetooth detailed elements. All components are based in the J2ME APIs: Generic Connection Framework and Bluetooth.

B. Framework Components

The main areas of functionality that include this framework are the following:

1) Bluetooth constants and data: Direct access to main constants, attributes and UUIDs (Unique identifiers for known services and protocols) defined by the bluetooth specification.

2) Service publishing: Defining and publishing standard and custom services.

3) Service discovery: To enable discovering and selection of remote services, access to attributes and properties and to generate the service connection.

4) Bluetooth service connection: To provide the basic implementation of service connection and for init custom data transfers.

5) Logging and debugging: To help developers keep track of the testing executions. Logging capacities for debugging are implemented.

All components are based in J2ME APIs: javax.microedition.io, javax.bluetooth and javax.obex. The main package (btlibrary.core) includes the only core components. Additional packages add specific implementations.

C. Implementation

The core library components implement all basic functionality areas, providing the basis for implementing or extending additional elements. Figure 2 shows the core classes and interfaces. To show some examples of the main components and their logic, the following sections describe the main relation and dependencies of these components, in typical tasks like service publishing, service discovery and client connections.

1) Service publishing: To publish a bluetooth service it is required the following tasks:

- Setup the device as visible (discoverable).
- Identify the service (name, UUID, attributes).
- Wait for connections.
- Process each incoming connection.
- End the service.

 ISBN: 978-1-61208-100-7

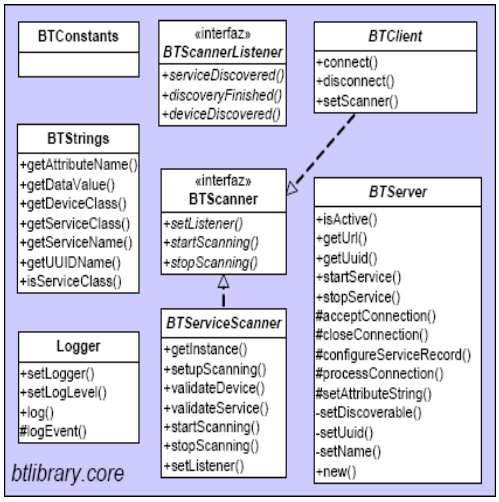


Figure 2. Classes and components of btlibrary.core

The class that implements this functionality is BTServer. It provides methods for creating, identifying, starting or stopping, and processes incoming connections, abstracting the internal logic for these tasks. Figure 2 shows its main methods.

2) Service discovery: The main tasks to discover remote services are:

- Setup the discovery search pattern (the service UUID and attributes required).
- Start the device discovery (or get a cached list of previously known devices).
- Start a service discovery for selected devices.
- Select a service discovered and get the service address.

The class that implements this process is BTServiceScanner. This class notifies applications while new devices and services are found, through the BTScannerListener interface. Figure 3 shows some relations and messages to accomplish this task, and the JSR-82 API calls involved.

3) Connecting bluetooth services: As a result of discovery service tasks available service addresses are obtained. To connect some some of these services we need the service address. BTClient class implements this functionality, including the discovery functions, allowing connection and use of the service operations available.

As an example, a bluetooth file transfer profile service (FTP OBEX) offers functions for browsing remote folders, sending, downloading or deleting files. The ObexFtpClient class, based on BTClient, implements these functions. This bluetooth client offers access to file transfer functions using simple calls, abstracting the protocol and specification logic. Figure 4 shows client operations (connection, file listing and transfer) and API calls involved in relation to bluetooth APIs.

4) User interface definitions: An additional topic is a definition of a generic user interface for use at interactive

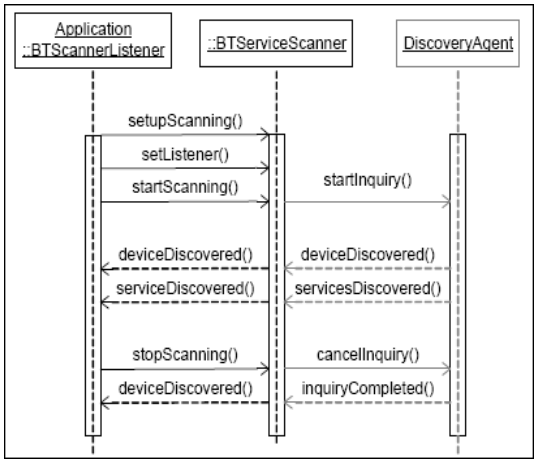


Figure 3. API calls involved in service discovery.

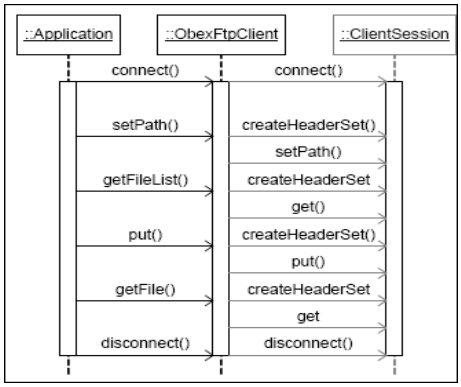


Figure 4. Message sequences in a OBEX FTP service.

service discovery applications, in order to provide a common GUI to control this process, to show discovery progress and status. A graphical service selection interface must provide the following functions:

- To show/hide the selection dialog.
- Show available devices/services while they are discovered (or previously discovered).
- To control the service discovery process (canceling, restarting).
- To select a device from the dialog list.
- To close the dialog and cancel the selection.

The formal definition is implemented in the ConnectionDialog interface. Additionally, there has been implemented a specific dialog for J2ME user interface (ConnectionDialogME), that looks like the model shown in Figure 5.

5) Design Patterns: In order to make an easier service design and application integration, there should be patterns to develop bluetooth services and clients to help the process to be simpler and faster. In general, there are simple steps to accomplish in order to get a successful development, adding basic and advanced examples in order to guide developers.

 ISBN: 978-1-61208-100-7

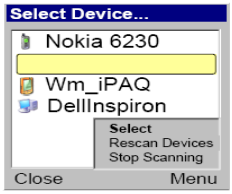


Figure 5. Service discovery connection dialog for J2ME.

As an example, the service creation and publishing process needs to consider some basic patterns and directions:

- To define the base/reference bluetooth profile in order to build the service: generally, a serial port profile (SPP) or an object exchange (OBEX) profile.
- To define the service identifier (UUID). If it is planned to implement some standard bluetooth service, it must assign the specific UUID designated.
- For standard bluetooth profiles or services, it must follow all the official requirements defined to ensure interoperability.
- For custom services, the primary task is to define the public interface, constants and data structures used to exchange messages.
- To optimize connection resources and bandwidth, data transfers must try to use the maximum packet length available, avoiding small data transmissions.
- Due to limitations in the number of concurrent connections, timeout mechanisms are required to avoid blocking new connections when some previous connection is lost or is still awaiting data.

V. Application Fields

Bluetooth services development can have multiple applications. There are several projects related to bluetooth technologies and ubiquitous environments. All of them could be simplified using the components of this framework. Some typical application domains are presented in next sections.

A. Localized Information Services

An information service can provide or process data using the user location as a context. For example a project that allows a museum visitant to receive instant information (text, images, audio/video) about paintings or other things located near the current location or room using his mobile phone with a J2ME application via bluetooth. The key aspect for success is to identify the users and manage profiles, to customize messages and the users experience.

B. Local Environment Interaction

Users can interact with the local environment when some bluetooth-enabled devices are located near the user. For example: a teacher enters the classroom and uses his mobile phone to transfer the class slides and then starts and controls the presentation using a bluetooth service running in a local computer connected to a media projector.

C. Machine to Machine Communication (M2M)

Due to high interoperability and compatibility of bluetooth protocols and profiles, it is possible to implement several classes of device-to-device communication: synchronization, data exchange, collaborative work and distributed applications.

D. Communications and Networks

In this area we may find extended applications to add new communication possibilities in a localized environment. Chat or instant messaging applications, networked games, Internet access for email, news, web browsers, database access, remote access, audio/video streaming. Many internet-based services are possible using a bluetooth service as a gateway.

VI. Practical Application in a Mobile Information System

As a demonstration of use of this framework in a real application for an ubiquitous environment, there has been implemented a mobile information system for academicals and communicational purposes. This system provides the following functions to teachers and students:

- To publish general or group announcements (for teachers, students, class groups, student groups, workgroups)
- To publish calendar events to achieve best attendance to key activities and to complete important tasks.
- To send and receive private messages.
- To store a global directory to access contact information from users (phone, email).
- To associate users with their mobile devices, in order to receive notifications directly.
- To register user profiles and groups, ensuring security and customized information access.

As the user interface for the service, there are some alternatives:

- A web interface accessible via common web browsers. With XHTML compliance allow mobile web browsers to access the portal content in a simple format [7].
- A J2ME client application installed on mobile devices to access portal content with a simplified browser with http connection using a local bluetooth service acting as gateway, giving some extra bluetooth-based features.
- A message interface using a local bluetooth service that receives commands to process and trigger some type of action, like changing user configurations or requesting contacts.

A. Service Architecture

There are three main areas of implementation for this service. Figure 6 shows the schematic composition of the complete service.

 ISBN: 978-1-61208-100-7

1) Data Server: A typical HTTP server is the web content provider and the data interface for some service components. A database server is the global data container. Only one computer acts as a data server.

2) Bluetooth Service: The bluetooth service is composed by different minor bluetooth services and processes for specific tasks: a HTTP gateway service for processing web requests and responses; a notification agent that discovers nearby devices, retrieves messages, news or events associated to the owner's device and sends only new notifications to each discovered device; a command reception service reads messages sent by mobile devices to request information, update profile properties and obtains objects like contacts, applications or events. There can be multiple instances of this bluetooth service installed in any machine connected to the Internet and with a bluetooth installed device.

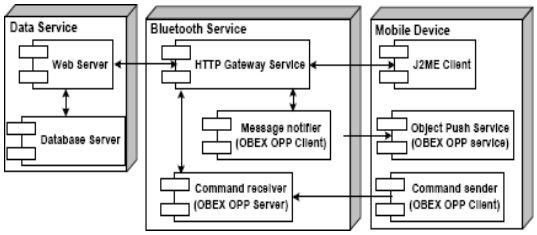


Figure 6. Mobile information service architecture.

3) Mobile Device: It is the main user interface for this service. It can get notifications automatically when it enters in a bluetooth service area, using the internal object push service activated in each device. Generally, this service integrates the reception of special objects with some "organizer" functions. When it receives some contact, event or task they are automatically saved or scheduled.

Also it allows sending commands to any bluetooth service to request or set information, seeking online contacts, messages and resources and receives via a bluetooth the transfer of the required content to get the J2ME client application and update the contact list.

Finally, the J2ME client application installed in this device can access the web site using the http gateway provided by any bluetooth service in range, allowing direct access to the web content in a simplified interface and providing additional mobile and bluetooth related functions directly like trigger phone calls or download content.

VII. Conclusions

J2ME technology allows developing services for a wide range of devices. Bluetooth connectivity is highly supported in newer devices and can interoperate with a variety of device classes due to industry-adopted standard protocols and formats. JSR-82 API is the access point to bluetooth technology in J2ME. Bluetooth service development through this API is a complex task.

This framework can simplify the service development providing direct access to common application requirements in bluetooth-enabled applications: service discovery, publishing and connecting to remote bluetooth services. There are many possible applications of this technology for ubiquitous computing environments: localized information services, environment interaction, machine-to-machine communication and online and network applications.

Through developing a mobile information service based on this framework shows one of many applications of bluetooth technologies to common tasks. In this service we have applied all aspects covered by the framework: creating, discovering and connecting to standard and custom bluetooth services. While in the design and implementation phases some elements of this framework were improved, completed and fixed. The final set of functionality provided by the framework has been enough to start a proper bluetooth service development. The future tasks are: optimizations for several processes like service discovery, bandwidth usage, multiple service synchronization, in order to simplify the code to reduce library size and timeout strategies.

VIII. Acknowledgement

This work has been partially funded by CONICYT through Fondecyt Project No. 11080284 and the Pontificia Universidad Católica de Valparaíso (www.pucv.cl) through Nucleus Project No. 037.215/2008 "Collaborative Systems".

References

[1] M. Weiser, "The Computer of the 21th Century", Scientific American, sept. 1991, pp. 94-100.

[2] Sun Microsystems, "JSR 139: Connected Limited Device Configuration Specification", Version 1.1, 2003, from http://jcp.org/en/jsr/detail?id=139 [accessed, May 3, 2010].

[3] Sun Microsystems, "JSR 271: Mobile Information Device Profile (MIDP)", 2005, from http://jcp.org/en/jsr/detail?id=271 [accessed, May 3, 2010].

[4] Bluetooth SIG, "Bluetooth Specification", Version 2.1 +EDR, The Bluetooth Special Interest Group, 2007, from http://www.bluetooth.com/SiteCollectionDocuments/Core_V21__EDR.zip [accessed, May 3, 2010]

[5] Bluetooth SIG, "Bluetooth Specification 1.1: Part K:10 Generic Object Exchange Profile (GOEP)", 2001, pp. 310-338 from http://www.bluetooth.com/SiteCollectionDocuments/GOEP_SPEC_V12.pdf [accessed, May 3, 2010]

[6] SO Sullivan, "JSR82: Past, Present and Future for Java/Bluetooth APIs", Rococo Software, 2007, from http://www.rococosoft.com/weblog/archives/java_bluetooth/jsr82_tips_techniques/ [accessed, May 3, 2010]

[7] W3C, "XHTML Basic 1.1", W3C Recommendation, 2008, from http://www.w3.org/TR/xhtml-basic/ [accessed, May 3, 2010].

[8] C. Bala Kumar, PJ Kline and TJ Thompson, "Bluetooth Application Programming with the Java APIs", Elsevier, 2004.

 ISBN: 978-1-61208-100-7

Generating Modest High-Level Ontology Libraries for Smart-M3

Dmitry G. Korzun, Alexandr A. Lomov, Pavel I. Vanag
Department of Computer Science
Petrozavodsk State University, PetrSU
Petrozavodsk, Russia
Email: {dkorzun, lomov, vanag}@cs.karelia.ru

Sergey I. Balandin, Jukka Honkola
Nokia Research Center
Nokia
Helsinki, Finland
Email: {sergey.balandin, jukka.honkola}@nokia.com

***Abstract*—Web ontology language (OWL) allows structuring smart space content in high-level terms of classes, relations between them, and their properties. In Smart-M3, a semantic information broker (SIB) maintains the smart space in low-level terms of triples, based on resource description framework (RDF). This paper describes SmartSlog, our solution for constructing Smart-M3 knowledge processors (KPs) that consume/produce smart space content according to high-level ontology terms. The solution is based on the code generation approach. Given an OWL ontology description, the SmartSlog generator maps OWL to the ontology library. It provides 1) API to communicate with SIB and 2) data structures to represent in KP code all ontology classes, relations, properties, and individuals. As a result, the developer easier constructs the KP code, thinking in high-level ontology terms instead of low-level RDF triples. SmartSlog is oriented to ubiquitous systems; the library is modest to the device capacity; it is written in ANSI C, supports even small embedded devices with restricted performance, and allows interoperable applications.**

***Keywords*-Smart spaces; Smart-M3; OWL/RDF ontology; code generator; knowledge processor; low-performance devices**

I. INTRODUCTION

A smart space is a virtual, service-centric, multi-user, multi-device, dynamic interaction environment that applies a shared view of resources [1], [2]. Information conforms to ontological representation with subject–relation(predicate)–object triples as in semantic web [3]. Triples are represented using Resource Description Framework (RDF). A number of devices may access information via semantic information brokers (SIBs), which also support information reasoning.

A client application consists of one or more knowledge processors (KPs) running on various user's devices (Figure 1). KPs act cooperatively forming a publish/subscribe system [4]. Each KP can be thought as an agent using the smart space as a shared knowledge space. The KPs produce (insert, update, remove) and/or consume (query, subscribe, unsubscribe) information in a smart space. The smart space access protocol (SSAP) implements the SIB ↔ KP communication, using operations with RDF content as parameters.

A KP may provide information for the smart space and use information provided by other KPs. The information content is not restricted in any way—it may be information relating to the physical environment, to the KPs themselves or anything. Thus, multiple KPs from multiple vendors may share ad-hoc information across numerous domains, enabling cross-domain and cross-platform interoperability.

Application examples include context gathering in meetings [5], meeting room smart space [6], smart home [7], gaming, wellness and music mashup [8] and social networks [9].

Smart-M3 [10] (Multidomain, Multidevice, and Multivendor) is an open software platform [11] that implements the smart space concept. Smart-M3 has been developed by a consortium of companies and within research projects: Artemis JU funded Sofia project (Smart Objects for Intelligent Applications) and Finnish nationally funded program DIEM (Device Interoperability Ecosystem).

Real-life scenarios often involve a lot of information, which leads both to largish ontologies and possibly complex instances that the KPs need to handle. Thus, programming KPs on the level of SSAP operations and RDF triples bring unnecessary complexity for the developers, who have to divert effort for managing triples instead of concentrating on the application logic. The OWL representation of knowledge as classes, relations between classes, and properties maps quite well to object-oriented paradigm in practice (but not so

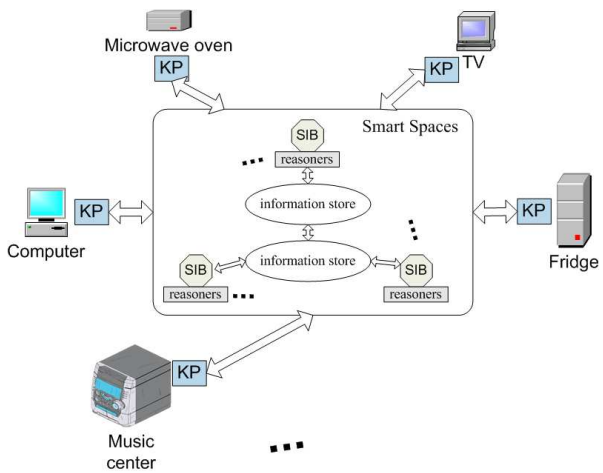


Figure 1. Smart spaces form a publish/subscribe system in a ubiquitous environment: KPs run on various types of computers and devices, the distributed knowledge store supports reasoning over cross-domain information

 ISBN: 978-1-61208-100-7

well in theory). Therefore, it is feasible to map OWL classes into OO classes and instances of OWL classes into objects[1] in programming languages. This approach effectively binds the subgraph describing an instance of an OWL class to an object in a programming language.

This paper describes the SmartSlog ontology library generator tool, our solution for allowing the construction of Smart-M3 KPs by programming with domain concepts that encoded in the relevant ontology.

SmartSlog is an ANSI C library generator for Smart Space ontology [12]. The generator maps an OWL ontology description to ANSI C code (ontology library), abstracting in KP code the ontology and communication with SIBs.

SmartSlog library simplifies constructing KP code. The code manipulates with ontology classes, relations, and individuals using predefined data structures and library API. The number of domain elements in KP code is reduced compared with the low-level triple-based scheme. The API are generic, hence does not depend on concrete ontology; all ontology entities appear as arguments in API functions. Search requests to SIB are written compactly by defining only what you know about the object to find (even if the object has many other properties).

The SmartSlog tool is constructed to take into account the limited resources available on small computers such as mobile and embedded devices. For example, the KP code does not need to maintain the whole ontology as unused entities can be removed. Also, triples are not kept indefinitely as the memory is freed immediately after the use. Furthermore, even if a high-level ontology entity consists of many triples, its synchronization with SIB transfers only a selected subset, saving on communication. These features make it possible to use SmartSlog when developing KPs for small devices—devices that are expected to play a central role in ubiquitous environments.

The rest of the paper is organized as follows. Section II briefly discusses related work Section III introduces the SmartSlog with its architectural and implementation details. Section IV presents generic SmartSlog library API and data structures. Section V describes SmartSlog optimizations. Section VI shows an example of application construction. Section VII concludes the paper.

II. Related Work

SmartSlog is closely related to code generators for C/GLib Smart-M3 KP API and Smart-M3 Python KP API, as they all use a common back-end for analyzing the ontologies and creating a model for code generation (Smart-M3 CodeGen in [11]). Also, the ontology APIs for generated by SmartSlog and the C/GLib generator are very similar. However, the code generated by SmartSlog is more concerned with adequate performance even on low-end devices. For example, dependencies are kept to minimum and memory usage is predictable and bounded.

Ontology based code generation facilities are also provided as part of the Sofia application development kit (ADK) [13] for Java-based KPs. The Sofia ADK is an Eclipse-based toolset for creating smart space applications. The view towards software developer is very similar to the SmartSlog, namely providing programming language view to the concepts defined in an ontology.

Similar ideas also exist in the semantic web world, with projects aiming to provide object-RDF mapping[2] libraries. These libraries are typically not tied to any ontology and implemented in interpreted languages, such as RDFAlchemy [14] in Python or Spira [15] in Ruby. Obviously the approach is very difficult both to implement and to use in statically typed compiled languages such as C, while very convenient in dynamically typed, interpreted languages.

III. Ontology Library Architecture

SmartSlog is built into the base Smart-M3 ontology library generation scheme (Smart-M3 CodeGen in [11]), see Figure 2). For a KP developer, the use scenario consists of two basic steps. First, she (thinking in ontology terms) provides a problem domain specification as an OWL description. The generator inputs the specification and outputs the ontology library. The latter is an interface that eliminates the developer from low-level triple-based details. Second, she uses the library when writing her KP code. The KP logic is implemented in high-level terms of the specified ontology. Note that the developer can easily start coding from KP template and Makefile generated optionally.

An ontology library generator uses a static templates/handlers scheme. Code templates are “pre-code” of data structures that implement ontology classes and their properties. Since names of ontology entities depend on a given ontology, each template contains a tag $\langle \mathrm{name} \rangle$ instead of every proper name. The generator has a set of handlers; each handler transforms one or more templates into final code replacing tags with the names taken from the ontology.

The transformation happens during ontology RDF graph traversal based on Jena OWL framework [16]. The latter constructs a meta-model to represent the graph. The generator comprehensively traverses this model, and those nodes are visited that a handler needs to transform its templates into final code.

Templates and their handlers are device-aware. The dependence is resolved on the level of a mediator library that implements triple-based ontology operations for RDF elements and SIB communication. SmartSlog uses KPI_low [17] as a mediator library, oriented to small embedded devices and ANSI C programming.

[1] These objects only have attributes, but no methods and thus no behavior

[2] In the spirit of object-relational mapping

 ISBN: 978-1-61208-100-7

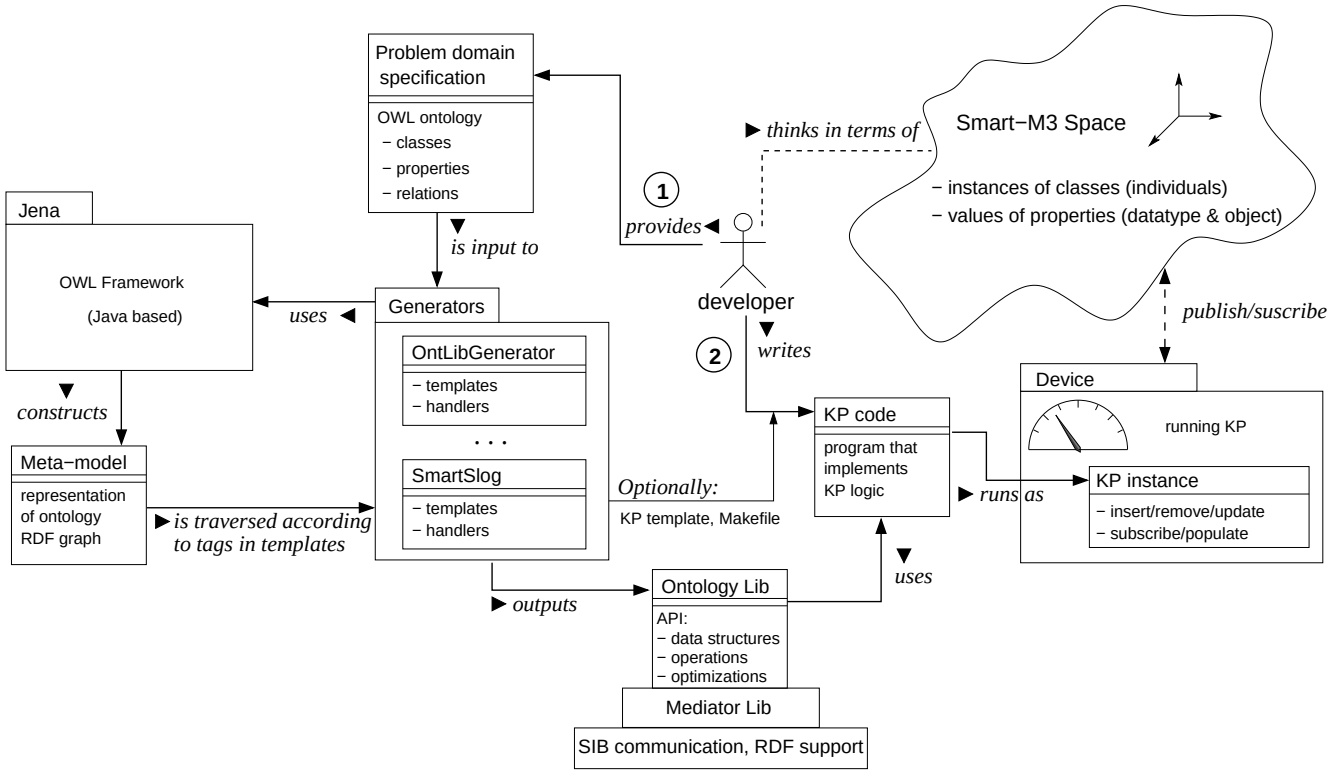


Figure 2. Smart-M3 ontology library generation scheme. Each OntLibGenerator implements own code templates and handlers oriented to a specific mediator library. SmartSlog extends the set of available generators for producing libraries on the top of KPI_low interface (for low-performance devices)

A SmartSlog library consists of two parts: dependent and independent on the given ontology (Figure 3). The SmartSlog generator produces ontology-dependent parts. It is implemented on the top of Smart-M3 CodeGen and uses own ANSI C templates (oriented to KP_low interface). The whole ontology can be represented in several files.

The generator iteratively calls the Jena meta-model. The corresponding templates are loaded and processed, and the final code is generated in files ⟨name⟩.c and ⟨name⟩.h, where ⟨name⟩ is the ontology file name. The code implements all ontology classes and properties as structures in C. Note that the generated code can be optimized further by removing ontology entities unused in the KP.

The ontology-independent part contains API: basic data structures (for generic ontology class, property, and individual) and functions for their manipulation. The code structure is shown in Table I. SmartSlog API uses KPI_low when communicating with SIB. Hence, the ontology-independent part implements all high-level ontology entity transformations to low-level triples and vice versa.

This library division into two parts allows constructing efficient applications. If the ontology changes the ontology-independent part does not require recompiling; it can be shared by several KPs that use different ontology. Ontology-dependent part can be shared by KPs with the same ontology. These cases are typical since multiple smart space applications with different ontology can run on the same device as

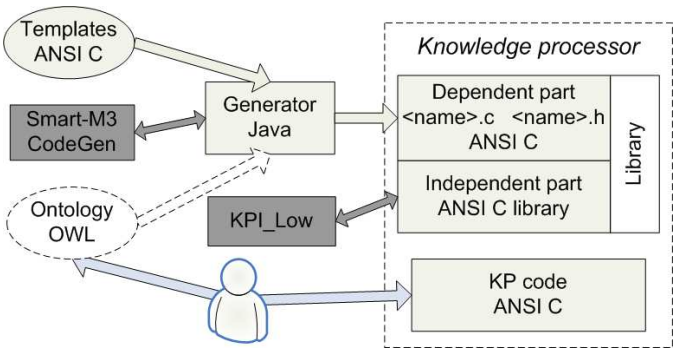


Figure 3. The SmartSlog ontology library architecture: ontology-dependent and ontology-independent parts

well as multiple KPs form one smart space application.

Optionally the generator produces a template for KP code (and Makefile) with function `main()`. In the beginning, it initializes local ontology structures and joins the smart space. In the end, it leaves the smart space gracefully. In between, the developer inserts own code (KP logic).

IV. Library API

SmartSlog API evolves over the generic API of Smart-M3 CodeGen [11]. “Generic” means that API does not depend on ontology: classes, properties, and individuals appear as arguments in API functions. Datatype and object properties

 ISBN: 978-1-61208-100-7

TABLE I
SMARTSLOG CODE STRUCTURE FOR ONTOLOGY-INDEPENDENT PART

Files *.c and *.h	Description	Fnc/Str	LOC/COM
generic.h	Declarations of all API data structures and functions.	0/0	13/21
structures	Base data structures and functions for them.	11 / 4	201 / 214
classes	Manipulation with classes.	20 / 0	318 / 221
properties	Manipulation with properties.	28 / 0	550 / 312
	Sum:	59 / 4	1082 / 768
ss_func	Access to smart space (joining, leaving, ...).	5 / 0	34 / 35
ss_classes	Manipulation with classes in smart space.	15 / 0	344 / 412
ss_properties	Manipulation with properties in smart space.	12 / 0	351 / 383
ss_populate	Population of individuals from smart space.	2 / 0	62 / 107
ss_subscribe	Subscribe containers and functions.	24 / 3	483 / 179
	Sum:	58 / 3	1274 / 1116
kpi_interface	Interface to KPI_low (triple transformation).	9 / 0	401 / 299
utils/*	Auxiliary defs&funcs. Unused directly in KP code.	51 / 3	540 / 641
	Sum:	60 / 3	941 / 940
	TOTAL:	177 / 10	3297 / 2824

Fnc/Str counts the number of functions and structures. LOC/COM was computed by the CCCC tool [18].

are treated similarly. Run-time checking must be performed for arguments.

In SmartSlog, each ontology class, property, and individual is implemented as a C structure (types `property_t`, `class_t`, and `individual_t`). The API has generic functions that handle such data objects regardless of their real ontology content. Currently supported OWL constraints are class, datatypeproperty, objectproperty, domain, range, and cardinality. For example, a class knows all its superclasses, OWL one of classes, properties, and instances (individuals); the implementation is as follows.

```
typedef struct class_s {
  int rtti;       /* run-time type information */
  char *classtype;      /* type of class, name */
  list_t *superclasses; /* all superclasses */
  list_t *oneof;        /* class oneof value */
  list_t *properties;   /* all properties*/
  list_t *instances;    /* all individuals */
} class_t;
```

API functions are divided into two groups: for manipulating with local objects and for communicating with SIB. The first group (local) includes functions for

- classes and individuals: creating data structures and manipulating with them locally;
- properties: operations set/get, update, etc. in local store (also run-time checks for correctness, e.g., cardinality and property values).

For example, creating individual and setting its properties:

```
individual_t *aino = new_individual(CLASS_WOMAN);
set_property(aino, PROPERTY_LNAME, "Peterson");
```

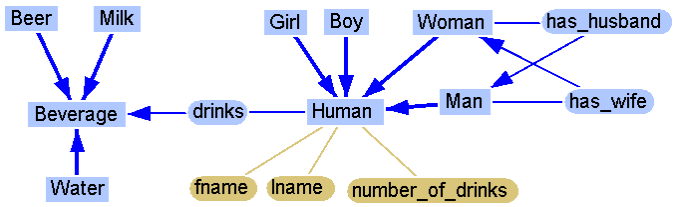


Figure 4. Ontology for humans and their drinks

In this example, the definitions of `CLASS_WOMAN` and `PROPERTY_LNAME` are in the library ontology-dependent part for the ontology shown in Figure 4. (We used GrOwl tool [19]: classes are in blue rectangles, datatype properties are in brown ovals, object properties are in blue ovals.)

The second group (to/from smart space) has prefix "`ss_`" in function names and allows accessing smart space for

- individuals: insertion, removal, and update;
- properties: similarly to the local functions but the data are to/from smart space (it requires transformation to/from triples and calling the mediator library);
- querying for individuals in smart space (existence, yes/no answer);
- populating individuals from smart space by query or by subscription.

For example, inserting an individual and then updating some of its properties:

```
ss_insert_individual(aino);
    . . .
ss_update_property(aino,
    PROPERTY_LNAME, "Ericsson");
```

Subscription needs more discussion. In advance, a subscription container is created to insert those individuals which to subscribe for. Optionally, the container inserts the properties whose values are interested only. Then KP explicitly subscribes for selected properties of selected individuals.

Subscription is synchronous or asynchronous. The former case is simplest; KP is blocked waiting for updates. Even devices without thread support allow synchronous subscription. The latter case is implemented with a thread that controls updates from smart space and assigns them to the containers. KP is not blocked, and updates come in parallel.

Internally, communication with SIB leads to the composition/decomposition of high-level ontology entities from/to triples and calling KPI_low for triple-based data exchange.

SmartSlog API covers all basic primitives of a publish/subscribe system. Compared with Smart-M3 CodeGen that provides similar primitives, SmartSlog API has the following advantages. Smart-M3 CodeGen API depends on glib library, e.g., using list data structures. Low-performance devices do not support glib. In contrast, SmartSlog has no such requirements for underlying libraries. Smart-M3 CodeGen currently does not allow asynchronous subscription important for some smart space applications.

 ISBN: 978-1-61208-100-7

SmartSlog extends generic API by patterns for ontology-based filtering and search. Each pattern is an `individual_t` structure and can be thought as an abstract individual where only a subset of properties is set. A pattern is either pattern-mask or pattern-request.

A pattern-mask is for selecting properties of a given a class or individual. It needs when a subset of properties is used, and the pattern includes only those properties. Then this pattern is applied to the given class or individual, e.g. for modest updating the properties. For example, let us update only the last name of "Aino" (see the ontology in Figure 4).

```
individual_t *aino_p = new_individual(CLASS_WOMAN);
set_property(aino_p, PROPERTY_LNAME, NULL);
ss_update_by_pattern(aino, aino_p);
```

As a result, only the last name value is transferred to smart space. Compared with `ss_update_property()` the benefit becomes obvious when KP needs to update several properties at once or it can form the property subset only in run-time. The same scheme works for population to transfer data modestly from smart space.

A pattern-request is for compact definition of search queries to smart space. A pattern is filled with those properties and values that characterize the individual to find. For example, let us find all men whose first name is "Timo" and wife's first name is "Aino".

```
individual_t *timo_p =
    new_individual(CLASS_MAN);
individual_t *aino_p =
    new_individual(CLASS_WOMAN);

set_property(timo_p, PROPERTY_FNAME, "Timo");
set_property(aino_p, PROPERTY_FNAME, "Aino");
set_property(timo_p, PROPERTY_HAS_WIFE, aino_p);

timo_list = ss_get_individuals_by_pattern(timo_p);
```

In this example, two patterns ("Timo" and "Aino") and two properties (datatype "fname" and object "has_wife") form a subgraph. The SmartSlog library matches the subgraph to the smart space content. As a result, a list of available individuals is returned. Currently, searching leads to iterative triple exchange and matching at the local side. In future, it can be implemented on the top of SPARQL [20], and the most processing moves to the SIB side.

V. Implementation Optimizations

SmartSlog is primarily oriented to low-performance devices [21] and uses a limited subset of ANSI C [22]. SmartSlog does not optimize its mediator library (KPI_low). Instead, SmartSlog optimizes local data structures, the (de)composition (to)from triples, and the way how the mediator library is used. Some of these optimizations are also usable for computers with no hard performance restrictions.

Each ontology entity is implemented as a C structure of constant size. For ontology with N entities the SmartSlog ontology-dependent part is of size $O(N)$. In many problem domains, however, the whole ontology contains a lot of classes and properties.

SmartSlog provide constants that limits the number of entities, hence the developer can control the code size. Furthermore, one KP often needs only a subset of them (see our example in Section VI). SmartSlog allows the developer to select what ontology entities she needs in KP code (or to deselect unneeded). Currently, it is implemented with a simple mechanism based on #{define, ifdef} C compiler preprocessor directives.

Inserting and receiving individuals to and from smart spaces lead to transferring a lot of triples. The network traffic can be reduced by transfer a subset only. SmartSlog API allows the KP developer to explicitly select (using patterns, see Section IV) what properties to use in an operation. That is, even if an individual in the smart space has dozens of properties, KP can populate only few of them (the others are unused this time). Moreover as we discussed above, KP can also deselect totally unneeded properties from its code.

As a result, KP works locally with a subset of properties required by KP semantics at current time instance. There is no need to load/save all properties from/to the smart space.

Smart-M3 CodeGen keeps a triple store—a local cache of smart space content. For large ontology it is expensive. In contrast, SmartSlog does not intend to store any triple for long time. Ontology entities are stored in own structures. When a triple is needed it is created and processed. Then the memory is freed immediately after the usage.

SmartSlog supports both types of subscriptions: synchronous and asynchronous. The latter case requires threading. SmartSlog uses POSIX threads [23] available on many embedded systems [21]. Nevertheless, SmartSlog allows switching the asynchronous subscription off if the target device has no thread support.

VI. Use Case Example

In this section, we show how SmartSlog can be used for constructing a simple Smart-M3 application. In spite of the simplicity, the example illustrates such SmartSlog features as patterns and subscriptions (synchronous and asynchronous). Both datatype and object properties are used.

Let Ericsson's family consist of Timo (husband) and Aino (wife). Timo likes drinking beer outside home. Aino has to control Timo's drinking via monitoring the amount of beer he has drunk already. If the amount is exceeding a certain bound (e.g., `MAX_LITRES_VALUE=3`) she notifies Timo by SMS that it's good time to come back to home.

The ontology for such personal human data was shown in Figure 4 above. When Timo starts drinking he associates his object property "drinks" with class "Beer". Then Timo keeps his drink counter "number_of_drinks" in smart space and regularly updates it. Aino can subscribe to this counter.

For messaging, the family uses the ontology shown in Figure 5. Aino sends SMS to notify Timo via smart space.

 ISBN: 978-1-61208-100-7

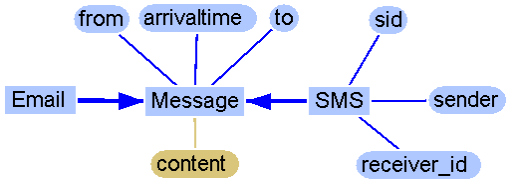


Figure 5. Ontology for messaging

Timo subscribes for SMS and checks each SMS he received for who sent it (by phone number). Hence Timo recognizes a notification SMS from his wife.

Given these two ontology files, SmartSlog generator produces files `drinkers.{c, h}`. Since the ontology includes more details than needed for this application, excessive classes and properties can be disabled in the final code by compiler preprocessor directives.

The KP code for Timo can be constructed with SmartSlog using the following scheme.

1. Create Timo, set his properties, and insert the individual to the smart space.

```
individual_t *timo = new_individual(CLASS_MAN);
set_property(timo,PROPERTY_FNAME, "Timo");
     . . .
ss_insert_individual(timo);
```

2. Timo keeps his counter in the smart space.

```
individual_t *beer = new_individual(CLASS_BEER);
ss_set_property(timo, PROPERTY_DRINKS, beer);
```

3. Timo subscribes to SMS from Aino: creating an individual for SMS and filling the subscribe container. Then asynchronous (parameter “true”) subscription starts.

```
individual_t *sms = new_individual(CLASS_SMS);
add_data_to_list(subscribed_prop_list,
     PROPERTY_FROM);
add_data_to_list(subscribed_prop_list,
     PROPERTY_TO);

subscription_container_t *container=
     new_subscription_container();
add_individual_to_subscribe(container,
     sms, subscribed_prop_list);

ss_subscribe_container(container, true);
```

4. Timo drinks, updates the counter, and checks SMS.

```
while(sms_notify(sms)) {
    amount += drink(timo);
    ss_update_property(timo,
         PROPERTY_NUMBER_OF_DRINKS, amount);
}
```

Similarly, the KP code for Aino is constructed as follows.

1. Aino searches Timo in the smart space by pattern.

```
individual_t *wife = new_individual(CLASS_WOMAN);
set_property(wife, PROPERTY_LNAME, "Ericsson");
set_property(wife, PROPERTY_FNAME, "Aino");

individual_t *timo = new_individual(CLASS_MAN);
set_property(timo, PROPERTY_FNAME, "Timo");
set_property(timo, PROPERTY_HAS_WIFE, wife);
     . . .

list = ss_get_individuals_by_pattern(timo);
```

2. Synchronous (parameter “false”) subscription waits for Timo is starting to drink.

```
subscription_container_t *container=
     new_subscription_container();
add_individual_to_subscribe(container, timo,
                            properties);
ss_subscribe_container(container, false)

property_t *drinks = get_property(timo,
                        PROPERTY_DRINKS);
if (drinks==NULL) wait_subscribe(container);
```

3. Monitoring Timo’s counter and checking the limit. Synchronous subscription is similar to the above.

```
/* Subscribing for Timo’s counter */
   . . .

while(1) {
    amount = get_property(timo,
              PROPERTY_NUMBER_OF_DRINKS);
    if (amount >= MAX_LITRES_VALUE) {
        /* Send SMS to Timo */
        break;
    }
    wait_subscribe(container_counter);
}
```

4. Create an individual for SMS and insert it to the smart space. Properties “to” and “from” are required.

```
individual_t *sms=new_individual(CLASS_SMS);
set_property(sms, PROPERTY_TO,
                  TIMO_PHONE_NUMBER);
set_property(sms, PROPERTY_FROM,
                  WIFE_PHONE_NUMBER);
ss_insert_individual(sms);
```

VII. Conclusion and Future Work

The addressed area of high-level ontology library generation for low-performance devices is very important. The realization of the ubiquitous computing vision will by definition include a lot of small devices around us. Allowing these small devices to easily share information with other devices and architectures, large or small, will be very important.

In this paper we described SmartSlog—a tool that supports efficient programming such devices for participating in smart space applications. The resulting code is compact due to high-level ontology style, portable due to adhering to ANSI C and POSIX standards, modest and optimizable to device capacity due to the design. We believe that SmartSlog will become an important element of the Smart-M3 platform.

The paper presented our work-in-progress. The future work includes more optimization depending on the needs of a concrete KP. For example, ontology metainformation allows defining what types of embedded devices can use a certain part of ontology. It leads to implementing various ontology manipulations that utilize metainformation on

 ISBN: 978-1-61208-100-7

versioning, namespaces, and other differentiation characteristics. Another important direction of our future work is optimization of the SIB ↔ KP communication. For example, a part of triple-based processing can be moved to the SIB side using SPARQL query language; its support will appear in Smart-M3 soon.

ACKNOWLEDGMENT

Authors would like to thank Finnish-Russian University Cooperation in Telecommunications (FRUCT) program for the provided support and R&D infrastructure. The special thanks to Nokia university collaboration program for providing publication grant and all FRUCT experts for commenting and reviewing the project. We would also like to thank Vesa Luukkala and Ronald Brown from Nokia Research Center for providing feedback and guidance during the construction of the SmartSlog tool.

REFERENCES

[1] I. Oliver, J. Honkola, and J. Ziegler, "Dynamic, localised space based semantic webs," in *Proc. IADIS Int'l Conf. WWW/Internet 2008*. IADIS Press, Oct. 2008, pp. 426–431.

[2] I. Oliver, "Information spaces as a basis for personalising the semantic web," in *Proc. 11th Int'l Conf. Enterprise Information Systems (ICEIS 2009)*, vol. SAIC, May 2009, pp. 179–184.

[3] D. Fensel, W. Wahlster, H. Lieberman, and J. Hendler, Eds., *Spinning the semantic web : bringing the World Wide Web to its full potential*. The MIT Press, 2005.

[4] R. Baldoni, M. Contenti, and A. Virgillito, "The evolution of publish/subscribe communication systems," in *Future Directions in Distributed Computing*, ser. Lecture Notes in Computer Science, vol. 2584. Springer, 2003, pp. 137–141.

[5] I. Oliver, E. Nuutila, and S. Törmä, "Context gathering in meetings: Business processes meet the agents and the semantic web," in *The 4th Int'l Workshop on Technologies for Context-Aware Business Process Management (TCoB 2009) within Proc. Joint Workshop on Advanced Technologies and Techniques for Enterprise Information Systems*. INSTICC Press, May 2009.

[6] A. Smirnov, A. Kashnevik, N. Shilov, I. Oliver, S. Balandin, and S. Boldyrev, "Anonymous agent coordination in smart spaces: State-of-the-art," in *Proc. 9th Int'l Conf. Smart Spaces and Next Generation Wired/Wireless Networking (NEW2AN'09) and 2nd Conf. Smart Spaces (ruSMART'09)*, ser. Lecture Notes in Computer Science, vol. 5764. Springer-Verlag, 2009, pp. 42–51.

[7] K. Främling, I. Oliver, J. Honkola, and J. Nyman, "Smart spaces for ubiquitously smart buildings," in *Proc. 3rd Int'l Conf. Mobile Ubiquitous Computing, Systems, Services and Technologies (UBICOMM 2009)*. IEEE Computer Society, 2009, pp. 295–300.

[8] J. Honkola, H. Laine, R. Brown, and I. Oliver, "Cross-domain interoperability: A case study," in *Proc. 9th Int'l Conf. Smart Spaces and Next Generation Wired/Wireless Networking (NEW2AN'09) and 2nd Conf. Smart Spaces (ruSMART'09)*, ser. Lecture Notes in Computer Science, vol. 5764. Springer-Verlag, 2009, pp. 22–31.

[9] S. Balandin, I. Oliver, and S. Boldyrev, "Distributed architecture of a professional social network on top of M3 smart space solution made in PCs and mobile devices friendly manner," in *Proc. 3rd Int'l Conf. Mobile Ubiquitous Computing, Systems, Services and Technologies (UBICOMM 2009)*. IEEE Computer Society, 2009, pp. 318–323.

[10] J. Honkola, H. Laine, R. Brown, and O. Tyrkkö, "Smart-M3 information sharing platform," in *The 1st Int'l Workshop on Semantic Interoperability for Smart Spaces (SISS 2010) in conjunction with IEEE ISCC 2010*, Jun. 2010.

[11] "Download Smart-M3 software for free at SourceForge.net," Release 0.9.4beta, May 2010. [Online]. Available: http://sourceforge.net/projects/smart-m3/

[12] "Download SmartSlog software for free at SourceForge.net," Release 0.22, Apr. 2010. [Online]. Available: http://sourceforge.net/projects/smartslog/

[13] P. Liuha, A. Lappeteläinen, and J.-P. Soininen, "Smart objects for intelligent applications - first results made open," *ARTEMIS Magazine*, no. 5, pp. 27–29, Oct. 2009.

[14] "RDFAlchemy," Jul. 2010. [Online]. Available: http://www.openvest.com/trac/wiki/RDFAlchemy

[15] "Datagraph's spira at github," Version 0.0.5, Jun. 2010. [Online]. Available: http://github.com/datagraph/spira

[16] "Jena – a semantic web framework for java," Jul. 2010. [Online]. Available: http://jena.sourceforge.net/

[17] "Download KPI_low software for free at SourceForge.net," Jun. 2010. [Online]. Available: http://sourceforge.net/projects/kpilow/

[18] T. Littlefair, "CCCC — C and C++ code counter," May 2010. [Online]. Available: http://cccc.sourceforge.net/

[19] S. Krivov, R. Williams, and F. Villa, "GrOWL: A tool for visualization and editing of OWL ontologies," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 5, no. 2, pp. 54–57, 2007.

[20] E. Prud'hommeaux and A. Seaborne, "SPARQL query language for RDF," W3C Recommendation, Jan. 2008. [Online]. Available: http://www.w3.org/TR/rdf-sparql-query/

[21] M. Barr and A. Massa, *Programming Embedded Systems: With C and GNU Development Tools*. O'Reilly Media, Inc., 2006.

[22] "The ANSI C standard (C99)," ISO/IEC, Tech. Rep., 1999.

[23] "Standard for information technology — portable operating system interface (POSIX)," Tech. Rep. 1003.1-2001/Cor 2-2004, 2004.

 ISBN: 978-1-61208-100-7

Experimental Comparison of Frequency Hopping Techniques for 802.15.4-based Sensor Networks

Luca Stabellini and Mohammad Mohsen Parhizkar
The Royal Institute of Technology (KTH), ICT/Communication Systems
Electrum 418, SE-164 40 Kista, Sweden
Email: {lucast,mmpa}@kth.se

***Abstract*—Frequency hopping communication schemes represent an attractive alternative for interconnecting low power wireless sensor nodes operating in unlicensed bands. The use of multiple communication channels can in fact mitigate the negative effects of interference induced by collocated wireless networks and potentially results in improved reliability. With this respect, quite a few adaptive variations, aiming at improving the resilience of frequency hopping toward interference have been recently proposed. In this paper we present the experimental evaluation of three different hopping schemes: we implement a traditional hopping algorithm and two adaptive variations on TMote Sky sensor motes and quantify their energy performance under different channel conditions. We also compare the effectiveness of these three hopping techniques against the one of a communication scheme making use of a single channel. Our results, obtained considering a two-node topology, indicate that our previously proposed utility based adaptive frequency hopping approach is the most effective in avoiding interference and can significantly reduce the overall energy consumption despite its higher complexity. The performed experiments also show that even though reliable single-channel communication might be possible, by using frequency hopping sensors can limit the performance degradation induced by interference while avoiding the energy overhead introduced by the spectrum sensing algorithms that nodes have to use for identifying clear channels.**

***Keywords*-Frequency Hopping; Adaptive Frequency Hopping; Interference Mitigation; Coexistence in Unlicensed Bands; Wireless Sensor Networks;**

I. INTRODUCTION

A. Background

Frequency hopping communication techniques represent a common solution for interconnecting wireless personal area network devices operating in unlicensed bands. The basic idea implemented by these schemes is to allow communication among two or more wireless terminals by means of synchronous hopping over a defined set of channels (also referred to as the *hopset*) that are selected for packet transmissions in a pseudo-random fashion. Such a strategy guarantees a certain degree of frequency diversity and potentially allows to mitigate the interference that might be induced by transmissions of collocated wireless networks, consequently improving reliability.

This nice feature is extremely attractive for low-power devices such as wireless sensor nodes. As outlined by recent surveys conducted in the context of industrial automation (see for instance [1, 2]) the potential unreliability of wireless communications is in fact perceived as one of the major barriers to the adoption of wireless sensing technologies for commercial applications. By exploiting multiple communication channels through frequency hopping, sensor devices can mitigate the negative effects of interference and potentially improve communication reliability. We remark that the attention towards frequency hopping transmission schemes has been constantly growing during the last years as witnessed by the recent proliferation of radio standards and communication protocols adopting this solution: examples are provided by IEEE 802.15.1 [3], WirelessHART [4] and ISA SP100 [5].

While frequency hopping techniques can guarantee a certain resilience against bad channel conditions, it is well known that performance of this kind of systems can be severely degraded if some of the channels belonging to the used hopset constantly experience bad communication quality. For dealing with this problem adaptive algorithms have been proposed: in particular two approaches have been investigated in the literature. The first one (see for instance the adaptive specifications included in [6], [7] and references therein) aims at identifying bad channels that are subsequently removed from the used hopset whose cardinality is thus reduced: note that this approach is implemented by a variety of adaptive algorithms such as ISOAFH [8] (that targets the identification of interference induced by WLAN devices) and EAFH [9] (that also adapts the size of transmitted packets to the particular channel conditions of each frequency band). The latter instead adopts a probabilistic approach that rather than removing channels from the hopset, uses all the available frequency bands but with probabilities that depend on channel quality (see for instance [10] and [11]).

B. Problem Formulation and Contribution

These two approaches introduce different overhead, present different complexity and provide different advantages. For instance removing *bad* channels from the hopset results in relatively low complexity: on the other hand this choice might introduce delays in the adaptation process (due to the need of identifying those bad channels with a certain accuracy) and frequency bands that are removed from the hopset might have to be periodically re-checked resulting in additional wastes of energy and time. The probabilistic approach introduced in [10] allows to overcome these limitations: adaptation can in fact start immediately after the first packets are transmitted/received and the available resources can be exploited in a more granular manner. This however results in higher computational complexity and requires frequent exchange of information among the communicating nodes in order to maintain synchronous hopping and avoid the multi-channel hidden terminal problem [12].

We remark that the impact of these different design choices over the performance of wireless devices has always been evaluated through simulations, and we are not aware of any

 ISBN: 978-1-61208-100-7

published research work aiming at quantifying and comparing through experiments on real hardware the effectiveness of different hopping algorithms. In this paper we provide such an experimental comparison. In particular, our contribution is two-fold:

- first, using TMote Sky sensor nodes we implement the two adaptive approaches previously described as well as a traditional hopping algorithm and quantify and compare their energy performance by means of extensive experiments under different channel conditions;
- we further compare the energy performance of frequency hopping against the one of a communication scheme making use of a single channel.

The use of (adaptive) frequency hopping has been envisaged for improving the performance of wireless systems under three different settings: (i) in presence of frequency static interference (such as for instance the one induced by IEEE 802.11 b/g devices), (ii) in presence of frequency dynamic interference (such as the one induced by collocated networks making use of frequency hopping) and (iii) in presence of bad channel conditions (for instance induced by multipath fading, frequency selective channel responses or other propagation anomalies). In this work we consider only frequency static interference. As done in many other papers (see for instance [13, 14]) we limit the focus of our investigation to a simple two-node topology: the extension to networks comprising more than two nodes is left for future work.

The rest of this paper is organized as follows. Section II outlines the setting of our experiments and Section III describes the hopping algorithms we implemented. Experimental results are presented in Section IV, while conclusions are drawn in Section V.

II. Experimental Setup

A. Network Scenario

The setting of our experiments is sketched in Figure 1: we focus on a simple two-node topology comprising two TMote Sky sensor nodes S_1 and S_2 located 1 meter far away from each other, and consider the exchange of a certain bulk of data, consisting of N packets, from node S_1 (acting as transmitter) to node S_2 (the receiver). Each transmitted packet has a payload of 100 Bytes. The two sensors run the Contiki operating system [15] and are connected to two PCs through a USB connection that allows to collect transmission statistics. Packet transmissions are implemented using the following handshake mechanism: on slot i, node S_1 sends a data block; S_2 verifies the correctness of the received packet by means of a 16-bit cyclic redundancy check (we used the CRC16 provided by the Contiki operating system [15]). An acknowledgement or a not acknowledgement (requesting the retransmission of the corrupted block) is then transmitted on slot $i + 1$.

The TMote Sky used for our experiments feature an IEEE 802.15.4 2420 Chipcon wireless transceiver operating in the 2.4 GHz ISM band: the available hopset comprises thus the 16 frequency bands $c_{11}, \ldots, c_{26}$ specified by the IEEE 802.15.4 radio standard. We implemented a simple MAC-layer synchronization routine where nodes hop to the channel that has to be used for the upcoming transmission either immediately after sending or receiving a packet or after a certain time-out t_{Max} expires: for our experiments we fixed $t_{\text{Max}} = 25$ [ms].

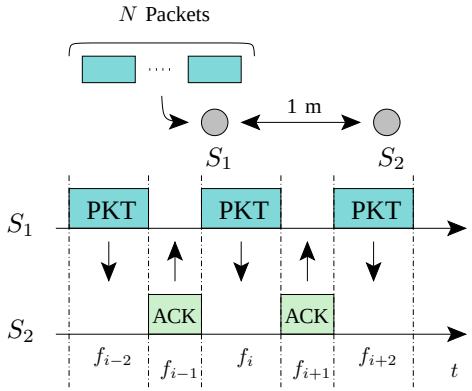


Fig. 1. Sketch of the considered scenario.

We focused on a simple two-node topology for two reasons. On one hand, the implementation of frequency hopping schemes over multi-node and potentially multi-hop networks requires that different issues, one of them being synchronization, are addressed. This is out of the scope of this paper: furthermore, we note that the same problems will arise independently on the used hopping technique. Considering only two sensors simplifies the implementation process and allows us to focus on the comparison of the energy performance of the different communication techniques. As a second aspect, we remark that networks comprising several sensors can potentially be organized in a countless number of different topologies. The choice of a particular topology (for instance a star rather than a tree or a mesh) might make the obtained results dependent on the particular considered setting: by focusing on the single link between two nodes instead it is possible to obtain general results that are not topology dependent.

B. Experimental Approach

We performed two different experimental campaigns. For the first one, we selected an interference-free environment: we set the transmission power of the nodes so as to achieve a negligible packet loss rate (we verified that an output power of -10 dBm was sufficient for this purpose) and we *artificially* controlled using software-defined values the probability p_i of receiving a corrupted packet over channel c_i (note that in fact all packets are correctly received however with probability p_i, a packet is discharged and considered lost). This approach, that has previously been used for instance in [16, 17], basically permits to *simulate* the performance of the considered hopping algorithms on real motes (thus allowing to quantify their exact energy consumption) while controlling the packet error probability experienced over the wireless channel.

Our second campaign was instead performed inside the office spaces of the Radio Communication Systems department of KTH where the 2.4 GHz ISM band is heavily used by several wireless terminals such as laptop, PDA and wireless keyboards/mouses: as an example, the variation of average channel occupancy over the 16 IEEE 802.15.4 channels during a 7-day period is shown in Figure 2. The spectrum is mainly

 ISBN: 978-1-61208-100-7

utilized by WLAN devices (i.e., operating within the IEEE 802.11g radio standard): on the plotted figure three non-overlapping WiFi carriers can be easily identified.

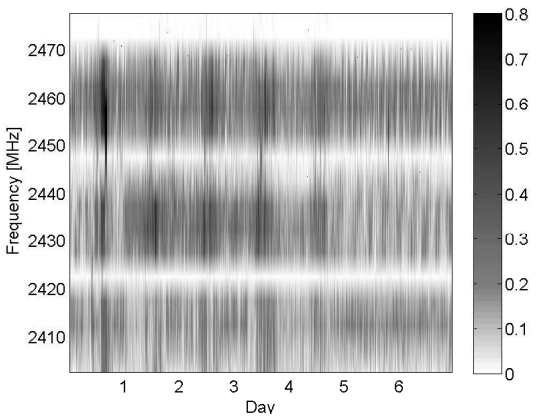


Fig. 2. Average channel occupancy for the 16 IEEE 802.15.4 channels during a 7-day period, from April 19 to April 25.

Transmissions of these devices, although not controllable, provide an example of interference pattern that sensors are likely to experience in real scenarios and thus represent an excellent source of interference for our motes. Our experiments have been performed over a time-frame of seven days: on each day we iterated several transmissions for each of the considered hopping algorithms in order to ensure that all of them were tested under a wide range of channel conditions. The purpose of this second campaign was to *qualitatively* assess the performance of the hopping techniques in a real scenario. Our conclusions are however mainly based on data obtained during the first round of controlled experiments.

C. Performance Metrics

We quantified the performance of the considered communication schemes by measuring the total energy E^{Tot} spent by the two-node system for the successful delivery of the specified bulk of data: this accounts for the energy spent while transmitting and receiving packets and control messages as well as for the energy required by the CPU of the two nodes. For this purpose, we used the online energy estimation routine [18] provided by the Contiki operating system: this allows to measure the time spent by nodes on each of the four following states: transmit, receive, CPU and LPM (Low Power Mode). The total consumed energy can then be computed multiplying the obtained times by the power consumption of sensors on each state (for TMote Sky we have: listen 60 [mW], transmit (-10dBm) 33 [mW], CPU 5.4 [mW] and LPM 0.1635 [mW] [19]). We normalized the obtained values to the amount of energy required to complete a packet exchange (comprising both the transmission of the data packet as well as the following acknowledgement) in interference-free conditions.

III. Hopping Algorithms

We now briefly describe the three hopping algorithms that have been the object of our evaluation. In particular we implemented:

- a traditional Frequency Hopping (FH) scheme;
- an Adaptive Frequency Hopping algorithm similar to the one defined in [6]; this adaptive approach is the one adopted by several radio standards such as for instance WirelessHART [4] and IEEE 802.15.1 [3];
- the Utility Based Adaptive Frequency Hopping (UBAFH) algorithm introduced in [10].

We also considered a simple communication scheme where only a single channel is used and no hopping strategy is implemented. The aforementioned hopping approaches will be detailed in the next subsections.

A. Traditional Frequency Hopping

If a traditional frequency hopping technique is implemented the channels belonging to the hopset are used in a pseudo-random fashion. For this purpose S_1 and S_2 share a common seed: this is used to generate random numbers and chose the frequency band that shall be used for the upcoming transmission. All the 16 available channels are equally likely to be selected in each time-slot.

B. Adaptive Frequency Hopping

We consider the adaptive frequency hopping algorithm specified in [6] (note that in [6] the focus was on the IEEE 802.15.1 radio standard: we here generalize the proposed scheme to IEEE 802.15.4): S_1 and S_2 estimate the packet error rate experienced on each channel using a certain number of transmissions N_E. In this way S_1 estimates the probability of receiving a corrupted ACK/NACK, while S_2 estimates the probability of receiving a corrupted data packet. After this *channel classification* procedure has been completed, S_1 reports to S_2 his estimates, S_2 computes average channel conditions (by averaging his estimates with the ones received from S_1) and updates the hopset by removing channels with packet error rate greater than a certain threshold p^{Max}. The updated hopset is then communicated to S_1 and adaptation can start. This procedure can eventually be repeated on a periodic fashion in order to deal with changes of channel conditions.

We remark that [6] do not specifies the values of N_E and p^{Max} which can therefore be vendor specific: for our experiments we assumed $N_E = 16 \cdot 20$ (thus channel conditions are estimated considering in average 20 transmissions for each of the available frequency bands) and $p^{\text{Max}} = 0.5$. We stress that different choices for these parameters can be used to implement different performance tradeoffs. A low value of N_E allows to shorten the time required to perform channel classification and thus reduces the adaptation delay: on the other hand if N_E is too small, channels might be classified in an inaccurate manner and for instance good frequency bands might erroneously be removed from the hopset while bad channels might not be properly identified. Similar considerations should be made when selecting the packet error rate threshold p^{Max}: a high threshold might lead nodes to hop over interfered frequencies while lower values might induce a very selective channel classification procedure where several channels are removed from the hopset decreasing the degree of frequency diversity. This might be undesirable if nodes experience both frequency static interference as well as multi-path fading. The value we assumed for our experiments i.e. $p^{\text{Max}} = 0.5$ has been suggested in [13] and has been used in other published works.

 ISBN: 978-1-61208-100-7

C. Utility Based Adaptive Frequency Hopping

The utility based adaptive frequency hopping algorithm proposed in [10] adopts a different approach: S_1 and S_2 constantly maintain estimates $\hat{p}(c_i)$ for the packet error rate experienced on each of the available frequency bands. These estimates are computed using a window moving average that evaluates $\hat{p}(c_i)$ over channel c_i accounting for the last $N_T = 32$ transmissions. The obtained values are then mapped to a probability mass function defining channel usage probabilities and assigning to channels with better conditions higher values. For complexity reasons we modified the mapping function defined in [10] and considered instead:

$$f : \hat{p}(c_i) \rightarrow f\left(\hat{p}(c_i)\right) = \frac{\nu(\hat{p}(c_i))}{\sum_{j=1}^{16} \nu(\hat{p}(c_j))} \quad (1)$$

where:

$$\nu(\hat{p}(c_i)) = \begin{cases} 20 \cdot (1 - \hat{p}(c_i)) \cdot 32 & \text{if } \hat{p}(c_i) \leq \frac{3}{32} \\ 5 \cdot (1 - \hat{p}(c_i)) \cdot 32 & \text{if } \frac{3}{32} < \hat{p}(c_i) \leq \frac{12}{32} \\ 3 & \text{if } \hat{p}(c_i) > \frac{12}{32} \end{cases} \quad (2)$$

Note that the factor 32 that multiplies $1 - \hat{p}(c_i)$ is introduced in order to obtain integer quantities and reduce computational complexity. To the payload of each packet S_1 and S_2 add two bytes (thus the payload in this case has a total size of 102 octets) containing the outcomes of the last 16 packets transmissions: this allows to keep synchronous estimates of packet error rate at the two nodes (the reader is referred to [10] for additional details). The channel to be used at time-slot k is then selected using the information that nodes have up to time-slot $k - 16$. Also in this case, synchronous channel selection is ensured by using a seed known to both nodes. Note that channels are still chosen in a pseudo-random fashion, however, while for a traditional hopping algorithm, all the channels are equally likely to be used, in this case channels with better conditions (i.e. lower packet error rate) are assigned higher usage probabilities (proportionally to $\nu(\hat{p}(c))$) and are consequently selected more often than frequency bands where nodes experience high packet error rate.

IV. Results

A. Interference Controlled Environment

We start our performance evaluation by quantifying the complexity added by the adaptive schemes in absence of interference. Under these conditions, the energy consumption of the traditional FH algorithm represents our reference case: in Figure 3 (top-left) we show the relative amount of energy consumed by the two-node system while in the CPU, transmitting and receiving states. Note that energy spent while receiving represents the major component. This is due to the fact that receiving (or idle listening) is more energy costly than transmitting; furthermore, prior to each packet transmission, the data that are to be sent have to be copied from the microcontroller to the radio transceiver: during this operation the radio of sensors is in the listening state, and as a result the time spent while listening is greater than the one spent transmitting (see [20] for additional details). On the top-right side of Figure 3 we consider instead the utility based adaptive frequency hopping algorithm proposed in [10]: the total energy consumption is in this case increased by approximatively 4%. This is due both to longer listening and transmitting times (note that nodes add to each transmitted packet a two-byte field for synchronization purposes) as well as to the higher computational complexity of the adaptive procedure which results in increased CPU energy consumption. The adaptive frequency hopping algorithm described in Section III-B basically presents the same energy performance as traditional FH (since the adaptation procedure we implemented is on demand, and in absence of interference no adaptation is performed, see the bottom-left plot in Figure 3). Finally, if the single channel approach is selected, the overall energy consumption is reduced by approximatively 7%: this is due to lower complexity (no generation of random numbers is performed) as well as to the fact that nodes do not need to switch frequency band after transmitting/receiving packets and acknowledgements[1].

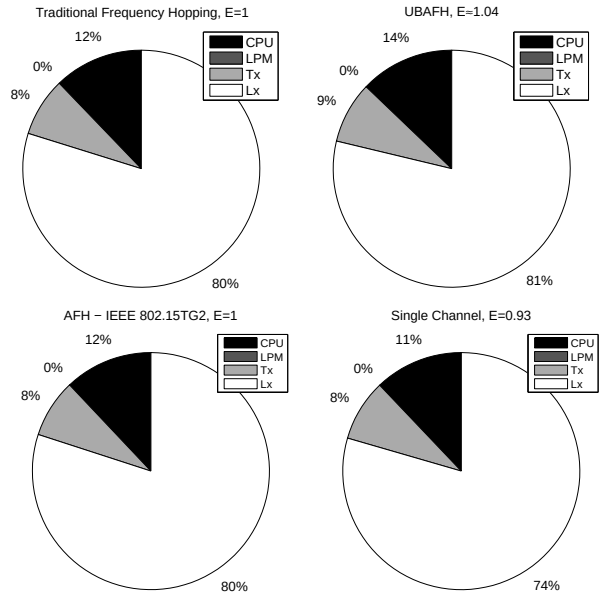


Fig. 3. Relative energy consumption in the four different energy states of the two node system for FH, UBAFH, AFH and Single Channel scheme in interference free conditions. Note that percentages of UBAFH and of the Single Channel scheme sum up to 104% and 93% respectively since we normalized the obtained values to the energy consumption of FH.

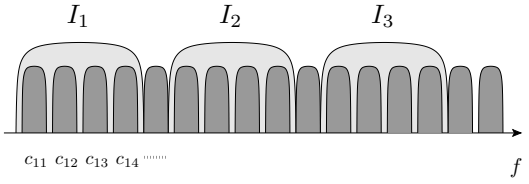


Fig. 4. Interference Scenario. Each interfering carrier I_i induces a certain packet error probability over the overlapping IEEE 802.15.4 channels.

Let us now start our performance evaluation. Using the methodology described in Section II we emulated the presence of three WLAN carriers (I_1, I_2, I_3, see Figure 4) overlapping

[1]For the CC2420 radio unit, channel switching time is in the order of 200 μs [19] and it is equivalent to the time required to transmit about 50 bits

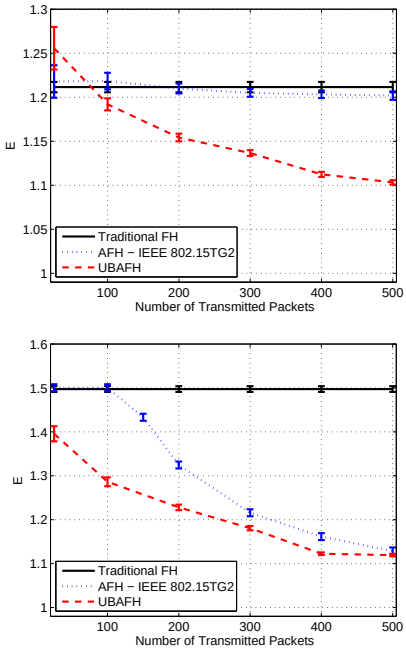


Fig. 5. Results for Scenario 1. Average energy per packet for $p = 0.4$ (top) and $p = 0.8$ (bottom).

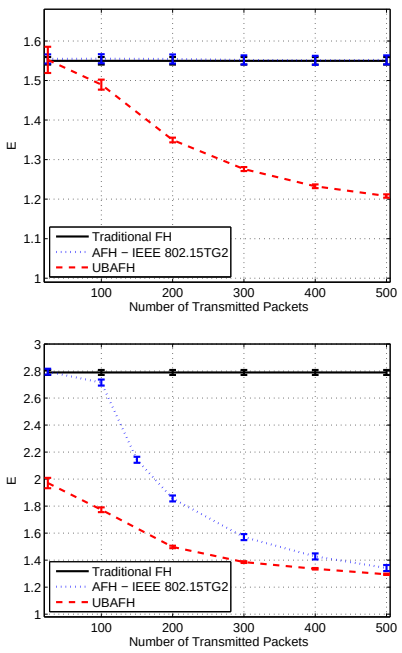


Fig. 6. Results for Scenario 2. Average energy per packet for $p = 0.4$ (top) and $p = 0.8$ (bottom).

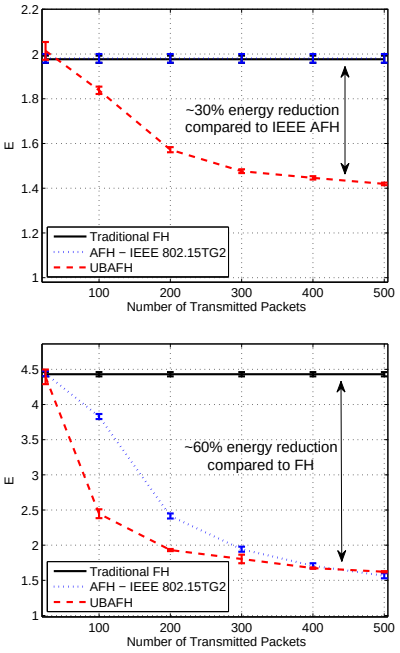


Fig. 7. Results for Scenario 3. Average energy per packet for $p = 0.4$ (top) and $p = 0.8$ (bottom).

with the channels used by sensors and we varied the packet error rate experienced over those channels. In particular we considered the following scenarios:

1) Scenario 1: only one WLAN carrier (I_1) is active. This overlaps with the IEEE 802.15.4 channels $11 - 14$;
2) Scenario 2: two WLAN carriers (thus both I_1 and I_2) are active. I_1 overlaps with channels $11 - 14$, while I_2 overlaps with channels $16 - 19$.
3) Scenario 3: all three WLAN carriers are active. These overlap respectively with channels $11 - 14$, $16 - 19$ and $21 - 24$.

For each of these scenarios we run our experiments for two different settings. In the first one, the packet error probability induced by the WLAN carriers is set to $p = 0.4$ while for the latter we consider $p = 0.8$: these two values are respectively below and above the threshold packet error probability used by the channel classification procedure defined by IEEE AFH (see Section III-B). In all cases, we considered symmetric channel conditions at the two nodes i.e. nodes experience equal packet error probabilities on the same channel.

Results for Scenarios 1, 2 and 3 are respectively presented in Figures 5, 6 and 7, where we show as a function of the amount of transmitted data N the average energy per packet E for the three hopping schemes. 95% confidence intervals are also plotted in all curves. While the energy performance of traditional frequency hopping do not significantly depend on the amount of transmitted data, the other algorithms can benefit from adaptation and in fact transmitting a larger amount of packets allows to improve energy efficiency. It should be remarked how the different adaptive approaches implemented by the two schemes we considered lead to different energy performance. The traditional adaptive algorithm proposed in [6], makes use of an ineffective channel classification procedure that allows the adaptation process to start only after a significant number of packets has been transmitted. We further remark that the use of a *binary* approach, where channels are either used for hopping or completely removed from the hopset, is very sensitive to the choice of the used threshold. Over a set of channels presenting only frequency static interference and for an appropriate packet error rate threshold, this adaptive strategy provides the best performance since nodes hop only over clear frequency bands: however energy efficiency can easily be deteriorated if the value of p^{Max} is not properly chosen. In our experiments, where we selected on purpose an improper threshold, a packet error probability equal to 0.4 was sporadically allowing to classify the considered frequency bands as interfered, preventing the algorithm from adapting. This behavior can potentially be improved by lowering the threshold used by the channel classification procedure, however the same problem might arise also with a lower value of p^{Max} if on some of the available channels nodes experience a packet error rate that is just slightly below the new threshold.

The probabilistic approach adopted by UBAFH overcomes these limitations. As shown by the plotted curves, adaptation can start as soon as a few packets are transmitted: this results in lower energy consumption. Moreover, the implemented algorithm allows a more granular exploitation of the available resources if compared to the binary strategy implemented by IEEE AFH. This is clearly shown by the energy performance in presence of low interfering activities (top plots of Figures 5, 6 and 7): channels experiencing low (but still significant) packet error rates are used less frequently than not-interfered channels and this allows to reduce energy

 ISBN: 978-1-61208-100-7

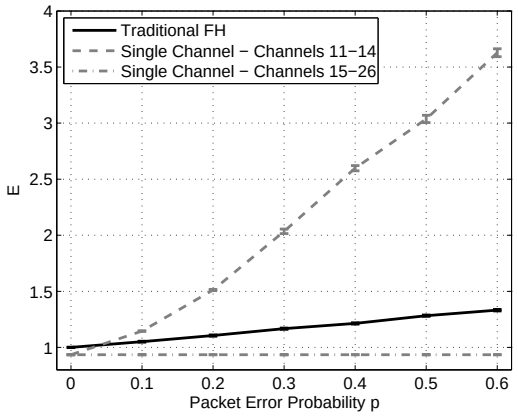


Fig. 8. Average Energy per packet as a function of the experienced packet error probability for frequency hopping and single channel scheme. Results are obtained considering $N = 100$ packets.

consumption of up to 30% if compared to IEEE AFH.

In order to compare the energy performance of frequency hopping against the one of a communication scheme making use of a single channel we performed a very simple experiment: we activated the WLAN carrier overlapping with the first four 802.15.4 channels (I_1 with reference to Figure 4) and varied the packet error probability experienced by sensors in the range $[0, 0.6]$. The average energy per packet for frequency hopping and single channel approach are shown in Figure 8. Note that if nodes operate over an interfered frequency, packet error probabilities as low as 10% are already sufficient to justify the use of frequency hopping proving that the overhead introduced by channel hopping is relatively small.

B. Real Environment

Results obtained in a real and uncontrolled wireless scenario validate the considerations made in the previous sub-section. Average energy per packet for FH, IEEE AFH and UBAFH are shown in Figure 9: the two plots are obtained considering the transmission of bulks of data consisting of 100 (top) and 500 (bottom) packets. For each algorithm we performed 200 experiments per day, 100 between 10 to 12 AM and 100 between 2 to 4 PM: during these hours the 2.4 GHz ISM band was mainly used by WiFi devices. For a relatively small amount of data, the three algorithms basically perform in the same way and lead to similar energy consumptions. However, while more and more packets are transmitted, adaptation plays an important role as shown in the bottom plot of Figure 9. IEEE AFH basically fails in identifying bad channels (in fact, only in a few cases we observed during the channel classification phase a packet error rate greater than the fixed threshold): these are consequently kept in the hopset and used as often as the good ones. The approach implemented by UBAFH instead allows to progressively decrease the probability of selecting frequency bands where sensors experience bad conditions and this results in lower energy consumption.

We finally compare always in the office spaces of the radio communication systems department of KTH the performance of the considered frequency hopping techniques against the one of a communication scheme making use of a single

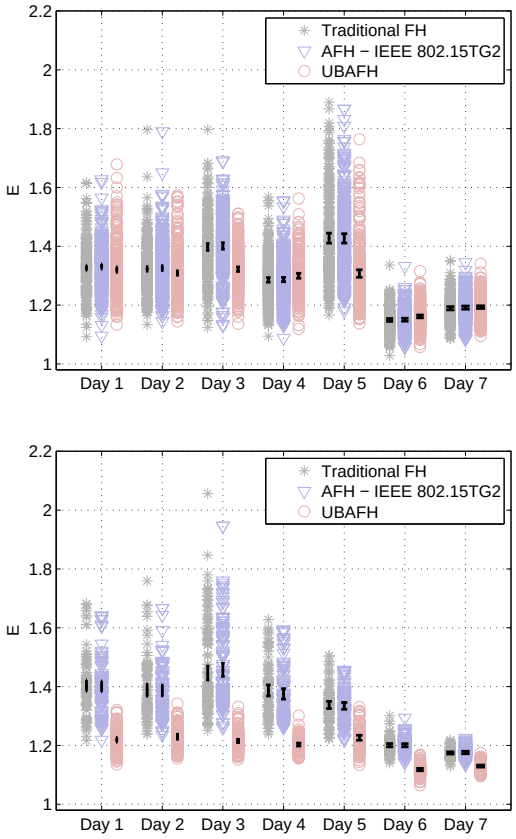


Fig. 9. Energy performance for traditional FH, IEEE AFH and UBAFH in a real environment. Results have been obtained considering 100 (top) and 500 (bottom) packet transmissions. 95% confidence intervals are also shown. Note that day 6 and 7 correspond to Saturday and Sunday

channel. Results for this scenario are presented in Figure 10 where we show the average energy per packet for the different communication schemes. For the single channel case, energy values are presented for all the 16 available frequency bands. For each of them, we performed 25 experiments per day (between 2 to 4 PM) and repeated these experiments on 7 different days. Note that on channels that overlap with the WiFi carriers used for internet access in the environment of our evaluation (see Figure 2), channel conditions can be extremely bad and energy consumption can be increased of up to 6 times. The use of frequency hopping allows to mitigate these problems by *averaging* channel conditions and reducing the high energy consumption that nodes experience in the worst case single-channel scenario.

We stress the importance of this last observation: recently published works (see for instance [21]) have questioned the utility of frequency hopping schemes in real environments pointing out that in typical settings, when multiple channels are available, it is likely that there is a non-empty set of clear and not interfered frequency bands. We remark however that while identifying those channels by means of dedicated spectrum sensing algorithms might be quite straightforward [22], the energy overhead introduced by this procedure might be significant and could be equivalent to the energy required to

 ISBN: 978-1-61208-100-7

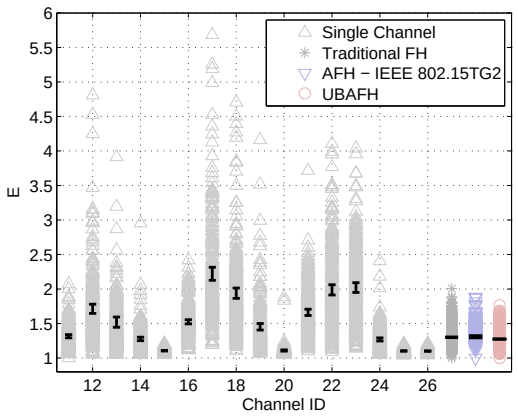


Fig. 10. Average Energy per packet for the single channel approach and for the three considered frequency hopping schemes. Results are obtained transmitting $N = 100$ packets. 95% confidence intervals are also shown.

transmit several tens of packets (see [17]). Thus, even though reliable single-channel communication is indeed possible (note for instance that in Figure 10, if channels 15, 20, 25 or 26 are selected, average energy per packet is lower than the one achieved by the frequency hopping scheme), the use of frequency hopping allows to limit the performance degradation that can potentially be induced by interference and at the same time permits to avoid the energy overhead introduced by the spectrum sensing algorithms that nodes have to run in order to identify clear channels.

V. Conclusions

In this paper we presented the experimental evaluation of different frequency hopping techniques for wireless sensor networks. Using TMote Sky sensor nodes operating within the IEEE 802.15.4 radio standard we implemented a frequency hopping algorithm and two adaptive schemes and quantified their complexity and energy performance under different channel settings. Our results have shown that traditional frequency hopping schemes, where channels are used in a pseudo-random fashion are very sensitive to interference that can severely degrade their energy efficiency. On the other hand, the utility based adaptive frequency hopping algorithm we recently proposed, introduces significant computational complexity but allows to effectively adapt the hopping pattern in presence of bad channel conditions. In our experiments this led to energy savings of up to 60 percent if compared to non-adaptive schemes and as high as 30 percent if compared to the other and more traditional adaptive approach that was considered in our evaluation. Comparison with a traditional communication scheme using a single channel has also outlined that frequency hopping is very useful in presence of interference and can be exploited in order to limit the performance degradation induced by transmissions of collocated wireless devices.

Promising directions for future work might include the extension of our experimental campaign for investigating the behavior of the considered hopping schemes in presence of frequency dynamic interference and propagation anomalies (such as the multi-path or frequency selective fading that might for instance arise in industrial settings). It could also be interesting to evaluate the effectiveness of the considered hopping techniques on networks comprising more than two nodes.

References

[1] J. Morse, “Market Pulse: Wireless in Industrial Systems: Cautious Enthusiasm“, Embedded Systems, Winter 2006.
[2] “WSN for Smart Industries: A Market Dynamics Report“, OnWorld, September 2007.
[3] “Part 15.1: Wireless medium access control (MAC) and physical layer (PHY) specifications for wireless personal area networks (WPANs)“, ANSI/IEEE Standard 802.15.1-2005.
[4] “Why Wireless HART? The Right Standard at the Right Time“, white paper, October 2007. Available online at www.hartcomm2.org.
[5] “The ISA100 Standards - Overview & Status“, October 2008, Available online at http://www.isa.org.
[6] “Part 15.2: Coexistence of Wireless Personal Area Networks with other Wireless Devices Operating in Unlicensed Frequency Bands“, ANSI/IEEE Standard 802.15.2-2003.
[7] P. Popovski, H. Yomo, and R. Prasad, “Strategies for Adaptive Frequency Hopping in the Unlicensed Bands“ in IEEE Wireless Communications, Vol. 13, No. 6, December 2006.
[8] M. C.-H. Chek and Y.-K. Kwok, “Design and Evaluation of Practical Coexistence Management Schemes for Bluetooth and IEEE 802.11b Systems“, in Computer networks, Vol. 51, Issue 8, June 2007.
[9] A. C.-C. Hsu, D. S. L. Wei, C.-C. J. Kuo, N. Shiratori, and C-Ju Chang, “Enhanced Adaptive Frequency Hopping for Wireless Personal Area Networks in a Coexistence Environment“, in Proceeding of Global Telecommunications Conference (GLOBECOM), 2007.
[10] L. Stabellini, L. Shi, A. A. Rifai, J. Espino, and V. Magoula, “A New Probabilistic Approach for Adaptive Frequency Hopping“, in Proceedings of International Symposium on Personal Indor and Mobile Radio Communications, PIMRC, 2009.
[11] K. J. Park, T. R. Park, C. D. Schmitz, and L. Sha, “Entropy-Maximization Based Adaptive Frequency Hopping for Wireless Medical Telemetry Systems“, in Proceedings of the 1st ACM International Workshop on Medical-Grade Wireless Networks, 2009.
[12] J. So and N. Vaidya, “Multi-Channel MAC for Ad Hoc Networks: Handling Multi-Channel Hidden Terminals Using A Single Transceiver“, in Proceedings of the Fifth ACM International Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc), 2004.
[13] N. Golmie, O. Rebala, and N. Chevrollier, “Bluetooth adaptive frequency hopping and scheduling“, in Proceedings of Military Communications Conference (MILCOM), Boston, USA, October 2003.
[14] K. J. Park, T. R. Park, C. D. Schmitz, and L. Sha, “Design of Robust Adaptive Frequency Hopping for Wireless Medical Telemetry Systems“, in IET Communications, Vol. 4, No. 2, 2010.
[15] A. Dunkels, B. Grönvall, and T. Voigt, “Contiki - a Lightweight and Flexible Operating System for Tiny Networked Sensors“, in Proceedings of the IEEE Workshop on Embedded Networked Sensors (Emnets-I), 2004.
[16] M. Rossi, G. Zanca, L. Stabellini, R. Crepaldi, A. F. Harris III, and M. Zorzi, “SYNAPSE: A Network Reprogramming Protocol for Wireless Sensor Networks using Fountain Codes“, in Proceedings of the IEEE Conference on Sensor, Mesh and Ad Hoc Communications and Networks (SECON), 2008.
[17] L. Stabellini and M. U. Javed, “Experimental Comparison of Dynamic Spectrum Access Techniques for Wireless Sensor Networks“, in Proceedings of Vehicular Technology Conference (VTC Spring), 2010.
[18] A. Dunkels, F. Österlind, N. Tsiftes, and Z. He, “Software-Based On-Line Energy Estimation for Sensor Nodes“, in Proceedings of the Fourth IEEE Workshop on Embedded Networked Sensors (Emnets IV), June 2007.
[19] ‘Tmote Sky Data Sheet‘ (2006) Moteiv, San Francisco, CA. Available online at: http://www.moteiv.com/products/docs/tmote-skydatasheet.pdf.
[20] F. Österlind and A. Dunkels, “Approaching the Maximum 802.15.4 Multi-Hop Throughput“, in Proceedings of the Fifth ACM Workshop on Embedded Networked Sensors (HotEmNets 2008), June 2008.
[21] J. Ortiz and D. Culler, “Multichannel Reliability Assessment in Real World WSNs“, in Proceedings of the 9th IEEE/ACM International Conference on Information Processing in Sensor Networks (IPSN), 2010.
[22] L. Stabellini and J. Zander, “Energy Efficient Detection of Intermittent Interference in Wireless Sensor Networks“, to appear in International Journal of Sensor Networks, 2010.

 ISBN: 978-1-61208-100-7

Impact of the Parameterization of IEEE 802.15.4 Medium Access Layer on the Consumption of ZigBee Sensor Motes

Eduardo Casilari, Jose M. Cano-García
Dpto. Tecnologia Electrónica
University of Malaga, Spain
ecasilari@uma.es, cano@dte.uma.es

***Abstract*—This paper presents an analysis of the impact that the parameterization of the CSMA/CA algorithm (employed by IEEE 802.15.4 Medium Access Layer) has on the consumption of ZigBee motes. For this purpose, the study introduces an analytical model that permits to compute the mean drain current of low duty cycle sensor motes. The results show that the energy required by the re-association process required after a packet loss cannot be neglected when setting the values of CSMA/CA parameters.**

Keywords-IEEE 802.15.4, ZigBee, CSMA/CA

I. INTRODUCTION

IEEE 802.15.4 (which describes the Physical Layer and Medium Access Control [1]) and ZigBee [2] specifications define a protocol stack for the development of short-range and low power communications for Wireless Personal Area Networks (WPANs) and Wireless Sensor Networks (WSNs). These protocols are basically intended to provide networking solutions for low-bandwidth sensor devices.

The low-cost and simplicity of IEEE 802.15.4/ZigBee compliant motes, together with their capability to configure self-organizing networks, has made this technology an attractive choice for a wide set of applications including domotic systems, health telemonitoring or industrial plant-process control.

IEEE 8021.15.4/ZigBee networks typically consist of a set of battery-powered sensor nodes ('motes'), which periodically (or sporadically) send their sensed data to one or several data sinks. To maximize the nodes' battery lifetime, the activity of the nodes' radio transceivers must be reduced so they remain most of the time in a low-power ('sleep') state. The idea is that the transceiver only has to 'wake up' (to be active) in order to sense and transmit (or receive) the data for a small fraction of time.

Many possible advantages of employing IEEE 802.15.4 are strongly affected by the configuration of the Medium Access Control (MAC) sublayer. IEEE 802.15.4 MAC employs CSMA/CA (Carrier Sense Multiple Access/Collision Avoidance) to regulate the participation of the nodes in the network. The 802.15.4 specification permits the setting of some parameters of the CSMA/CA contention algorithm (although default values are proposed). This paper investigates the effects of this parameterization on the power consumption of the motes. With this goal, an analytical model is proposed to characterize the mean drain current in an IEEE 802.15.4 node as a function of the CSMA/CA parameters, the node duty cycle and the radio transmission conditions. In contrast with other works in the literature, the model pays a special attention to the power required by the re-association procedures that packet losses provoke.

This paper is organized as follows: Section II reviews the CSMA/CA algorithm, which defines the behavior of the 802.15.4 MAC. Section III reviews the existing works that analyze the effects of the parameters of CSMA/CA on the performance of 802.15.4 networks. Section IV presents the model that computes the mean consumption of an 802.15.4 node. Section V discusses the results obtained with the proposed model. Finally, Section VI summarizes the main conclusions of the paper.

II. REVIEW OF CSMA/CA ALGORITHM

IEEE 802.15.4 standard discriminates two classes of nodes: Full-Function Devices (FFD) and Reduced-Function Devices (RFD). FFDs can assume the role of network 'Coordinators' and be in charge of the communications of a set of nodes (the 'children' or 'leaf' nodes) according to a star or a cluster-tree topology. Conversely the role of RFDs (devised for very simple 'motes' with limited resources) just allows the node to communicate (as an 'end' node) with a single FFD acting as its Coordinator. Simultaneously, the MAC layer of IEEE 802.15.4 offers two alternative operational modes: (1) Under the beacon mode, the Coordinator periodically broadcasts a special frame (a 'beacon'), which announce the existence of the Coordinator (and the corresponding WPAN) while enabling the synchronization of the children nodes. The beacon informs the children if they have any pending packet. If this is not the case and the children have not data to send (or after sending the data), both the children and the Coordinator can enter a sleep (low-consumption) mode. (2) Under the beaconless mode (which is massively implemented in commercial 802.15.4 motes as it avoids the need for synchronization with the Coordinator), the children can wake up from the sleep mode in any moment to send (or ask for) data. This obliges the Coordinator to be active at any time. Beacon mode is recommendable when the Coordinator (or the intermediate routers in a multi-hop cluster-tree) is powered by batteries. Conversely non-beacon mode typically suits applications which can be deployed by a simple star topology formed by a set of wireless sensors and a Coordinator powered from the main source

The Medium Access Control (MAC) in beaconless 802.15.4 networks is governed by non-slotted CSMA/CA. According to this protocol, nodes desiring to transmit a packet have to wait a random time chosen between 0 and $(2^{BE}-1)$ backoff periods. A backoff period is 0.32 ms, which

 ISBN: 978-1-61208-100-7

is the time corresponding to 20 symbols (the duration of a symbol is 16 µs when the nodes operate in the 2.4 GHz band, with a rate of 250 kbps or 64.5 Ksymbols/s). *BE* is the Backoff Exponent, an increasing variable that regulates the limit of the CSMA waiting times. Its initial value is set by the parameter *macMinBE*. Once this random time is elapsed, the node checks the availability of the radio medium through a Clear Channel Assessment (CCA). If the channel is detected to be busy, the exponent *BE* is incremented by 1 (up to a maximum value *macMaxBE*) and a new random waiting time is chosen and executed before performing the next CCA. This process can be repeated *macMaxCSMABackoffs* times. Thus, if the CCA fails *macMaxCSMABackoffs*+1 consecutive times, a channel access failure is assumed to have occurred, the packet is dropped and the transmissions concludes. Otherwise, if a CCA succeeds, the channel is considered to be free and the node switches its radio transceiver from the receiver state to the transmitter state. For this purpose a turnaround time of 0.192 ms (12 symbols) is reserved. Then, the device proceeds to transmit the packet and (optionally) waits for an acknowledgment (ACK) message from the receiving node (after switching again the radio transceiver from the transmission to the reception mode). CSMA/CA wait is not accomplished for the sending of an ACK, so that the receptor sends the acknowledgement as soon as it receives the packet. However, the transmitted packet or the ACK message can experience a collision due to interferences, reflections, shadowing effects or the activity of other nodes in the same 802.15.4 network. So, if the ACK is not received in a predetermined period, the node retransmits the packet after executing the aforementioned backoff algorithm of CSMA (resetting the initial value of BE to *macMinBE*). The number of times that a packet can be retransmitted is limited by the parameter *macMaxFrameRetries*. Thus, a sending failure is assumed after transmitting the packet *macMaxFrameRetries* +1 times without receiving the corresponding ACK.

Consequently the dynamics of the MAC layer of 802.15.4 standard heavily depend on these four parameters: *macMaxCSMABackoffs*, *macMaxFrameRetries*, *macMinBE* and *macMaxBE*, which are set to constant values in the nodes. The ranges and default values recommended by the standard for these parameters are tabulated in Table I.

TABLE I. ALLOWED RANGES FOR 802.15.4 MAC PARAMETERS

Parameter	***Range***	***Default Value***
macMinBE	[0-7]	3
macMaxBE	[3-8]	5
macMaxFrameRetries	[0-7]	3
macMaxCSMABackoffs	[0-5]	4

III. RELATED WORK

The effects of the MAC parameter setting on the performance of 802.15.4 networks have been recently studied by different research papers.

The study in [3] compares the reliability of 802.15.4 cluster-trees when three different sets of values are employed to define the parameters *aMacFrameRetries*, *macMaxCSMAbackoffs*, *macMaxBE* and *macMinBE*. By means of simulations with NS-2 Network Simulator tool, the study shows that 802.15.4 cluster-trees may severely underperform if the default set of parameters is utilized. The performance is computed in terms of packet delivery ratio, message latency and energy consumption per node. The same authors present similar conclusions in [4]. In this case the study, based on both simulations and some experimental results in a real testbed, are focused on single-hop topologies. Both studies employ the battery consumption model of a CC2420 radio transceiver but assuming that nodes remain in the sleep mode during the backoff periods (which is not true in actual 802.15.4 motes as recovering from this state requires a non-negligible time).

The influence of the parameter *macMaxFrameRetries* (called maximum number of retransmission times) on the throughput, packet delivery ratio and energy consumption in 802.15.4 beacon enabled networks is examined in [5]. The study proposes a Markovian chain model to characterize the performance of the network although most analysis are based on simulations with NS-2. The study concludes that a low value for *macMaxFrameRetries* reduces the power consumption. As the traffic load increases, if just one packet attempt is permitted, the throughput increases.

In [6] authors evaluate the impact of the parameterization of *macMaxCSMAbackoffs*, *macMaxBE* and *macMinBE* on the packet loss probability and packet latency in 802.15.4 beaconless 802.15.4 star topology (under unslotted CSMA/CA). The analysis (which considers different traffic loads) is also based on NS-2 simulations. The paper does not evaluate the performance in terms of battery consumption.

Most of these papers address the problem of the scalability of 802.15.4 networks. The employed simulation or analytical model normally assumes that channel occupation and packet collisions are uniquely due to the activity of other 802.15.4 sensor motes. The network scale is evaluated considering an elevated concentration of nodes in the same transmission area under relatively heavy traffic load, which are not always the actual application scenario for a ZigBee network.

In most implementations of the transceivers used for 802.15.4 networks, devices operate in 2.4 GHz ISM band. Thus 802.15.4 communications are exposed to the interferences of other popular standards such as Bluetooth and especially 802.11 (Wi-Fi). The effects of these interferences are becoming more unavoidable [7] with the expansion of the new versions of IEEE 802.11 (such as 802.11n), which employ a higher bandwidth (two channels of 20 MHz instead of the 20 MHz single channel of the previous versions). Thus, packet collisions and channel occupancy in many 802.15.4/ZigBee networks which just require a few sensors (e.g. some biosensors belonging to the WPAN of the same patient) may be basically determined by external interferences. In those scenarios network scalability is not the most relevant issue when analyzing the performance of 802.15.4 technology. The power required by

 ISBN: 978-1-61208-100-7

the sensors will be mainly linked to the operations executed by a single node during the duty cycle and the transmission conditions imposed by the interferences.

In any case, the aforementioned studies do not take into consideration that packet losses may oblige the sensor nodes to re-associate with the Coordinator. This operation of re-association must be considered to compute the mean drain current in the nodes as they may introduce an important extra consumption in the case of frequent losses.

IV. ANALYTICAL MODEL FOR BATTERY CONSUMPTION

In this section we offer an analytical expression that permits to compute the main current drained in a sensor mote which periodically sends a data to the Coordinator.

In our analysis we consider that no polling takes places so that the only existing data traffic is upstream (i.e. from the mote to the Coordinator). The current required for the initial start-up phase is also ignored. Similarly, we assume that the power required by sensing (data acquisition and processing) can be neglected when compared with the current drained by wireless communications. However, studies such as [8] reveal that (depending of the employed sensor and the sampling frequency) the sensing process may suppose an important part of the battery consumption in the wireless mote. Under this assumption, the battery consumption basically depends on the state of the radio transceiver. In general terms, for most commercial 802.15.4-enabled motes, four states are possible: transmission, listening, idle (during CSMA/CA backoffs and turnaround time) and sleep states (for which the consumption is minimized).

Basing on the drain current and the time spent in these states, we can estimate the mean current that must be supplied to the mote to transmit a packet of n bytes flowing from the application layer:

$$I_{active}(n) = \frac{t_{onoff} I_{onoff} + t_{listening} I_{listening} + t_{tx}(n) I_{tx} + t_{idle} I_{idle}}{t_{act}(n)} \quad (1)$$

where $t_{listening}$ ($I_{listening}$), $t_{tx}(n)$(I_{tx}) and t_{idle} (I_{idle}) are the mean time (and mean current) that the mote requires in the listening, transmission and idle states, respectively, to transmit the n user data bytes. Besides t_{onoff} and I_{onoff} are the total time and current necessary to wake up and turn off the transceiver as well as to transmit the data from/to the processing unit (e.g: the microcontroller) connected to the transceiver. Finally, $t_{act}(n)$ indicates the time of the complete activity period:

$$t_{act}(n) = t_{onoff} + t_{listening} + t_{tx}(n) + t_{idle} \quad (2)$$

If T is the update period of the data (i.e. the time between two consecutive transmissions of the sensed magnitudes) we have that the mean current at which the battery is drained is:

$$I_{drain}(n) = \frac{t_{act}(n)}{T} I_{active}(n) + \left(1 - \frac{t_{act}(n)}{T}\right) \cdot I_{sleep} \quad (3)$$

where I_{sleep} is the current in the sleep mode while the term $\left(\frac{t_{act}(n)}{T}\right)$ actually represents the duty cycle of the mote.

The drain current in the different states, as well as the time t_{onoff}, are determined by the particular mote that is being utilized. Conversely, the times in the different states can be calculated as a function of the data size (n), the dynamics imposed by CSMA/CA algorithm (illustrated in Figure 1) and the frequency of the collisions and the channel access failures.

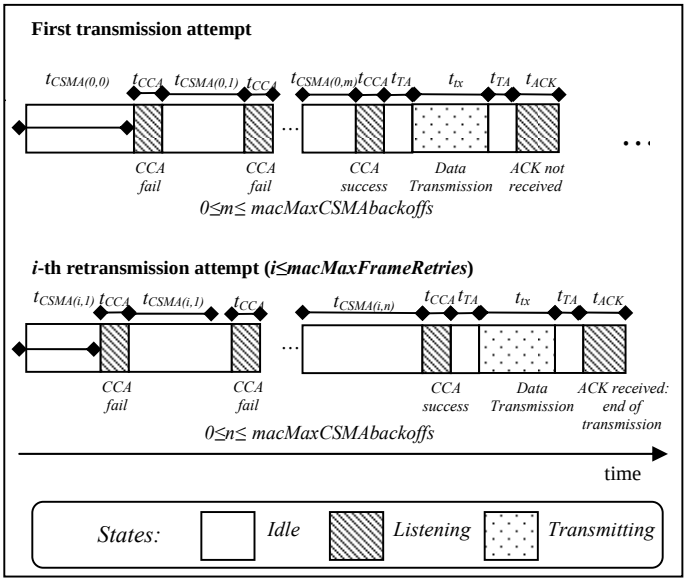


Figure 1. Timing and evolution of the transceiver state during the application of the CSMA/CA algorithm

In our analysis both processes (access failures and collisions) are assumed to follow independent and self-uncorrelated stochastic processes) [1]. In particular, if p_o denotes the probability that the channel is occupied when the CCA operation is performed and p_c is the probability of a packet collision (i.e. the probability that a packet is not acknowledged after being transmitted), we have that the average listening time (i.e. the periods in which the sensor performs a CCA or waits for an ACK) required to transmit a packet can be computed as:

$$t_{listening} = \sum_{i=0}^{mMaxF} (1 - p_{CSMAfail})^i \cdot p_c^i \cdot \cdot \left\{ p_{CSMAfail} \cdot mMaxb \cdot t_{CCA} + (1 - p_{CSMAfail}) \cdot \left(\overline{n}_{CCA} \cdot t_{CCA} + t_{ACK} \right) \right\} \quad (4)$$

where:

-$mMaxF$ is *macMaxFrameRetries* (the maximum number of times that a transmission can be retried)

[1] Authors in [9] offer an analytical expression to compute the probabilities p_o and p_c as a function of the number of nodes contending in the 802.15.4/ZigBee network. The expression does not take into account the presence of other interfering sources. See [10] for an empirical characterization of the bit error probability as a function of the received power.

 ISBN: 978-1-61208-100-7

-*mMaxb* is *macMaxCSMABackoffs*, the maximum number of times that the CSMA algorithm is repeated before a CCA failure is considered.
-t_{ACK} is *macAckWaitDuration*, the maximum time (0.864 ms or 54 symbols) that the receiver waits for the ACK before proceeding with the next attempt.
-$p_{CSMAfail}$ defines the probability of suffering a channel access failure (after *mMaxb*+1 failed CCA operations):

$$p_{CSMAfail} = p_o^{mMaxb+1} \tag{5}$$

- $\overline{n}_{CCA}$ is the mean number of CCA operations which are executed in an attempt that does not finish in a channel access failure. It can be computed as:

$$\overline{n}_{CCA} = \left(\frac{1-p_o}{1-p_{CSMAfail}} \right) \sum_{i=0}^{mMaxb} p_o^i (i+1) \tag{6}$$

Similarly, the time in the idle state (t_{idle}) imposed by the CSMA waits and the turnaround time can be computed as:

$$t_{idle} = \sum_{i=0}^{mMaxF} (1-p_{CSMAfail})^i \cdot p_c^i \cdot \cdot \left\{ p_{CSMAfail} \cdot t_{CSMAfail} + (1-p_{CSMAfail}) \cdot (t_{CSMAnofail} + t_{TA}) \right\} \tag{7}$$

where:
-t_{TA} is the turnaround time (0.192 ms or 12 symbols), reserved for the transceiver to switch from reception to transmission (in the opposite sense the turnaround time is included in *macAckWaitDuration*).
-$t_{CSMAfail}$ describes the mean time required by the (*mMaxb*+1) CSM/CA waits of a transmission attempt that concludes in a channel access failure (after (*mMaxb*+1) CCA failures):

$$t_{CSMAfail} = \sum_{i=0}^{mMaxb} \left(\frac{1}{2} (2^{\min(macMinBE+i, macMaxBE)} - 1) \cdot t_{backoff} \right) \tag{8}$$

being $t_{backoff}$ the duration of a backoff period (0.32 ms or 20 symbols)
-$t_{CSMAnofail}$ stands for the mean expected delay introduced by the CSMA/CA waits of an attempt that does not finish in a channel access failure (that is to say, an attempt with a successful CCA). This time can be computed [9] as:

$$t_{CSMAnofail} = \left(\frac{1-p_o}{1-p_o^{mMaxb+1}} \right) \sum_{i=0}^{mMaxb} p_o^i \cdot \cdot \left\{ \sum_{j=0}^{i} \left(\frac{1}{2} (2^{\min(macMinBE+j, macMaxBE)} - 1) \cdot t_{backoff} \right) \right\} \tag{9}$$

On the other hand, the mean time ($t_{tx}(n)$) that the radio transceiver is in the transmission state (for a packet payload of *n* data bytes) is:

$$t_{tx}(n) = \left(\frac{8 \cdot (O_H + n)}{r} \right) \cdot \sum_{i=0}^{mMaxF} (1-p_{CSMAfail})^{i+1} \cdot p_c^i \tag{10}$$

where *r* is the binary rate of 802.15.4 (250 kbps when operating at ISM 2.4 GHz band[2]) while O_H is the total packet overhead (preamble, frame delimiter, headers of MAC, Network and Application Sublayer and CRC field) of the 802.15.4/ZigBee data packet. For our study we assume that O_H is 31 bytes. Note that in the expression (10) the summation $\sum_{i=0}^{mMaxF} (1-p_{CSMAfail})^{i+1} \cdot p_c^i$ is the mean number of times that a packet is transmitted.

A. Effect of the node re-association

In the previous model, the mean drain current of the 802.15.4 node has been computed Assuming that nodes just associate to the 802.15.4/ZigBee network during the initial start-up. Consequently the battery consumption is only caused by the cyclic transmission of user data bytes. Thus, the presented equations neglect the current required by the exchange of messages that take place during the different phases of the star-up. These phases basically consist of the active scanning phase (to detect the presence of the Coordinator), the association phase to join the Coordinator WPAN and the ZigBee binding phase (which is necessary to connect compatible ZigBee endpoints at the application layer). However, in most cases, after a packet loss (induced by collisions or by a channel access failure), the node will try to re-associate with a Coordinator[3] (if the orphan scanning is not implemented or if the realignment command is not received after the orphan scan). This re-association process may take several seconds with a mean current consumption of more than 20 mA. Aiming at incorporating the extra consumption caused by the re-associations, the mean activity time needed to transmit a packet has to be recomputed as:

$$t_{act}(n) = t_{onoff} + t_{listening} + t_{tx}(n) + p_f \cdot t_{reassoc} \tag{11}$$

where $t_{reassoc}$ is the time required for the whole re-association process (including the binding and active scan phases) and p_f defines the probability of packet loss. This probability can be directly derived [9] from the probabilities of packet collision (p_c) and channel access failure ($p_{CSMAfail}$):

$$p_f = (1-p_{CSMAfail})^{aMaxF+1} \cdot p_c^{aMaxF+1} + + \sum_{i=0}^{aMaxF} \left(p_{CSMAfail} \cdot (1-p_{CSMAfail})^i \cdot p_c^i \right) \tag{12}$$

[2] The 2006 revision of the standard allows different modulations when the node works in the 868/915 MHz ISM bands. These new modulations permit to improve the bit rate up to 100 Kbps (for the 868 MHz band) and 250 Kbps (for the 915 MHz band). Conversely, in 802.15.4 devices operating in 2.4 GHz ISM band, the only permitted instantaneous bit rate is 250 kbps as long as just QPSK modulation (with 2 Megachip/s and 62.5 Ksymbol/s) is enabled

[3] It is up to the developer to define the number of losses that must take place before the device can be assumed to be orphan so that the MAC has to be reset and a new association procedure is triggered (or an orphaned device realignment procedure is performed). Our model assumes that any loss generates a re-association .

 ISBN: 978-1-61208-100-7

Similarly the mean current ($I_{active}(n)$) required to transmit a packet can be redefined as:

$$I_{active}(n) = \frac{t_{onoff} I_{onoff} + t_{listening} I_{listening} + t_{tx}(n) I_{tx} + t_{idle} I_{idle} + p_f \cdot t_{reassoc} I_{reassoc.}}{t_{act}(n) + p_f \cdot t_{reassoc}} \quad (13)$$

where $I_{reassoc}$ is a new term that defines the mean current required during the whole re-association phase.

Again the values of $I_{reassoc}$ and $t_{reassoc}$ rely on the employed mote but also on the number of scanned channels to detect the presence of the coordinator, the values of the CSMA/CA parameters and the probability of suffering packet losses during the re-association.

V. ANALYSIS OF THE IMPACT OF THE MAC PARAMETRISATION

In this section, we show and comment some numerical results obtained with the previous model. To compute these results we employ the battery consumption model of the Texas Instrument CC2480 ZigBee processor which we have presented in [11]. The CC2480 processor utilizes the Z-Stack of Texas Instrument, which is one of the most widely employed implementations of 802.15.4/ZigBee stack for the deployment of wireless sensor networks. In contrast with other chips that only implement an 802.15.4 transceiver, the CC2480 processor provides full ZigBee functionality, as far as it integrates the whole Z-Stack in a single chip. The current absorbed in the different states is summarized in Table II. The analyzed device keeps the transceiver in the listening mode during the idle periods (which is not the case of other commercial ZigBee motes), so the values of I_{idle} and $I_{listening}$ coincide. The Table also includes the mean drain current and time required by the re-association process. These values were measured in the most favorable case (without losses) in which the re-association always successes. The measurement of $I_{reassoc}$ and $t_{reassoc}$ were obtained in a network configured with the default values of the CSMA/CA parameters) so that they can be regarded as a rough approximation of the typical consumption during the re-association of a 802.154/ZigBee mote. In the presented results, in order to analyze the limit case in which association has the lowest impact, we assume that just one channel is scanned.

In the presented analysis we consider the typical case of a WSN formed by sensors with a low duty cycle and a low user data payload (2 bytes). In particular we tested our model with data rates lower than 1 packet per second which implies that duty cycle is always below 4% for all results (with a percentage of time in the transmission mode always under 0.25%). Unless the network is composed by hundreds of nodes within the same transmission range, a node will create very low interference in other nodes. Consequently packet losses will be provoked by 'external' factors (e.g.: Wi-Fi or Bluetooth interferences) which can be characterized by the probabilities p_o and p_c.

We firstly analyze the impact of the initial and maximum values of the backoff exponent (BE) by changing the limits *macMaxBE* and *macMinBE* (and configuring the other parameters with the default values). Results for three different values of p_o and p_c are depicted in Figures 2 and 3. As we assume that the radio conditions (modeled by p_o and p_c) are not affected by the contention CSMA/CA algorithm, the larger the values of *macMaxBE* and *macMinBE*, the higher the consumption. Obviously this is due to the fact that CSMA/CA random waits increase for higher values of *macMaxBE* and *macMinBE.* In any case the graphs show that the increase in the consumption is especially remarkable for extremely noisy environments. On the other hand, as the noise is reduced the results rapidly converge. This is especially true in the case where the modified parameter is *macMaxBE* as long as most transmission will be executed after the first CSMA wait (which is decided by *macMinBE*) and the effect of the parameter *macMaxBE* is minimized. Figures 2 and 3 do not include the consumption due to the re-association. The previous conclusions about the effects of *macMinBE* and *macMaxBE* are completely different if that consumption is added. Fig. 4, estimated for the case in which *macMinBE* is changed (results are similar for *macMaxBE*) shows that the impact of the election of both parameters is almost negligible.

TABLE II. SUMMARY OF DRAIN CURRENT FOR DIFFERENT 802.15.4/ZIGBEE OPERATIONS IN THE MOTE

Operation	State	Mean Required Current (mA)	Duration (ms)
Inactivity	Sleep mode	I_{sleep}=0.00075	Variable
Transmission of a packet of *n* bytes with sensed data	Transmission of a packet	I_{tx}=30.5 mA	Variable
	Listening (& idle)	$I_{listening}$= I_{idle} =32.5 mA	Variable
	Activation/deactivation of the ZigBee processor (radio transceiver is off)	I_{onoff}=13 mA	t_{onoff}=13
Association to the coordinator (without packet losses and default CSMA parameters)	Scanning in 1 channel	$I_{reassoc}$ =26.6 mA	$t_{reassoc}$ = 2000 ms
	Scanning in 16 channels	33.8 mA	up to 27500 ms

If we consider that p_o and p_c as processes that do not depend on the CSMA/CA dynamics, the parameters *macMaxBE* and *macMinBE* do not affect the probability of having a packet loss. On the contrary, the parameters *macMaxFrameRetries* and *macMaxCSMAbackoffs* clearly determine the loss probability (see equation (12)) and consequently the consumption provoked by the re-association. Figures 5 and 6 shows the impact on the drain current of the selection of the maximum number of transmission attempts (*macMaxFrameRetries*) when the other parameters are set as recommended by the specification. Results indicate that a low selection of *macMaxFrameRetries* (under the default value of 3) may dramatically impact on the consumption, in particular as the noise decreases. This is because in less noisy environments, 3 or 4 transmission attempts are enough to avoid the packet loss and, consequently, the cost of the re-association. The importance of the packet losses can be detected if we repeat the analysis without taking into account the battery

 ISBN: 978-1-61208-100-7

consumption provoked by the re-association. Results for this case are represented in Figure 7. Now, the value of *macMaxFrameRetries* seems to be irrelevant for environments with low noise. As the noise augments, the increase of *macMaxFrameRetries* increments the activity of the node and consequently the consumption.

The analysis of the impact of the parameter *macMaxCSMAbackoffs* (which can be observed from figures 8 and 9) offers a similar conclusion: if the number of maximum allowed CSMA waits is selected under the default value (4) the battery consumption is dramatically impacted due to the current needed by the frequent re-associations.

VI. CONCLUSIONS

This paper has investigated the effects of the parameterization of 802.15.4 MAC on the current required by IEEE 802.15.4/ZigBee sensor motes. The study is based on an analytical model that fully characterizes the dynamics of CSMA/CA algorithm. As a novelty the model also computes the power consumption provoked by the node re-association when packet loss occurs. Taking into account this extra component in the consumption, the obtained results seem to indicate that in typical WSNs (where sensors have a low duty cycle) the default values of the CSMA parameters *macMaxFrameRetries* and *macMaxCSMAbackoffs* proposed by the IEEE 802.15.4 specification exhibit a reasonable performance in terms of the expected battery lifetime.

ACKNOWLEDGMENT

This work was partially supported with public funds by the Spanish National Project No. TEC2009-13763-C02-01.

REFERENCES

[1] IEEE-TG15.4. Part 15.4: Wireless Medium Access Control (MAC) and Physical Layer (PHY) Specifications for Low-Rate Wireless Personal Area Networks (LR-WPANs), IEEE standard for Information Technology, 2003.

[2] ZigBee-Alliance. ZigBee specification, 2005.

[3] G. Anastasi, M. Conti, M. Di Francesco, and V. Neri, “Reliability and Energy Efficiency in Multi-hop IEEE 802.15.4/ZigBee Wireless Sensor Networks”, Proceedings of the IEEE Symposium on Computers and Communications (ISCC 2010), Riccione, Italy, June 22-25, 2010.

[4] G. Anastasi, M. Conti, and M. Di Francesco, “The MAC unreliability problem in IEEE 802.15.4 wireless sensor networks”, Proc. of ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems (MSWiM 2009), October 2009, pp. 196-203.

[5] P. K. Sahoo and J.P. Sheu, “Modeling IEEE 802.15.4 based wireless sensor network with packet retry limits”, Proc. of the 5th ACM symposium on Performance evaluation of wireless ad hoc, sensor, and ubiquitous networks (PE-WASUN’08), October 2008, pp. 63-70.

[6] D. Rohm, M. Goyal, H. Hosseini, A. Divjak, and Y. Bashir, "Configuring Beaconless IEEE 802.15.4 Networks Under Different Traffic Loads", Proc. of International Conference on Advanced Information Networking and Applications (AINA 2009), 2009, pp. 921-928.

[7] B. Polepalli, W. Xie, D. Thangaraja, M. Goyal, H. Hosseini and Y. Bashir, “Impact of IEEE 802.11n operation on IEEE 802.15.4 operation”, Proc. of International Conference on Advanced Information Networking and Applications Workshops (WAINA '09), 2009, pp. 328-333.

[8] C. Alippi, G. Anastasi, M. Di Francesco, and M. Roveri, "An Adaptive Sampling Algorithm for Effective Energy Management in Wireless Sensor Networks with Energy-hungry Sensors", IEEE Transactions on Instrumentation and Measurement, Volume: 58 Issue: 11, November 2009, pp. 335-344.

[9] M. Goyal, D. Rohm, H. Hosseini, K.S. Trivedi, A. Divjak, and Y. Bashir, “A stochastic model for beaconless IEEE 802.15.4 MAC operation”, Proc. of International Symposium on Performance Evaluation of Computer & Telecommunication Systems (SPECTS 2009), July 2009, Vol. 41, pp. 199- 207.

[10] B.Bougard, F. Catthoor, D.C. Daly, A. Chandrakasan, and W. Dehaene, “Energy Efficiency of the IEEE 802.15.4 Standard in Dense Wireless Microsensor Networks: Modeling and Improvement Perspectives”, Proc. of Design, Automation and Test in Europe (DATE'05), 2005, Vol. 1, pp.196-201.

[11] E. Casilari, G. Campos-Garrido, and J.M. Cano-García, “Characterization of Battery Consumption in 802.15.4/ZigBee Sensor Motes”, Proc. of IEEE International Symposium on Industrial Electronics (ISIE 2010), July 2010, pp. 3471-3476.

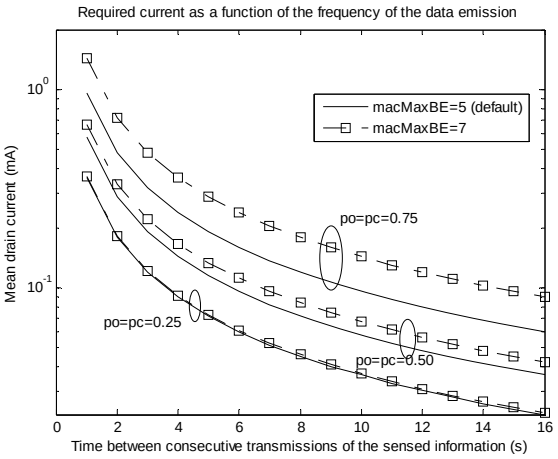


Figure 2. Mean drain current as a function of the frequency of data emission and two values of *macMaxBE.* Power consumption due to re-associations not considered

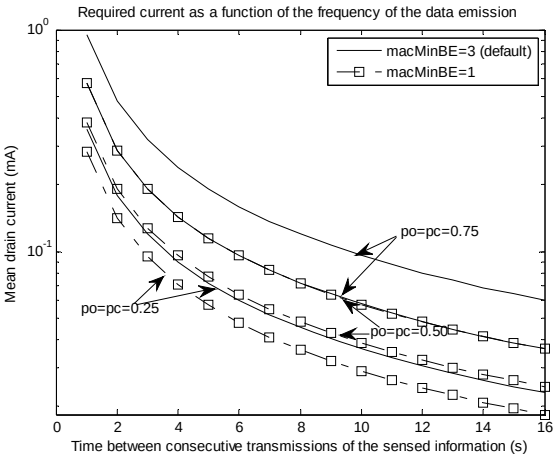


Figure 3. Mean drain current as a function of the frequency of data emission and two values of *macMinBE.* Power consumption due to re-associations not considered

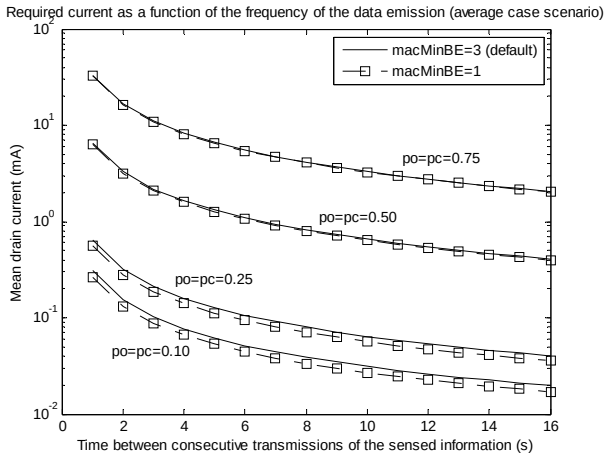


Figure 4. Mean drain current as a function of the frequency of data emission and two values of *macMinBE (*re-associations are considered)

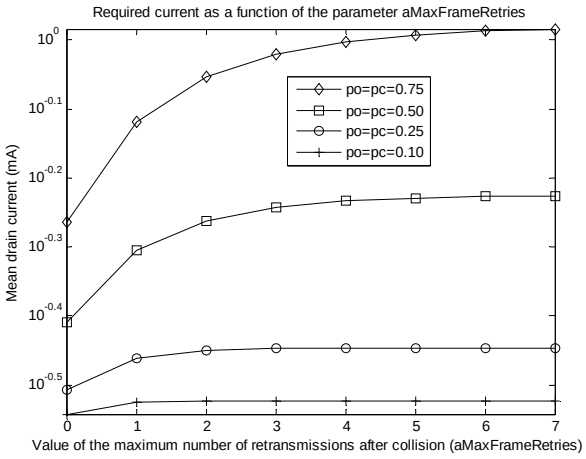


Figure 7. Mean drain current as a function of *macMaxFrameRetries* (rate=1 packet/s). Power consumption due to re-associations not considered

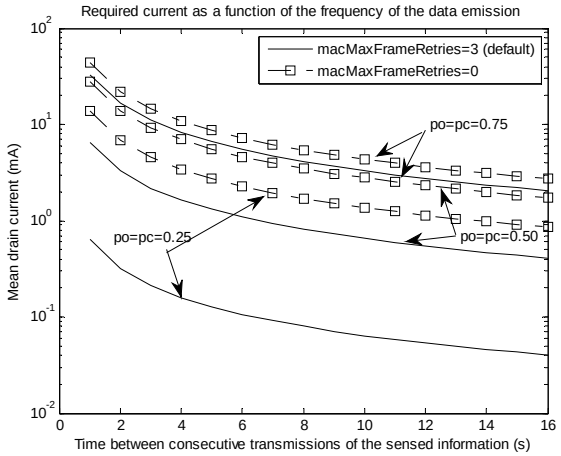


Figure 5. Mean drain current as a function of the frequency of data emission and two values of *macMaxFrameRetries (*re-associations are considered)

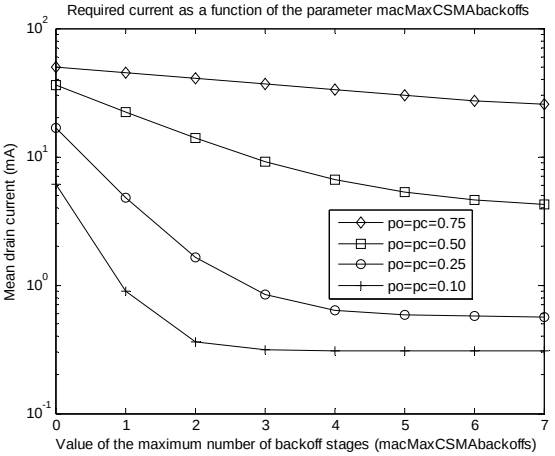


Figure 8. Mean drain current as a function of *macMaxCSMABackoffs* (rate=1 packet/s) *(*re-associations are considered)

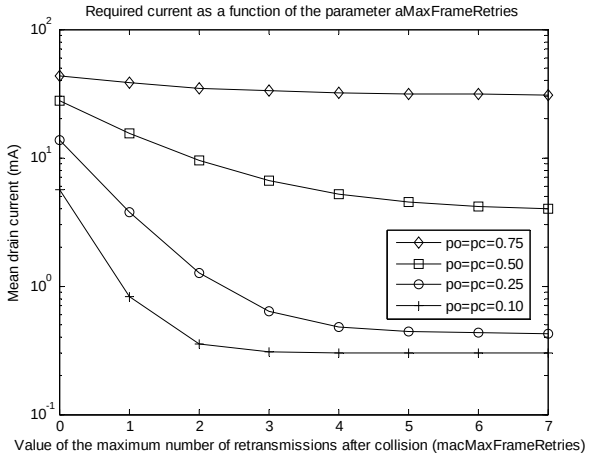


Figure 6. Mean drain current as a function of *macMaxFrameRetries* (rate=1 packet/s) *(*re-associations are considered)

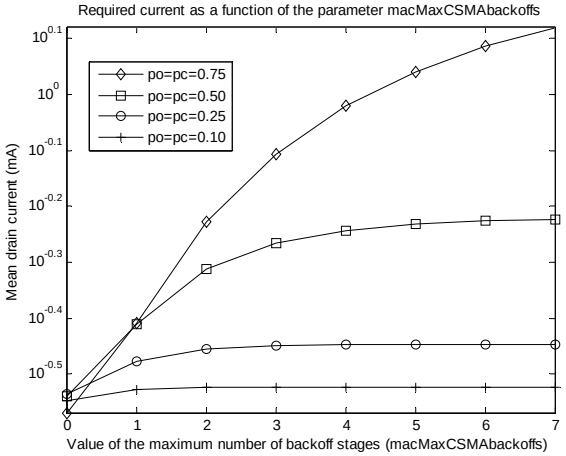


Figure 9. Mean drain current as a function of *macMaxCSMABackoffs* (rate=1 packet/s). Power consumption due to re-associations not considered

 ISBN: 978-1-61208-100-7

Game Theory-based Dynamic RRM for Reconfigurable WiMAX/WLAN System

Ognen Ognenoski [#1], Liljana Gavrilovska [#2]
[#] *Faculty of Electrical Engineering and Information Technologies, Ss Cyril and Methodius University Skopje, Macedonia*
[1] ognen@feit.ukim.edu.mk
[2] liljana@feit.ukim.edu.mk

***Abstract*— This paper proposes, analyzes and evaluates a novel user-oriented Radio Resource Management (RRM) strategies for coexisting WiMAX/WLAN reconfigurable system. The proposed strategies are based on game theory concepts, where games are dynamically played depending on the system condition with the applications on the terminal modeled as players. User's terminals have two interfaces, which can be used for enforcing reconfigurability by deploying RRM strategy that results in appropriate network selection for each application. The proposed RRM strategies are distributed and they map rationality behavior by adopting two game theory concepts: Nash and Quantal Response Equilibrium (QRE). These RRM strategies are compared with SNR based strategy. The effects of particular strategy are evaluated in terms of user's average throughput and the initiated application handovers. The results present guideline framework for distributed terminal management in reconfigurable systems.**

Keywords-Game theory; Reconfigurable systems; RRM

I. INTRODUCTION

In recent years, there are number of design and realization strategies for reconfigurable wireless systems due to their paramount importance in delivering 4G concepts [1]. Reconfigurable systems are envisioned as the future providers of rich multimedia content via various wireless links. The dynamic resource allocation strategies in such systems tackle crucial problems that lead to performance degradations in the wireless medium [2]. Furthermore, these strategies can be used to utilize particular network in the system by appropriate selection that will result in performance increase. The challenging task is the design of such strategies.

This paper proposes, analyzes and evaluates novel user oriented dynamic allocation strategy for user reconfigurability based on *game theory* concepts. The strategy is distributed at each user terminal and evaluated in a simulated WiMAX/WLAN reconfigurable system. User's terminals have two interfaces [3], one for each available network in the reconfigurable system. The result of the strategy is an appropriate network selection for the applications started on the terminal. This selection is made by game theory modeling, where players are the applications on the terminal and they can select WLAN or WiMAX. The result of the game differs due to the *degree of rationality* in the decision (Nash, QRE) and due to changes in the system (fluctuations of the received SNR from the networks). The general idea is to test how rationality of the distributed decision influences the overall user performances. Completely rational decision is modeled with the Nash equilibrium solution, whereas specific stochastic decision is modeled with QRE equilibrium. Furthermore, these two decisions (Nash based and QRE based) are compared with SNR based strategy.

The paper is organized as follows. Section II gives short background of game theory and explains the decision making for Nash and QRE equilibrium solutions. Section III explains the deployment of both game theory based RRM strategies (Nash and QRE) and the classical SNR based strategy. Section IV elaborates details of the system model with simulation setup, analytical definition of utility and presents the performance results in terms of application throughput and the application handover initiation. These results are used for comparison between the three decisions based on Nash, QRE and SNR. Finally, Section V concludes the paper.

II. GAME THEORY SCOPE

Game theory is a branch of applied mathematics. By using mathematical description and modeling, game theory concepts provide useful tool to analyze scenarios in which particular features can be modeled with games. The participants in the game are labeled as *players*. Each player has possible set of actions (*strategies*) that it can deploy and a *preference* that should be achieved during the game. The conflict of interest between the players implies that each player decision influences the preferences of the other participants in the game. Player's preferences are modeled with a *utility function* that depends on the scenario and the game model [4]. The main objective of the players in the game is to maximize their preferences by optimizing the utility function with appropriate deployment of their strategies. The solution concept in which referent player in the game is assumed to know the equilibrium strategies of the others and none of the players has any gain when unilaterally changes the strategy is Nash equilibrium [5]. The Nash equilibrium is a result of *totally rationality* of the

 ISBN: 978-1-61208-100-7

players, meaning that each player behaves in utility maximizing manner by playing the *best possible response* to other player's strategies.

Alternative approach to solve a game is the concept of *bounded rationality*. When the player's behavior is modeled with bounded rationality they do not always play the best response to other player's strategies. This approach implies stochastic decision for the strategy deployment where better actions are played with higher probability than worse actions. Stochastic decisions with bounded rationality can be more effective in particular cases. Nash equilibrium is optimal for rational players, however it may not be optimal for irrational players. The reason for irrationality occurrence may be due to dynamic changes of the modeled system or due to incomplete knowledge of all parameters that may influence the game outcome. Quantal Response Equilibrium (QRE) [6] is a solution concept of a game based on bounded rationality in which each player strategy is a *stochastic best response* to other player's strategies. There is a *precision parameter* (λ) within the QRE choice function that defines the degree of rationality in the decision making process. When this parameter is set to zero, the decisions become totally random, whereas for infinity value of the precision parameter the QRE equilibrium converges to Nash equilibrium.

The Nash solution is optimal for total rationality of the players. This rationality requires selection of the optimal strategy. For two possible strategies (*A* and *B*) and *utility* function *u* that maps the strategies in preferences, the rational player will do the following:

- $u(A) > u(B)$, play *A*;
- $u(A) < u(B)$, play *B*;
- $u(A) = u(B)$, play random;

The QRE solution is based on stochastic choice, where for two possible strategies (*A* and *B*) the player will do the following:

- $u(A) + \xi a > u(B) + \xi b$, play *A*;
- $u(A) + \xi a < u(B) + \xi b$, play *B*;
- $u(A) + \xi a = u(B) + \xi b$, play random;

where ξ is a random variable with zero mean value that can be observed as *additional shock* in utility calculation (mistakes in mapping the strategies to the preferences). The following equation denotes the probability for a player to play strategy *A* with the QRE solution:

$$P(A) = \frac{e^{\lambda u(A)}}{e^{\lambda u(A)} + e^{\lambda u(B)}} \quad (1)$$

The two game theory concepts map the degree of rationality in the decisions. The RRM strategies are based on this decision process, which is completely distributed. Hence, the obtained results evaluate the effect of rationality on overall user's performances (application throughput and handover initiation) in a reconfigurable system.

III. GAME THEORY BASED RRM

The user oriented RRM strategies proposed in this paper are based on Nash and QRE solution concepts. The implementation of the RRM strategies is for a terminal that has two interfaces (WiMAX, WLAN) that enable communication with the reconfigurable WiMAX/WLAN system. There are *two data applications* on each user terminal labeled as players in the game. This is done only for simplicity. The same approach can be used for more applications which can only lead to more complex computation of the equilibrium. Possible strategies for the players are to use WiMAX or WLAN network. The player's preferences in the game are defined with utility (application based metric) that depends on the application type and the specific PHY and MAC parameters of the component networks in the reconfigurable system. The general game setup is given in Table I, where P_1 and P_2 denote the players, *WiMAX/WLAN* are the possible strategies and with *U* as the appropriate utility for the particular strategy (e.g. $U_{wimax(1)}$ is the utility of P_1 for the *WiMAX* strategy).

TABLE I. Game notation

P_1 ; P_2	WiMAX	WLAN
WiMAX	$U_{wimax(1)}$; $U_{wimax(2)}$	$U_{wimax(1)}$; $U_{wlan(2)}$
WLAN	$U_{wlan(1)}$; $U_{wimax(2)}$	$U_{wlan(1)}$; $U_{wlan(2)}$

The outcome of the game is a network selection per application on each terminal. When *Nash based strategy* is deployed the decision for selecting network for the application is totally rational. In contrast, the QRE strategy deploys stochastic decision that depends on the precision parameter in the QRE choice function. The *QRE strategy* adopted in the paper deploys *opposite strategy* of the one selected with Nash. This can be achieved by adaptive tuning of the precision parameter that defines the probability for playing a certain strategy. To compare, if the Nash strategy selects the two applications to remain on one network, the modeled QRE strategy parts the applications on both networks. The game on each terminal is dynamically played and the decisions are updated, since the variations in the wireless channel and network availability change the utility perception of the mobile user.

The two game theory based RRM strategies are compared with classical *SNR based strategy*. This approach selects the interface with higher SNR for both applications. Additional intelligence in the SNR decision making is introduced by mechanism that parts the applications on both networks only when the SNR on the inactive interface exceeds the SNR value on current active interface (both interfaces have very high SNR). Comparing the SNR and the game theory RRM, the latter provides further intelligence by specifying why and which application should be run on particular network (since they are modeled as players).

 ISBN: 978-1-61208-100-7

IV. SYSTEM MODEL & PERFORMANCES

The following subsections elaborate the system model (simulation platform, utility functions) and present the evaluation of the decisions rationality within the game theory RRM. The results are compared with the SNR approach.

A. Simulation platform

The simulations are made in combined manner by utilizing three environments: *QualNet* [7], *Java* and *Gambit* [8]. The simulation platform is presented in Figure 1. Qualnet simulator is used to configure WiMAX/WLAN reconfigurable system and the terminals with two interfaces.

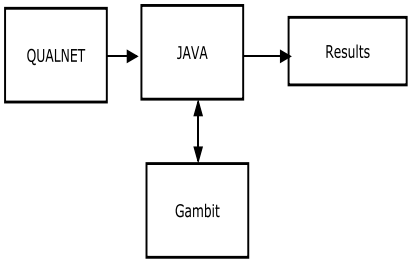


Figure 1.Simulation concept

The WiMAX cell has full coverage over the simulation area, whereas the WLAN has smaller coverage area which is completely within the WiMAX cell. As a result, the WiMAX network is always available for the users, whereas the WLAN is not because according to the mobility model the users can leave the WLAN coverage. The configuration parameters for the networks are given in Table II.

TABLE II. Parameters for WiMAX/WLAN

WiMAX	Antenna gain [dBi]	Tx power [dBm]	Antenna height [m]
BS	*18*	*43*	*35*
Channel	*3.3 GHz*		
WLAN	Antenna gain [dBi]	Tx power [dBm]	Antenna height [m]
802.11a	*8*	*15*	*1,5*
Channel	*5 GHz*		

When the simulation in QualNet is finished, the following parameters for every terminal (user) are transferred in Java environment:

- The time when an SNRwlan value is received
- The appropriate SNRwlan value in dB
- The time when SNRwimax value is received
- The SNRwimax value in dB

These values are used for definition of the user. The user is created from a defined *Java class*. Additionally, there are Java classes for the *applications*, *strategy deployment* and *results monitoring*. The strategy deployment in Java is done by coordination with the Gambit environment that provides Nash and QRE solution for the games. In Java, the parameters from QualNet and Gambit are combined and the results for average application throughput and the initiation of application handovers are obtained with event driven simulation.

B. Analytical utility model

The utility is application depended metric that shows how the application quality is perceived by the users. There are different approaches for estimating utility value for a particular application. References [9, 10] elaborate the concept of utility functions design and their application as metrics within RRM and optimization strategies in wireless systems. It shows that the utility function depends on the application type. For example, the on/off nature of the voice applications results in a steep utility function. References [11, 12] show approaches for estimation of the utility function for voice applications. The utility function for data application is a smooth curve due to the data elasticity. There are different utility functions designs depending on various technical parameters that "describe" the application flow in the wireless links.

This paper proposes RRM strategies for elastic data applications with *adaptive throughput* that can vary due to the conditions in the channel and available network resources. Such model can be used for variable data services corresponding to file download. In addition, they can be used within the game theory model since they do not have delay constrains. The utility depends only on the throughput. The adopted function for modeling the utility is given in [13, 14]:

$$u = a * \log[b * R * (1 - PER)] \qquad (2)$$

where R denotes the application throughput, PER is the packet error rate (predefined value 0,01) and a,b are dimensioning constants that depend on the maximum and minimum possible throughput in each network. The utility function depends only on the throughput and the specific modeling of the wireless system. The utility in WiMAX and WLAN is presented as follows, with a framework method that can be used for more complex approaches in utility based network dimensioning.

1) WLAN

The WLAN system in QualNet is based on IEEE 802.11a standard. The PHY modes that define the communication with the appropriate rate and SNR given in Table III [15].

TABLE III. Parameters for PHY mode in WLAN

PHY	1	2	3	4	5	6
Rate(Mbps)	*6*	*12*	*24*	*36*	*48*	*54*
SNR(dB)	*0-10*	*10-14*	*14-21*	*21-31*	*31-32*	*32-38*

The effective throughput depends on the procedures on MAC layer. The adopted approach is to use only half of the physical bitrate as effective throughput due to the MAC procedures. This means that maximum available throughput (*Rmax*) for the application will be when the AP PHY mode is 6 and serves only one user. Minimal available throughput (*Rmin*) will depend on the predefined number of maximum

 ISBN: 978-1-61208-100-7

possible users on WLAN who share the network capacity. This number can be predefined with admission control procedure (adopted value *Nmax*=10). Adopting this approach, the equations (3, 4) are used for calculating the minimal and maximal throughput in the WLAN network. The maximum and minimum throughputs are then used to create system of two logarithmic equations according to (2).

By setting maximum (=1) and minimum (=0.1) possible values of the utility function, system of logarithmic equations is formed.

$$R_{\max} = \frac{PHY_{\max}(54Mbps)}{2} \quad (3)$$

$$R_{\min} = \frac{PHY_{\min}(6Mbps)}{2}\frac{1}{N_{\max}} \quad (4)$$

The logarithmic equations for the system are given with (5) and (6), as follows:

$$u_{\max} = a\log_{10}(bR_{\max}(1-PER)) = 1 \quad (5)$$

$$u_{\min} = a\log_{10}(bR_{\min}(1-PER)) = 0.1 \quad (6)$$

These equations are used to define the utility values. The adopted approach for estimation of WLAN utility is done in simplistic manner by using only half of the physical throughput. Alternative approach for utility estimation of the WLAN can be with OFDM analysis. This type of estimation is done for the WiMAX system as elaborated in the following subsection.

2) WiMAX

The WiMAX system in QualNet is based on IEEE 802.16e standard. Estimation of the utility for WiMAX is performed by PHY and MAC analysis of the WiMAX technology. The received SNR in WiMAX defines the modulation and the coding used for sending appropriate number of bits/symbol. This mapping is given in Table IV [16]. The number of subcarriers in the system is *N*=2048 and the channel bandwidth is *B*=20MHz. The effective duration of the symbol is obtained in standard manner according to [17]:

$$T_N = NT_s \quad (7)$$

where *Ts=1/fs*, and *fs=8/7B*. The guard interval between the uplink and downlink of the MAC frame is calculated with:

$$T_g = \{\frac{1}{4};\frac{1}{8};\frac{1}{16};\frac{1}{32}\}T_N \quad (8)$$

Total symbol duration is given with the following equation:

$$T_{Ntotal} = T_g + T_N \quad (9)$$

The total number of symbols (*Tot_Sym*) is obtained from the MAC frame duration which can be adaptive according to standard, but for the analysis duration of 5ms is used.

This total number of symbols (*Tot_Sym*) is obtained when the MAC frame duration is divided with the total symbol duration. From the total number of symbols only the symbols used for data transfer are labeled as effective and are used for defining the utility function. The effective number of symbols (*Eff_Sym*) is given with the following equation:

$$Eff\_Sym = Tot\_Sym - (\sum_{i=1}^{5} X_i) \quad (10)$$

where

$$(X_1 + X_2 + X_3 + X_4 + X_5) = (3,1,7,1,1) \quad (11)$$

are five groups (indexed with *i*) of symbols for broadcast, contentions slots and preambles in the MAC frame [18].

TABLE IV. Mapping SNR in bits/symbol in WiMAX

Modulation and Coding	SNR	Bits per symbol
BPSK: 1/2	*3*	*96*
BPSK: 1/4	*6*	*192*
QPSK: 1/2	*8,5*	*288*
QPSK: 1/4	*11,5*	*384*
16 QAM:3/4	*15*	*576*
16 QAM:2/3	*19*	*768*
64 QAM	*21*	*864*

The approach for estimation the utility of the WiMAX system is done according to the analysis elaborated in [18], where the total number of symbols for all active users (*Us_TotSym*) in WiMAX depends on the frame duration (*Fr_Dur*) and the SNR for every user:

$$Us\_TotSym = R * Fr\_Dur * (\sum_{i=1}^{7} \frac{Nss_i}{bits\_sym_i}) \quad (12)$$

where *R* is the available throughput for the incoming user, Nss_i is the number of serving users with appropriate bits/symbol depending on their channel (SNR). The possible values for $bits\_sym_i$ in equation (12) correspond to the seven mappings (index *i* denotes particular row) in Table. IV. For example, if there are 3 users with SNR < 3, than for *i*=1 in (12), Nss_i =3 and $bits\_sym_i$ =96 for the three users. The throughput is used to model the utility for WiMAX with system of logarithmic equations (13, 14) :

$$u'_{\max} = a\log_{10}(bR'_{\max}(1-PER)) = 1 \quad (13)$$

$$u'_{\min} = a\log_{10}(bR'_{\min}(1-PER)) = 0.1 \quad (14)$$

where, due to predefined system dimensioning with admission control M_{max} *=20* (similar to N_{max} value adopted for WLAN), the maximal (Nss_i*=1*, $bits\_sym_i$*=864*) and minimal throughput are given with the equations (15,16):

$$R'_{\max} = \frac{Eff\_Sym}{Fr\_Dur * (\sum_{i=1}^{7} \frac{N\_SS_i}{bits\_sym_i})} \quad (15)$$

$$R'_{\min} = \frac{Eff\_Sym}{Fr\_Dur * (\frac{M_{\max}}{bits\_sym})} \quad (16)$$

C. Performance evaluation

The simulations with the proposed simulation platform are performed with 8 users in a square simulation environment (1000x1000m) with 16 running applications. The simulation duration is 600s in order to provide sufficient statistical regularity of the results. This implies that further increase of the number of users (and thus the number of the applications) will not change the trend of the results. The user's follow random waypoint mobility model and communicate with the networks on two separate channels. The path loss model is two-ray and the shadowing factor is 4dB. The performance evaluation of the three RRM strategies based on Nash, QRE and SNR are presented with the following parameters of interest:

- Dynamic allocation of applications per networks
- Average application throughput on WiMAX
- Average usage of WLAN
- Application handovers

The results for the *dynamic allocation* of the applications with Nash/SNR/QRE decisions observed on WiMAX and WLAN networks are given on Figure 2. and Figure 3.

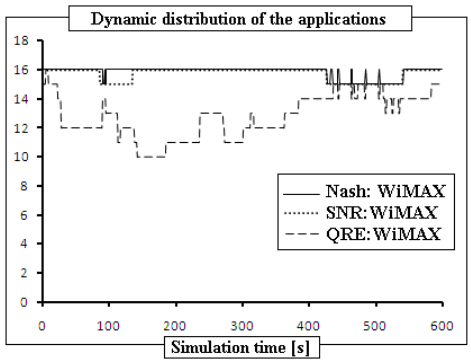


Figure 2. Dynamic distribution of the user's application on WiMAX network with Nash/SNR/QRE decisions

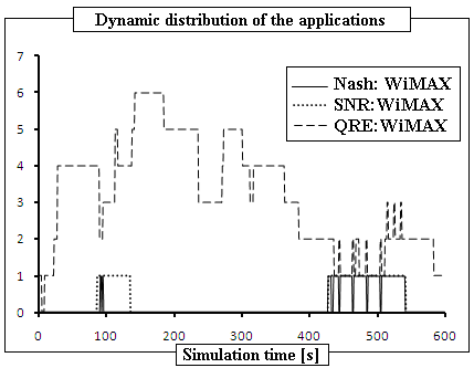


Figure 3. Dynamic distribution of the user's application on WLAN network with Nash/SNR/QRE decisions

The analysis of this parameter shows the load balancing feature of the distributed RRM in the reconfigurable system. Because the terminals do not have network related information (number of serving users, applications, etc.), they estimate their utilities dynamically for the game according to static network parameters and on their input in the network (two applications). The results show that using SNR and Nash strategies similar results are obtained for the load balancing, resulting in very rare usage of the WLAN. The QRE strategy, since it is tailored to do the opposite of Nash, results in frequent WLAN usage. In this sense, the QRE strategy provides better load balancing in the distributed decision making process.

The *average application throughput* per user on WiMAX is shown in Figure 4. This parameter is averaged during the simulation time because WiMAX is always available (full coverage) for the mobile users in the scenario. The average application throughput per user is a mean value of the throughput that both applications experience on the user terminal. The results show that QRE yields highest application throughput per user compared to SNR and Nash strategies which have similar performances.

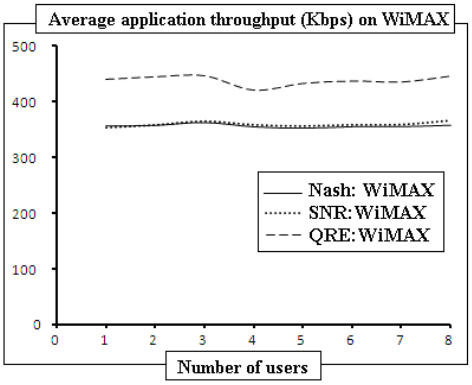


Figure 4. Average throughput per application for each user on WiMAX network with Nash/SNR/QRE decisions

The *average usage* of WLAN AP is shown on Figure 5, measured as throughput. It shows how much the AP was used during the simulation time on average by the users. The QRE strategy invokes highest usage of the WLAN AP due to the dynamic distribution of the applications on both networks. SNR and Nash have similar average AP usage. This parameter is also linked to the previous parameter that showed the average throughput per application for each user on WiMAX. When the deployed RRM yields more frequent AP usage in the scenario it produces lesser load on WiMAX and thus WiMAX throughput increase on average basis.

The application handovers are shown on Figure 6, in a normalized manner, where the strategy with the highest number of handovers (QRE) corresponds to 100%. This parameter shows how much handovers are occurred in the reconfigurable system for all of the user's applications with each RRM strategy. In addition, this parameter shows how the RRM strategy is prone to performing vertical handovers in the system. This parameter is highest with QRE

technique, whereas Nash results in slightly higher handover initiation compared to SNR. There is an imminent degradation of quality when vertical handover occurs because the reconfiguration of the communication protocol stack on the link user-network disrupts the service quality. In this sense, the QRE results in highest performance degradation compared to SNR and Nash.

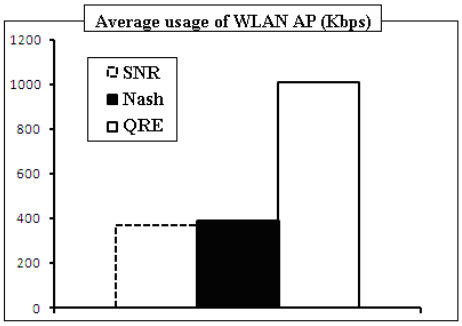


Figure 5.Average usage of WLAN AP: Nash/SNR/QRE

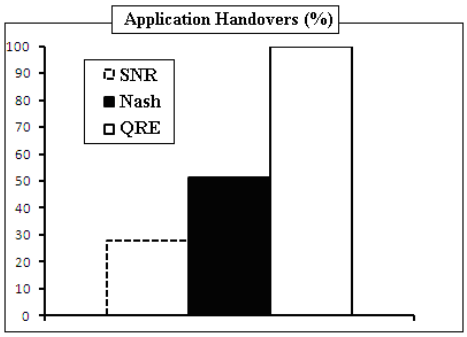


Figure 6.Handovers in (%) with Nash/SNR/QRE decisions

The performance results depend on the analytical modeling of the utility, the analysis of the networks and the scenario model. However, the presented framework approach is general and applicable in similar environments where terminals deploy distributed RRM to explore reconfigurability benefits and maximize their preferences. Furthermore, the results show the tradeoff between the throughput of the application and the handovers initiations as degradation factor. Such tradeoff should be considered when managing terminals in reconfigurable systems.

V. CONCLUSION

This paper proposes and evaluated novel user oriented RRM strategies for reconfigurable WiMAX/WLAN system. The strategies are distributed, based on game theory concepts (Nash and QRE), where the game is played by the applications on each terminal modeled as players. The RRM strategies result in appropriate network selection for each application on the user's terminals and are compared to SNR based strategy. The results show the tradeoff between the load balancing, user's throughput and handover initiation between the strategies. The provided framework can be used in distributed RRM strategies to tune user preferences by balancing application's throughput demands and handover initiation between the networks in the reconfigurable system.

REFERENCES

[1] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of L. Gavrilovska and V. Atanasovski, "Interoperability in Future Wireless Communications Systems: A Roadmap to 4G," Microwave Review, 13(1), June 2007, pp.19 - 28

[2] Z.Han and K.Liu, "Resource Allocation for Wireless Networks, Basics, Techniques and Applications", Cambridge university press, 2009.

[3] White Paper, "WiMAX and Wi-Fi Together: Synergies for Next-Generation Broadband, Intel Corporation, July 2008. Downloaded from http://download.intel.com/netcomms/technologies/wimax/wimax_and_wifi_together.pdf, on February 12, 2009.

[4] M. Felegyhazi and J. Huubaux, "Game Theory in Wireless Networks: A Tutorial," Tehnical Report , submmited on Feb. 21, 2006 February 12, 2009.

[5] A. MacKenzie and L. DaSilva, "Game Theory for Wireless Engineers", Publication in Morgan&Claypool , series: synthesis lectures on communications, lecture #1.

[6] T. Turocy, "Using Quantal Response to Compute Nash and Sequential Equilibria", *Economic Theory* 42(1): 255-269, 2010.

[7] "QualNet 4.5 User's Guide" Scalable Network Technologies, Inc.,March 2008

[8] McKelvey, Richard D., McLennan, Andrew M., and Turocy, Theodore L.(2007) Gambit: Software Tools for Game Theory, Version 0.2007.01.30 http://www.gambit-project.org.

[9] J. Riihijärvi, M. Wellens, and P. Mähönen, "Link-Layer Abstractions for Utility-Based Optimization in Cognitive Wireless Networks," *in Proceedings of CrownCom 2006*, Mykonos, Greece, June 2006.

[10] L. Badia and M. Zorzi, "On Utility-based Radio Resource Management with and without Service Guarantees," *in Proceedings of the 7th ACM International Symposium on Modeling, Analysis and Simulation of Wireless and Mobile Systems*, pp. 244-251, Venice, Italy, 2004.

[11] "The E-Model, a computational model for use in transmission planning," ITU-T Recommendation G.107, May 2000.

[12] C. Boutremans and J.-Y. Le Boudec, "Adaptive Joint Playout Buffer and FEC Adjustment for Internet Telephony," *in Proceedings of IEEE INFOCOM 2003*, vol.1 pp. 652- 662, San Francisco, April 2003.

[13] S. Khan, S. Duhovnikov, E. Steinbach, and W. Kellerer, "MOS-Based Multiuser Multiapplication Cross-Layer Optimization for Mobile Multimedia Communication," *Advances in Multimedia*, Hindawi Publishing Corporation, vol 2007, Article ID 94918.

[14] F. P. Kelly, "Charging and rate control for elastic traffic," *European Transactions on Telecommunications*, vol. 8, no. 1, pp. 33–37, 1997.

[15] Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications, ANSI/IEEE Std 802.11a/b, 1999, Supplement to ANSI/IEEE 802.11 Std 802.11, 1999 Edition.

[16] IEEE 802.16e Standard for Local and Metropolitan Network, Part 16: Air Interface for Fixed and Mobile Broadband Access, http://standards.ieee.org/getieee802/download/802.16e-2005.pdf

[17] Loutfi Nuaymi, "WiMAX: Technology for Broadband Wireless Access", Wiley, March 23, 2007.

[18] P.Mach and R.Bestak, "WiMAX throughput evaluation of convetional relaying," *Telecommunication Systems*, vol. 38, no. 1-2, pp.11-17, March 15, 2008.

 ISBN: 978-1-61208-100-7

A Beacon Cluster-Tree Construction Approach For ZigBee/IEEE802.15.4 Networks

Mohammed.I. Benakila, Laurent George
LACSC Laboratory
ECE Paris, school of engineering
Paris, France
e-mail: Benakila@ece.fr, lgeorge@ieee.org

Smain Femmam
Academy of Strasbourg
University of Haute Alsace
France
e-mail: femmam@ieee.org

***Abstract*— Wireless Sensor Networks (WSN) based on the IEEE 802.15.4 standard are in a constant expansion. Applications like production control, building control are more and more based on WSN because of their energy efficiency, self organization capacity and protocol flexibility. The IEEE 802.15.4 standard defines 3 network topologies: the mesh topology, the star topology and the Cluster-Tree topology. However, the construction of Cluster-Tree networks based on the beacon mode is still undefined by the IEEE 802.15.4 standard. A Beacon Cluster-Tree topology has the advantage of giving all the benefits of the Beacon mode (Synchronization, QoS support through Guaranteed Time Slots) and at the same time, allows the construction of large networks to cover large areas. In order to offer to network architects more flexibility in designing WSN, we present, in this paper, a Beacon Cluster-Tree topology construction approach. Our approach is different from what was proposed until now. We summarize our contribution in three points: (1) there are no conditions on the Beacon mode SuperFrame structure, (2) despite the size of the networks, the construction of a beacon Cluster-Tree topology is always possible and (3) no scheduling is done on SuperFrames or even on Beacon frames transmissions. Indeed, in this paper, we present a novel approach that exploits wireless receivers capability in dealing with multipath to retrieve transmitted data in order to avoid scheduling problems.**

Keywords - IEEE 802.15.4; Beacon mode; Beacon frame ; scheduling; SuperFrame scheduling; Cluster-Tree.

I. INTRODUCTION

Nowadays, the need of controlling human environment is strongly present in people's mind. This need aims at introducing more comfort in people's life, assuring an ambient assisted living (AAL), building automation or factory automation. Such diversified applications require communicating nodes that ensure an efficient data processing, limited energy consumption and must be based on a flexible protocol stack to fit with the requirements of each application.

ZigBee is considered to be a suitable network for sensing and control applications. ZigBee standard defined by the ZigBee Alliance [1] is a communication protocol for Wireless Sensor Networks (WSN). It provides mechanisms for network establishment, device communication and packets routing. Networks implementing this standard are low energy consumption and self-organized. The ZigBee standard is based on the IEEE 802.15.4 standard for the medium access control (MAC sub-layer) and for wireless transmissions and receptions (physical layer).

The IEEE 802.15.4 MAC sub-layer allows two modes for transmitting and receiving data: beacon enabled mode and non-beacon enabled mode [2]. The former can guarantee transmission determinism within Guaranteed Time Slots (GTSs), but needs a synchronization between all the devices forming the beacon enabled network. Non-beacon mode does not give any traffic guarantee and does not need synchronization between devices (see Section 3).

Three topologies are available in the IEEE 802.15.4: mesh topology, star topology and Cluster-Tree topology (see Section 3). The beacon mode has been designed to work with a star topology. However, no mechanisms have been defined in the IEEE 802.15.4 standard to enable the beacon mode using a Mesh or a Cluster-Tree topology. In this paper, we are interested in the construction of a Beacon Cluster-Tree topology, i.e., constructing a Cluster-Tree topology using the beacon mode.

In the present paper, Cluster-Tree mechanisms will not be modified, i.e., no modifications will be made on the association mechanism or data transmission mechanism. All the network devices will transmit during the same SuperFrame, which means that the beacon order (BO) and the SuperFrame order (SO) parameters will be the same for the whole network [2]. Nevertheless, using our approach, SuperFrames scheduling will be avoided and the construction of a Beacon Cluster-Tree network will be always possible without introducing any constraints on the SuperFrames parameters.

The rest of the paper is organized as follows: the next section presents some related works. Section three contains a brief overview of the IEEE 802.15.4 MAC sub-layer. Section four introduces the beacon collision problem within a Cluster-Tree topology, and in Section five we present the most known approaches to resolve this problem. The core of the proposed approach is presented in Section six. Simulation results are presented in Section seven, and finally, we conclude.

 ISBN: 978-1-61208-100-7

II. RELATED WORKS

Sensor networks applications are in importance nowadays. Sensors applications are much diversified, from temperature sensing to health care. In addition, sensor networks are intended to operate in different environments (factories, hospitals, museums) [3]. Consequently, flexibility in sensor networks design becomes a crucial property that standardization organizations are trying to ensure when defining their protocols.

Proposing new functionalities for a given network topology is a way to provide more flexibility for sensor networks design. We have proposed in [4] the definition of a new device, called the beacon-aware device. This device allows beacon and non-beacon networks cohabitation. The beacon-aware devices permits to create a network composed of a mix of beacon and non-beacon devices. The solution presented in [4] guarantees the integrity of the beacon network traffic by introducing a channel access priority mechanism.

Beside the network size and topology, the QoS is an important parameter to take into consideration. Some sensor applications that require a bounded transport delay can use Guaranteed Time Slots (GTSs) mechanism defined by the IEEE 802.15.4 standard within a fully beacon network. In [5], the authors propose a modelling methodology for Cluster-Tree networks in order to compute worst-case end-to-end delay, buffering and bandwidth requirements. This modelling method enables the network designers to create Cluster-Tree networks that fit with their application constraints.

IEEE 802.15.4/ZigBee standard defines the Cluster-Tree topology as a special case of a peer-to-peer network. But the realization of beacon Cluster-Tree networks is not defined in the standard. Some works have been done in order to model Cluster-Tree topologies [6] [7], failure recovery [8] and to allow the construction of Beacon Cluster-Tree networks.

The RFC submitted to the Task Group 15.4b (see [9]) propose enhancements to the IEEE 802.15.4 standard. The construction of beacon Cluster-Tree topologies was one of the document topics. The authors of [9] classify beacon frames conflicts into two categories: direct conflict and indirect conflict, and propose some approaches to solve each category of beacon conflict.

We present the beacon conflict categories and the approaches proposed by [9] in Section 4.

In [10], Koubaa, Cunha and Alves propose a SuperFrame scheduling algorithm to enhance the approach introduced in [9]. Indeed, the authors of [9] do not introduce any scheduling algorithm. [10] tackles the problem by introducing the constraints and the algorithm needed to provide a strong scheduling mechanism. [11] propose a mechanism to schedule beacon frames transmissions called beacon-only period approach. In this approach, there is no need to schedule SuperFrames, i.e., all the SuperFrames start at the same time. Only beacon frames transmissions are scheduled within a new SuperFrame period called the beacon-only period (for more details, see Section 4 of this paper). In addition, [11] introduces a GTS collision avoidance mechanism which guarantees a certain traffic QoS.

In this paper, we present and discuss the drawbacks of those solutions for the Cluster-Tree network management in Sections 4 and 5. We then describe the approach we propose in Section 6.

III. IEEE 802.15.4 MAC SUB-LAYER OVERVIEW

The IEEE 802.15.4 standard is a suitable protocol for low rate wireless networks. A lot of efforts has been done to make the standard low-power and self organized.

Tow kinds of devices have been introduced in [2], reduced function devices (RFD) and full function devices (FFD). A FFD is a device that implements all the functions defined by the standard. However, a RFD implements the basic functions (join a network, leave a network, transmit, etc.) defined in the standard. Two topologies are allowed by the standard:

Mesh topology:

In a mesh topology, there is one coordinator and a set of nodes associated to it. Each node is a router and permits other nodes association. A given node can communicate directly with other nodes if they are in its POS (Personal Operating Space, i.e., in-range transmission), or, passing by other nodes (acting as routers in this case) to reach its target node (see Figure 1). Using this topology, no synchronization is needed between the devices.

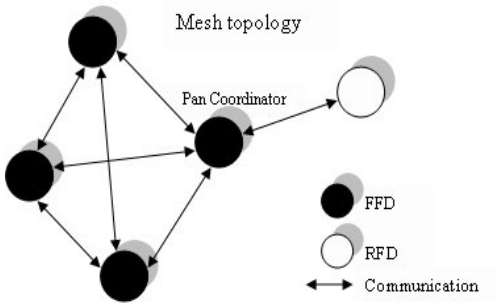


Figure 1. Mesh topology.

Furthermore, enabling the synchronization in such a topology can be problematic since synchronization mechanisms for mesh topology are not defined in the standard.

Star topology:

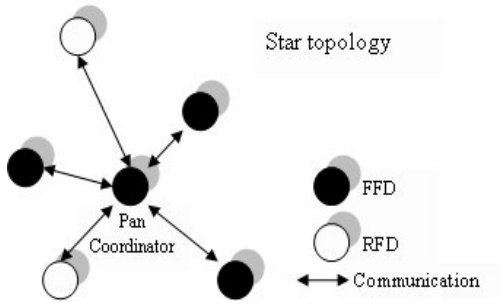


Figure 2. Star topology.

The coordinator is the main node in the network. All other nodes must be associated to it, and all communications between nodes must pass through it, even if the communication initiator node and the target node are in the POS of each other (see Figure 2). Performing synchronization (beacon mode) using this topology has been well defined by the IEEE 802.15.4 standard.

A third topology can be considered also, the Cluster-Tree topology. This topology is not defined in the 2006 version of the standard, but, was defined in the 2003 version of the standard. The Cluster-Tree topology is very interesting for time sensitive applications. Here after, the Cluster-Tree topology is introduced.

Cluster-Tree topology:

A Cluster-Tree network is a network in which most devices are FFDs. A RFD connects to a cluster tree network as a leaf device at the end of a branch (RFDs do not allow other devices to associate). A FFD device may act as a coordinator and provides synchronization services to other

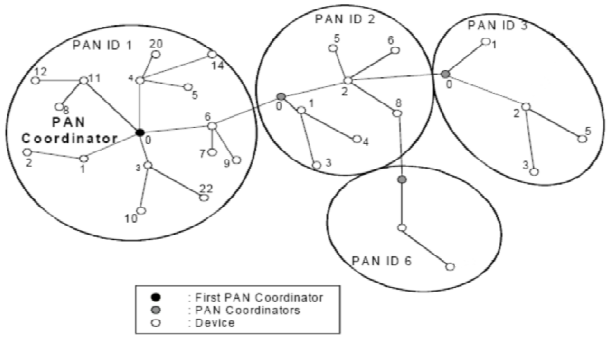


Figure 3. Cluster-Tree topology.

devices or other coordinators. The PAN coordinator forms the first cluster by choosing an unused PAN identifier and broadcasting beacon frames to neighboring devices. A candidate device receiving a beacon frame may request to join the network at the PAN coordinator. If the PAN coordinator permits the device to join, it adds the new device as a child device in its neighbor list. Then the newly joined device adds the PAN coordinator as its parent in its neighbor list and begins transmitting periodic beacons; other candidate devices may then join the network at that device. The simplest form of a cluster tree network is a single cluster network, but larger networks are possible by forming a mesh of multiple neighboring clusters. Once predetermined application or network requirements are met, the first PAN coordinator may instruct a device to become the PAN coordinator of a new cluster (Cluster head) adjacent to the first one. Other devices gradually connect and form a multicluster network structure (see Figure 3).

A. Non-beacon enabled network

In a non-beacon mode, the three topologies can be used. This mode assumes that every node can communicate directly with other nodes without any synchronization requirements. A node can transmit at any time, and can go to sleep at any time following its own energy consumption policy. All transmissions are done after performing the unslotted CSMA/CA algorithm to check if the channel is clear for a transmission or not.

A non-beacon device transmits the beacon frame only as a response to a beacon request command. Devices operating in this mode do not need to synchronize with other devices.

B. Beacon enabled network

In a beacon enabled mode, the coordinator plays a crucial role. It defines periods of time in which transmissions can be done and intervals of time where all nodes associated to it must go to sleep.

In this mode, the time is divided into a succession of ”SuperFrames”. A SuperFrame is a time interval that contains an active period and an inactive period. The Beacon Interval (BI) parameter indicates the interval between two successive beacon frames. The length of the active period is indicated by SD (SuperFrame Duration) parameter. The active period is divided into a fixed number of 16 time slots of equal sizes. All beacon network communications are done within this period. The active period is divided into a contention access period (CAP) and a contention free period (CFP).

The CAP is the period where all nodes compete for channel access using the slotted CSMA/CA algorithm.

The CFP gathers GTSs (Guaranteed Time Slots). A GTS is one or more slots of time reserved for a particular node.

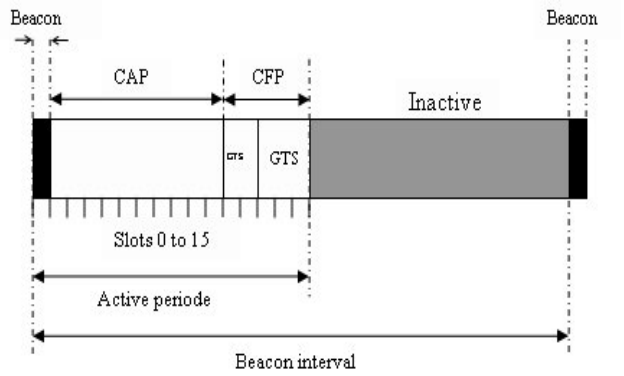


Figure 4. IEEE 802.15.4 SuperFrame structure.

 ISBN: 978-1-61208-100-7

A GTS is unidirectional, i.e., only for receptions or only for transmissions. The coordinator starts allocating GTSs from the last time slot to the first slots respecting a maximum size of the CFP. GTS transmissions do not need the use of CSMA/CA algorithm for channel access since the slots are reserved for one node. The structure of a SuperFrame is illustrated in Figure 4.

Beacon mode forces all the devices to synchronize with the coordinator. This is done by the reception and the process of the beacon frame. The most important parameters for synchronization are: the Beacon Order (BO) which is a parameter for computing BI, the Superframe Order (SO) which is a parameter for computing SD and the final CAP slot parameter. The final CAP slot indicates the end of the CAP. After this slot, only devices owning a GTS can transmit. Figure 5 presents the format of the beacon frame.

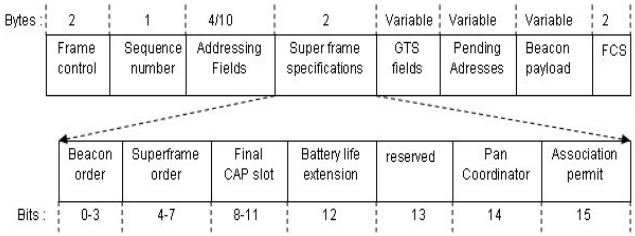


Figure 5. Beacon frame structure.

The BO is a parameter used by the associated nodes for calculating the beacon interval, which is computed using this formula :

$$BI = aBaseSuperF\ rameDuration * 2^{BO} \qquad (1)$$

with: aBaseSuperFrameDuration = 60 symbols.
The BO value should be between 0 and 14.
A BO with a value of 15 indicates that the device operates in non-beacon mode.
The SO is a parameter used for calculating the active period duration.

$$SD = aBaseSuperF\ rameDuration * 2^{SO} \qquad (2)$$

with: $0 \leq SO \leq BO \leq 14$.

IV. BEACON FRAME COLLISION IN A CLUSTER-TREE TOPOLOGY

In this section, we present the beacon frame collision problem when a Cluster-Tree topology is considered. This problem has been addressed as a request for comment (RFC) by the Task Group 15.4b in [9].

Two types of beacon frame collision have been identified in [9].

A. *Direct beacon frame collision*

A direct beacon frame collision occurs when more than one coordinator are in the transmission range of each other, and the beacon transmission occurs in approximatively the same time (see Figure 6(a)).

B. *Indirect beacon frame collision*

An indirect beacon frame collision occurs when a given node is in the transmission range of two or more coordinators. The coordinators send their beacon frames at approximatively the same time, i.e., at a given time, the node receives more then one beacon frame which results in the collision. This situation is a typical hide nodes transmission situation (see Figure 6(b)).

For more details about the beacon frame collision problem, see [9].

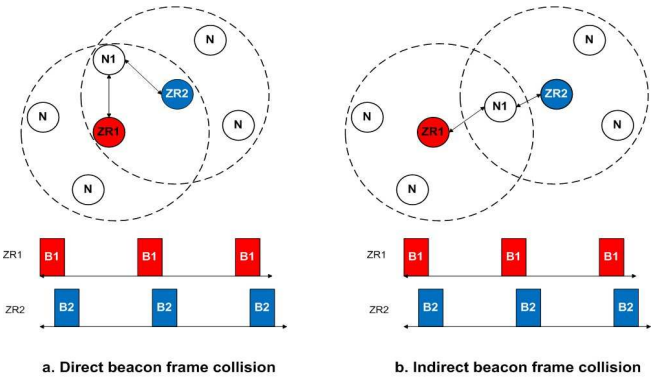


Figure 6. Beacon frame collision situations.

V. BEACON FRAME COLLISION AVOIDANCE USING THE TEMPLATE

This section contains a presentation of beacon frame collision avoidance mechanisms for each of the situations presented in the previous section.

A. *Direct beacon frame collision avoidance*

Two main approaches have been proposed to solve the problem of direct beacon frame collision: the time division approach and the beacon-only period approach. These approaches were defined in [9] and developed in [10] and [11].

The time division approach:

Basically, this approach consists in the following principals:

A given coordinator transmits its beacon frame and spends its active period during the inactive period of its neighbor coordinators. The Task Group 15.4b does not propose any scheduling algorithm in order to increase the mechanism efficiency. This lack has been tackled in [10].

 ISBN: 978-1-61208-100-7

The authors of [10] propose a SuperFrame scheduling algorithm in order to maximize the number of clusters in the network.

This approach suffers from several problems:

- To enable parent/child communications, a coordinator is activated during its active period and during the active period of its parent coordinator.
- The increasing density of devices in the network makes the problem more complicated, and the scheduling algorithm proposed in [10] may return an ”unschedulable set” response, which means that the Cluster-Tree topology can not be used.
- To make the SuperFrame scheduling algorithm more efficient, the authors of [10] have made restrictive constraints on the SO and the BO parameters, which could perturb the execution of some applications in the network.

The beacon-only period approach:

In this approach, the SuperFrame structure is modified. A time period, called ”Beacon-Only period”, is added at the beginning of the SuperFrame. The beacon-only period is divided into time slots called ”Contention-Free Time Slot” (CFTS). Each coordinator transmits its beacon frame within its CFTS (see Figure 7). Thus, beacon frame collisions will be avoided.

This approach has been presented, first, by the Task Group 15.4b, and then it was developed in [10] and [11].

However, dimensioning the beacon-only period is complicated since the duration of the period must be evaluated dynamically depending on association and leaving actions of beacon coordinators. In addition, the beacon-only approach does not avoid indirect beacon frames collision, or it does, but, with including global CFTS scheduling algorithm.

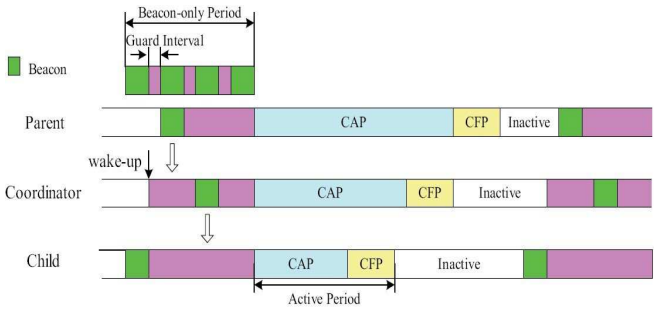


Figure 7. The beacon-only period approach.

B. *Indirect beacon frame collision avoidance*

Indirect beacon frame collision can be solved by the following two approaches:

The reactive approach:

Enabling this approach, the network is started normally and the coordinators do not do much to prevent from beacon frame collisions. Once a collision occurs, the node (the node in the POS of more than one coordinator, see Figure 6) will start orphan scans to try to re-synchronize with its coordinator. However, if after a number of orphan scans, the node is still unable to receive correctly the beacon frame, it initiates a beacon conflict command. Coordinators receiving the beacon conflict command will adjust their beacon transmission time in order to solve the problem.

This approach is simple, but, the recovery from a beacon conflict can take a long time.

The proactive approach:

In the proactive approach, coordinators try to avoid beacon frames conflict before starting their beacon frames transmission. A coordinator listens to the channel and collects its neighbours beacon frame transmission time. Nevertheless, if a beacon frame collision is reported, the network is able to solve the problem using the reactive approach.

Notice that this approach is more complicated than the first one.

C. *Discussion*

Previously, we presented different approaches proposed to enable the construction of a beacon Cluster-Tree topology. This section contains a discussion about the presented solutions and our motivations to present a new approach.

The Time Division Beacon Scheduling (TDBS) [10] is an improvement of the solution proposed by the Task Group 15.4b. This approach is based on a scheduling algorithm to avoid beacon collisions. However, the TDBS approach suffers from several lacks (see Section 5.1).

The second approach presented was the beacon-only period approach. This approach suffers also from several lacks presented in Section 5.1.

The approach we present, in the next section, aims at enabling the following properties:

- It’s clear that scheduling algorithms (for SuperFrames or Beacon frames) are inappropriate for networks with high nodes density. Our approach does not introduce any scheduling algorithm which means that the solution presented in this paper can be applied to construct beacon Cluster-Tree networks whatever the size of the network.
- Using our approach (same case for the beacon-only period approach), parent/child communications are easily enabled. In fact, to communicate with its parent node, a given node does not need to be active during its active period and its parent active period. Therefore, this mechanism simplifies the protocol implementation and reduces the energy consumption of the nodes forming the network.

 ISBN: 978-1-61208-100-7

VI. OUR APPROACH

In this section, we propose a new approach that allows forming Cluster-Tree networks without regard to the density of the network. Indeed, our approach is not based on SuperFrames or even Beacon frames transmission scheduling.

In our approach, the Cluster-Tree network is constructed thanks to the following:

- The same BO and SO values for all the nodes of the network [2].
- All the nodes are synchronized thanks to beacon frames transmission.
- All the nodes transmit during the same SuperFrame.

These points are detailed in the rest of this section.

A. *Beacon frame transmission*

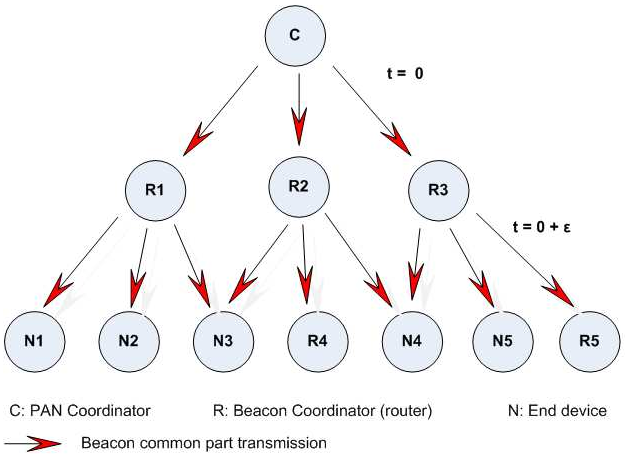


Figure 8. Beacon common part transmission.

To enable collision-free beacon transmissions, we are adopting a novel approach, described in this section. The beacon frame is divided into two parts:

- A common part: It is the part that does not change from a beacon coordinator to another. It contains the SO and the BO parameters.
- A specific part: It is the part specific to a beacon coordinator, i.e., changes from a given beacon coordinator to another.

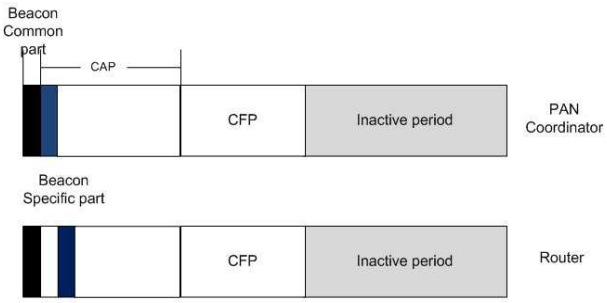


Figure 9. SuperFrame structure.

Each part is a separate frame. The transmission mechanism is described in Figure 8.

Nodes synchronization is achieved by the transmission of the common part. It gives the start signal to sensor nodes to begin the Beacon mode SuperFrame.

The common part contains only synchronization information (BO and SO).

The specific part is transmitted during the CAP period and it contains the traditional beacon information (GTSs, Pending addresses, etc.).

The PAN Coordinator broadcasts the beacon common part.

When the beacon common part is received, a beacon coordinator begins its SuperFrame and forwards the same frame (i.e., the beacon common part) to its neighbour nodes. Using this mechanism, beacon coordinators at the same level of the Cluster-Tree can transmit the common part at the same time. This should not cause a reception problem if a node receives more than one frame at the same time.

Indeed, all the beacon routers are broadcasting the same frame, the same bit configuration which means that all the beacon routers are transmitting the same RF signal. When a node receives more than one RF signal it can extract the message because all the RF signals are considered as multipath RF signals.

Multipath propagation occurs when RF signals take different paths from a source to a destination. A part of the signal goes to the destination while another part bounces off an obstruction, and then arrives to the destination. As a result, a part of the signal encounters delay and travels a longer path to the destination. Multipath can be defined as the combination of the original signal plus the duplicate wave fronts that result from reflection of the waves off obstacles between the transmitter and the receiver [12].

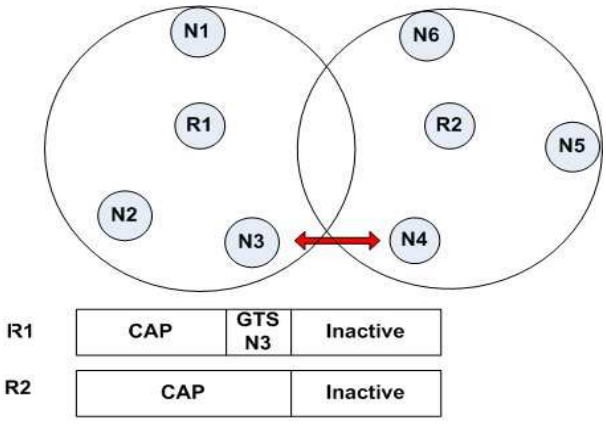


Figure 10. Example of a GTS collision case.

Multipath propagation occurs even with only one transmitter and one receiver. Nowadays receivers are able to retrieve the information from a RF signal perturbed by reflected RF signals.

Thus, a node receiving the beacon common part from more than one beacon coordinator (i.e., more than one RF signal) is able to retrieve the common part information since the node is able to deal with multipath RF signals, as confirmed in the experiments we have done in Section 7.

B. *GTS Allocation*

In our approach, each beacon coordinator manages independently its CFP period. It is able to accept or reject GTS requests and it is responsible for assigning GTS time slots to its children nodes. However, a mechanism must be introduced to avoid GTS transmission disruption by neighbor nodes transmissions. Figure 10 illustrates a case where a GTS transmission can be perturbed. We can see that if node"N4" is transmitting while "N3" is in its GTS period, there will be frame collisions.

To avoid this problem, GTS collision avoidance mechanisms should be implemented. We can conceive proactive or reactive approaches. These approaches are out of the scope of this paper.

Figure 11. MicroChip sniffer (left), PICDEM Z (left).

VII. SIMULATION AND EXPERIMENTS

This section aims at proving the well-functioning of the mechanism.

In the first part of this section we present experimentations we have done to show that receivers' capability to deal with multipaths could be exploited to transmit the beacon common part.

In the second part of this section, NS2 simulations are done to show that our approach can be implemented on sensor nodes.

A. *Multipath exploitation*

The core of our approach is the exploitation of multipath phenomena to avoid beacon transmission scheduling. Indeed, since receivers are able to deal with RF reflected signals to retrieve the information, they can retrieve information from RF signals (same RF signal) transmitted by different nodes at approximately the same time. All these RF signals will be considered as multipath signals by the receiver.

To put the stress upon this point, we conceived a real experimentation using the MicroChip PICDEM Z modules (see Figure 11). The principle is to send the same frame by several nodes to one receiver. For visual considerations, we choose a MicroChip sniffer as a receiver. The network architecture is presented in Figure 12.

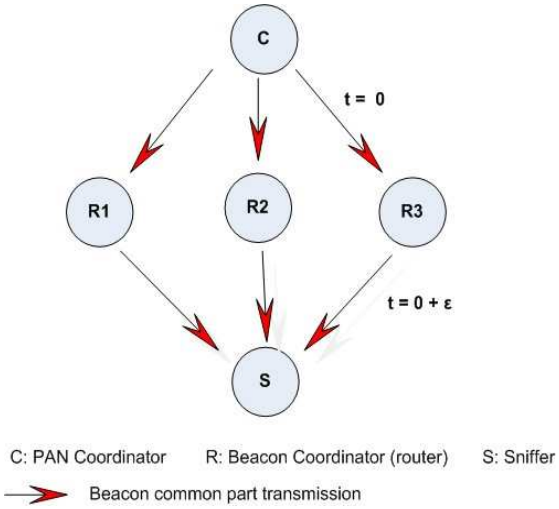


Figure 12. Network architecture for the test.

The receiver is in the transmission range of the PAN Coordinator and the three routers, i.e., the sniffer's software shows two frames which will allow us to measure the time offset between the reception and the transmission of the frame. When a router receives the frame from the PAN Coordinator, it retransmits it immediately. As it is shown in Figure 13, the sniffer receives two frames: the first one is the

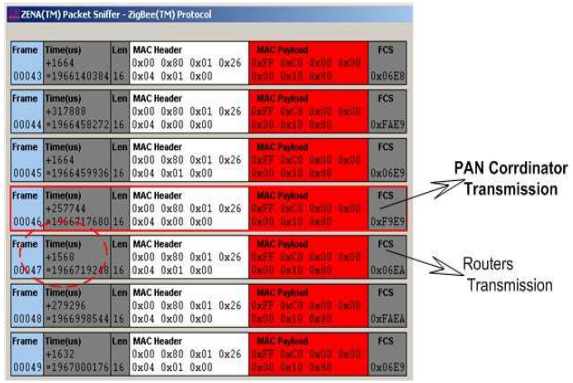


Figure 13. Frames received by the MicroChip sniffer (same bit configuration).

frame transmitted by the PAN Coordinator and the second one is the frame transmitted by the routers. The sniffer receives only one frame from the routers although there are three transmissions. Consequently, the receiver considers all the transmissions as only one transmission.

This receiver capability in processing multipath avoids introducing beacon or SuperFrames scheduling mechanisms.

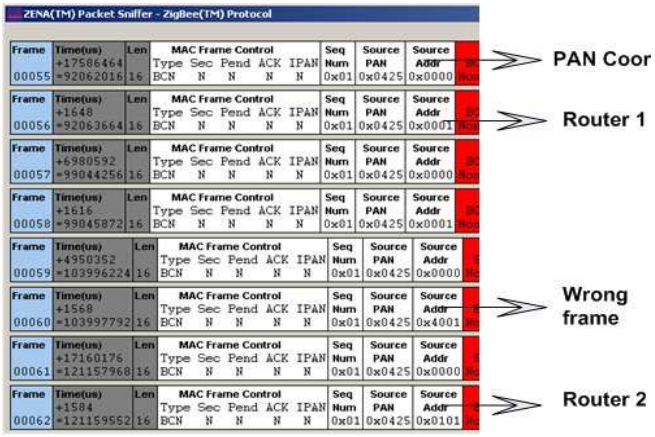


Figure 14. Frames received by the MicroChip sniffer (different bit configuration).

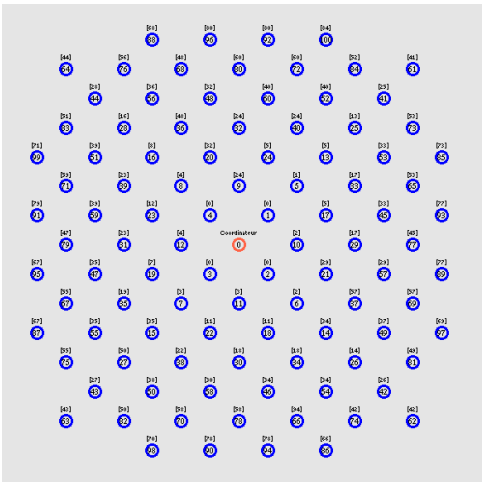


Figure 15. Network topology for NS2 simulation.

To show the impact of receiving different frames by the sniffer, the same network architecture as in Figure 12 is considered. Except, we are using only two routers instead of three. The frame transmitted by each router is different.

Instead of using the same bit configuration, each device (routers and PAN coordinator) in the network uses its 16-bits address. Only the addresses are different, the rest of the frame is the same for both devices.

Figure 14 shows the frames received by the sniffer when the frames are different. The PAN Coordinator's address is 0x0000, router1's address is 0x0001 and router2's address is 0x0101. From Figure 14 we can see that the sniffer receives only one frame instead of two frames. In addition, the received frame could be corrupted. In the figure, the sniffers interprets a received frame as a frame sent by a node with the address 0x4001 which does not exist in the network.

B. Network simulation

In this section, the presented approach is implemented in NS2 simulator. NS2 source code for IEEE 802.15.4 [13] is modified to support the mechanisms presented in this paper. The goal of this section is to show that our approach is implementable on a sensor node. Performances comparison is not presented.

The considered network topology under NS2 is presented in Figure 15. There are 101 nodes and all the devices are routers that accept nodes association.

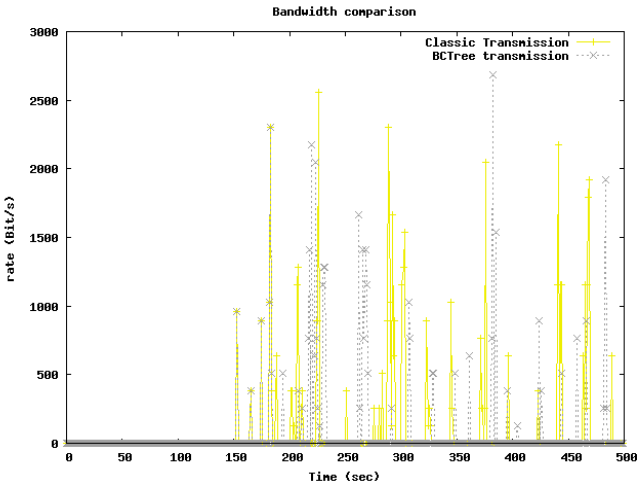


Figure 1. Bandwidth comparison (NS2 implementation vs our approach).

Our approach does not introduce any changes on association mechanisms. During the association, classic beacon frames are transmitted. Once the network is established, the beacon common part is used to give the start signal of the SuperFrame and the specific beacon part is used to send information concerning GTS, Final CAP Slot, pending addresses, etc.

Figures 16 and 17 represent bandwidth measures are represented. Figure 16, shows a comparison between bandwidth for a transmission from node 20 to node 64. It is a FTP over TCP traffic transmitted first using our approach and in a second time using the original implementation of the IEEE 802.15.4 in NS2. Both transmissions starts at t=150 sec and continue until approximately t=500 sec. In this simulation, we used the same BO and SO values in both cases (our approach and IEEE 802.15.4 NS source code). The goal is to validate the changes introduced in the IEEE 802.15.4 NS2 source code.

 ISBN: 978-1-61208-100-7

Figure 17 shows the bandwidth of different traffic flows. Four FTP over TCP flows are considered: node 20 to node 64, node 31 to node 24, node 33 to node 50 and node 80 to node 39.

Thus, from Figures 16 and 17 we can say that the introduced approach does not affect frames transmission mechanism.

VIII. CONCLUSION AND FUTURE WORKS

In this paper, we presented a new approach for the construction of ZigBee/IEEE 802.15.4 Cluster-Tree networks. The presented approach tackles the problems of beacon frames and SuperFrames scheduling. It allows the construction of Cluster-Tree topology without introducing constraints on SuperFrames structure and without taking into account the nodes density in the network.

We proposeed a collision-free beacon transmission approach that exploits node's capabilities in extracting the information from a signal perturbed by simultaneous transmissions of several beacon coordinators. These transmitted RF signals are considered as reflected RF signals since all the beacon coordinators are broadcasting the same signal.

Future works will deal with adapting the presented approach to enable the construction of Beacon mesh networks. For time sensitive applications a GTS collision avoidance mechanism must be introduced to grant the GTS traffic.

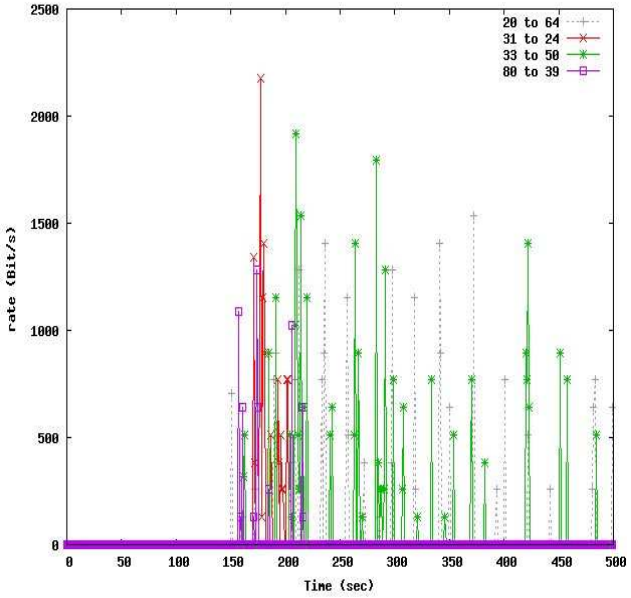


Figure 2. Bandwidth in the case of multiple transmissions.

REFERENCES

[1] ZigBeeAlliance, (Jan. 2008), ZigBee Specification [Online]. Available: www.zigbee.org

[2] IEEEComputerSociety, (Sept. 2006), Part 15.4: Wireless Medium Access Control (MAC) and Physical Layer (PHY) Specifications for Low-Rate Wireless Personal Area Networks (WPANs) [Online]. Available: http://standards.ieee.org/getieee802/802.15.htm

[3] M. Ilyas and I. Mahgoub, Handbook of sensor networks: compact wireless and wired sensing systems, CRC PRESS, 2004

[4] M. I. Benakila, L. George, and S. Femmam, "A Beacon-Aware device for the interconnection of ZigBee networks", in Proc. IFAC 2009, May 2009.

[5] P. Jurcik, R. Severino, A. Koubaa, M. Alves, and E. Tovar, "Real-Time Communications over Cluster-Tree sensor networks with mobile sink behaviour", in Proc. RTCSA 2008, Aug. 2008

[6] A. Koubaa, M. Alves, and E. Tovar, "Modeling and Worst-Case dimensioning of Cluster-Tree wireless sensor networks", in Proc. RTSS 2006, 2008.

[7] V. Mhatre and C. Rosenberg, "Design guidelines for wireless sensor networks: communication, clustering and aggregation", Ad Hoc Networks.

[8] G. Gupta and M. Younis, "Fault-Tolerant Clustering of wireless sensor networks", in Proc. WCNC 2003, Mar. 2003.

[9] T.G.15.4b, (2004), http://www.ieee802.org/15/pub/TG4b.html [Online]

[10] A. Koubaa, A. Cunha, and M. Alves, "A time division beacon scheduling mechanism for IEEE 802.15.4/ZigBee Cluster-Tree wireless sensor networks", in Proc. ECRTS 2007, July 2007.

[11] J. Francomme, G. Mercier, and T. Val, "Beacon synchronization for GTS collision avoidance in an IEEE 802.15.4 meshed network", in Proc. IFAC 2007, 2007.

[12] CISCO, (2008), Multipath and Diversity: http://www.cisco.com/application/pdf/paws/27147/multipath.pdf [Online]

[13] J. Zheng and M. J. Lee, "NS2 simulator for 802.15.4 (release v1.1)", http://www-ee.ccny.cuny.edu/zheng/pub/.

 ISBN: 978-1-61208-100-7

Modeling Energy Consumption for RF Control Modules

Barış ORHAN
VESTEL Electronics
Manisa, Turkey
baris.orhan@vestel.com.tr

Engin KARATEPE
Dept. of EEE, Faculty of Engineering
Ege University Izmir, Turkey
engin.karatepe@ege.edu.tr

Radosveta SOKULLU
Dept. of EEE, Faculty of Engineering
Ege University Izmir, Turkey
radosveta.sokullu@ege.edu.tr

***Abstract* — Due to the restricted energy capacity of wireless nodes, development of strategies and protocols to reduce power consumption is a current research topic. Providing a simulation model of the power consumption of a wireless node taking into consideration the specifics of its hardware platform profile and the algorithm, which is running on it before the design stage will help designers predict the possible problems and adaptivity of the software and hardware with less time and effort. In light of this, the article presents a Semi-Markov Chain based model and the related lifetime analysis of a wireless module operating in an event-triggered application scenario. The wireless node's operation pattern is modeled mathematically and equations are derived to provide its lifetime prediction using MATLAB. The lifetime is examined as a function of the channel noise and the arrival rate with defined hardware and software parameters. A case study is presented with simulation results and real life measurements, based on SoC (System on Chip) ICs (Integrated Circuits), communicating using RF4CE(Radio Frequency for Consumer Electronics) protocol.**

Keywords - Energy consumptio, Node Lifetime, RF4CE, Semi Markov Chains

I. INTRODUCTION

In the last couple of years, a combination of consumer demand for more efficient integrated home networking systems and a steep drop in the price of hardware fueled by manufacturing process improvements has resulted in a noticeable upward cycle of research in the field of RF (Radio Frequency) devices for consumer electronics. Because of the increased range of applications and the inherent physical restrictions, wireless networks and RF communications continue to attract the attention of the research community. Major design factors like limited power capacity and long lifetime necessity combined with the variability of the wireless communication media push further the search for new hardware platforms to meet these requirements. This continuous cycle of adapting new platforms to new wireless applications requires minimization of the design and evaluation time, which in turn creates the demand for suitable models and tools to assure that.

In this paper a numerical and simulation model for the estimating the lifetime of wireless RF control modules used in consumer electronics is presented. The paper is organized as follows: in the first part a short survey of related work is provided followed by detailed description of the proposed mathematical model. Section IV presents real life measurements taken to validate the proposed model. In Section V, the simulation results are presented followed by discussion on the specific relation between the RF module's hardware platform lifetime, the event arrival rate and the communication channel characteristics. Hardware and software parameters are defined for the MATLAB simulation model based on datasheets of the considered hardware platforms and the RF4CE specifications. [1-3]

II. RELATED WORK

The technological restrictions inherent to the IR (InfraRed) devices used in consumer electronics, together with the fact that the components used in the production of the RF devices have come to a price level suitable for mass production have triggered the design and development of new RF devices for consumer electronics [4]. Among the major advantages the RF technology brings to the market are the increased functionality and the potential for reducing power consumption. This in turn makes the question of wireless modules' power consumption design and analysis an important research issue [5]. However, research on wireless nodes so far has focused mainly on extending the lifetime of the network as a whole, thus research attention has been predominantly given to the MAC(Media Access Control) layer and network layer operation, protocols and optimization [8-10]. Together with this, little has been done related to modeling, studying and comparing the performance of different wireless node platforms from the point of view of long term lifetime evaluation. In some recent sources [6, 7, 11] the question of power consumption of wireless modules at the hardware level and its more detailed

 ISBN: 978-1-61208-100-7

functioning has been modeled and investigated with the purpose of defining ways for its possible reduction. Several different suggestions can be found regarding the way power consumption at the hardware level should be modeled. In [6] analysis of the network wide power consumption is carried out based on hybrid automata modeling and the calculation of the consumption of each node. The lifetime of the network is evaluated as a function of the distance from the nodes to the sink. The authors of [7] propose Markov Chains to be used where each possible state of the operation cycle is modeled as part of the chain. For their study the authors define 6 energy states and the transitions between them. In another work, [11], details of PowerTOSSIM, a simulation tool specifically oriented towards modeling power consumption in wireless sensor networks is presented. In calculating power consumption PowerTOSSIM provides possibilities for taking into consideration the details of the hardware characteristics of the nodes. Different from these the authors of [12, 13] present a model for evaluating the network lifetime based only on the hardware characteristics of the nodes and the application scenario. According to them the combination of hardware characteristics and the requirements imposed by the application scenario are the most important factors in determining the lifetime of a wireless node.

Wireless sensor networks design and operation is strongly application oriented. Thus, since most applications can be classified as either monitoring or event tracking, their operation can be related to either scheduled or event triggered scheme. For the RF control module under consideration the wireless node operates in an event triggered mode. In this mode the node spends a comparatively long time in sleep or idle mode in order to preserve energy and wakes up for receiving or transmission only when there is an event detected. After processing the event and transmitting the required information to the sink or to a neighbor node it returns to low power state. On the other hand monitoring applications require that the node awakes at specific predefined intervals to collect and process information. While the aim in event triggered operation is optimizing both the sleep and processing energy consumption, in schedule driven operation most important is optimizing the schedule of each individual node in order to increase the lifetime of the network as a whole.

In this work, we present a Semi-Markov Chain based model of a wireless module, the new generation RF remote control module, which has recently become very popular in consumer electronics. Furthermore we use this model and the specific hardware characteristics of two well accepted hardware platforms to analyze its long term power consumption, especially stressing on its relation to the event arrival rate and the conditions of the transmission environment. The model is based on wireless SoC (integrated CPU(Central Processing Unit) and radio modules), ,which are among the latest achievements in consumer electronics hardware and the only accepted so far in this field network communication standard, RF4CE [1].

III. MODEL OVERVIEW

In this section, the proposed Semi-Markov Chain model of the wireless node is discussed. A wireless sensor node's main functions are to sense an event and transmit information about it. For event based applications, like the one discussed in this paper, each event will trigger a sequence of operations to be carried out, which form an application specific execution cycle. Furthermore, the cycle can be decomposed into a series of computational and communication tasks. Each of the tasks in its turn can be further broken down into states related to the operation of the CPU, the radio and the specific communication standard used like receiving, listening to the channel, transmitting and processing. Each state can be uniquely associated with a specific power mode. These states (power modes) are defined in Table 1. The long term power consumption of the node will depend on the specific state, its duration and the number of times it is executed. So our aim is to model these states from their power consumption point of view taking into account their duration and frequency of occurrence. For this purpose the following assumptions have been made: a) the arrival of the events is assumed to be Poisson distributed b) the duration of time a node spends in each of the states is random.

TABLE 1 POWER STATES OF THE RF MODULE

Power Mode	CPU	Radio
S1	OFF	OFF
S2	ON	OFF
S3	IDLE	TX
S4	IDLE	TX
S5	IDLE	TX
S6	IDLE	OFF
S7	IDLE	RX
S8	IDLE	RX
S9	IDLE	RX

In the long run, the randomly distributed times a node will be found in each of the above states and the duration spent in each state for a certain time interval will reach a steady state. Based on the assumptions made in [12], such a process can be described as a Semi-Markov process with Poisson event arrivals. A Semi-Markov process is a stochastic process that changes states in accordance with a Markov chain but takes a random amount of time between changes.

According to the application discussed, the operation of the wireless node is event-triggered and so it keeps its CPU and radio at the lowest power level to reduce power consumption until it detects an event. Thus a simple Semi-Markov Chain model is given in Fig. 1.

For the transmission procedure, according the RF4CE standard, CSMA(Carrier Sense Multiple Access) is used with ARQ(Automatic Repeat Request) retransmission mechanism. States S1, S2, S3 correspond to sleep, processing and transmission, while S4 and S5 are virtual states, which

 ISBN: 978-1-61208-100-7

represent the ARQ mechanism. The maximum number of retransmissions without acknowledgment is 3. If unsuccessful the packet is dropped and the node returns to its initial state.

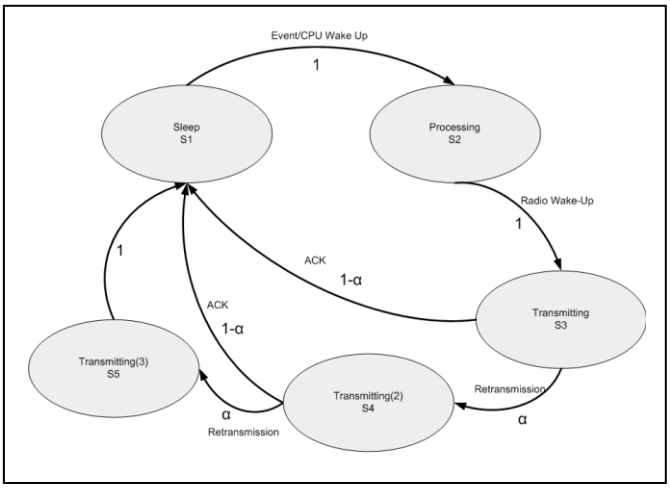


Figure 1.Simple Semi-Markov Chain model of event triggered operation

As defined by the algorithm of the underlying MAC protocol radios should sense the channel before transmission they tend to spend considerable energy listening to the channel. In [14] it is stated that nodes using IEEE 802.15.4 spend less than 50 percent of the energy for actual transmission, while channel listening accounts for more than 40 percent of the energy consumption. During the transmission radios are actually switching between listen and transmit states (S7, S8, S9) because of the requirement to sense the channel before actual transmission of the packet and to listen waiting for acknowledgement after the packet has been transmitted. On the other hand, the CPU makes transitions from processing to idle state while processing (S2-S4). By adding these factors into account, the Semi-Markov diagram is further detailed and updated with embedded chains as shown in Fig. 2.

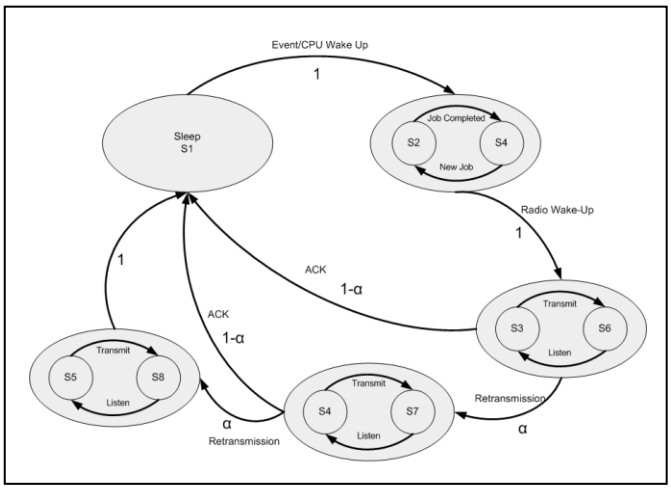


Figure 2.Semi-Markov Chain model with embedded chains

Based on the assumptions and the state definitions for the embedded Semi-Markov Chain made above we now proceed with the analytical model. Let X(t) denote the power state at time t, then $\{X(t), t\geq 0\}$ is a Semi-Markov process [15], where the expected value is calculated as;

$$\mu_i = \int_0^\infty x H_i(x) dx \quad (1)$$

and H_i is the random time the semi-Markov process spends in state *i* before making a transition; X_n denotes the nth state visited by the Markov Chain. X_n, $n\geq 0$ forms a Markov Chain with transition probability p_{ij}, where $T_{ii}$is the time between successive transitions into state *i* and its expected value is

$$\mu_{ii} = E[T_{ii}] \quad (2)$$

LetT_t, be the total time in *i* during [0,t], then the overall time spent in *i* over the combined time spent in all states (long term amount of time in *i*) is given by:

$$p_i \equiv \lim_{t\to\infty} P[X(t) = i | X(0) = j] = \lim_{t\to\infty} \frac{T_t}{t} \quad (3)$$

If we suppose that the embedded Markov Chain is positive recurrent for $n\geq 0$, a stationary probability exists, which is the frequency of visiting each state for a given infinite time duration. Then π_j is the stationary distribution of the embedded Markov chain $j\geq 0$. π_j has a unique solution:

$$\pi_j = \sum_i \pi_i p_{ij}, \sum_j \pi_j = 1 \quad (4)$$

π_j, can be interpreted as the proportion of transitions into state *j* over the sum of all state transitions. Then the following equation holds:

$$p_i = \frac{\mu_i}{\mu_{ii}} = \frac{\pi_i \mu_i}{\sum_j \pi_j \mu_j} \quad (5)$$

The proportion of time spent in *i* to the time spent in all states could be found using (4) and (5). Considering the initial simple model in Fig. 1 and using (2), we can compute:

$$[\pi_1\ \pi_2\ \pi_3\ \pi_4\ \pi_5] \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 1-\alpha & 0 & 0 & \alpha & 0 \\ 1-\alpha & 0 & 0 & 0 & \alpha \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix} = [\pi_1\ \pi_2\ \pi_3\ \pi_4\ \pi_5] \quad (6)$$

$$\sum_{1\leq j\leq 5} \pi_j = 1 \quad (7)$$

(6) and (7) has the unique solution

$$\pi_1 = \pi_2 = \pi_3 = \frac{1}{\alpha^2+\alpha+3}$$

$$\pi_4 = \frac{\alpha}{\alpha^2+\alpha+3} \quad (8)$$

$$\pi_5 = \frac{\alpha^2}{\alpha^2+\alpha+3}$$

Let p_i be the steady state of Semi-Markov Chain, $\hat{X}$ average sleep time, $\hat{Y}$ average processing time and $\hat{Z}, \hat{V}, \hat{W}$ average communication times, then:

$$D_\alpha = \hat{X} + \hat{Y} + \hat{Z} + \alpha\hat{V} + \alpha^2\hat{W} \quad (9)$$

$$p_1 = \frac{\hat{X}}{D_\alpha}, p_2 = \frac{\hat{Y}}{D_\alpha}, p_3 = \frac{\hat{Z}}{D_\alpha}, p_4 = \frac{\alpha\hat{V}}{D_\alpha}, p_5 = \frac{\alpha^2\hat{W}}{D_\alpha} \quad (10)$$

If T is long enough, the total time elapsed in state *i* is $\lim_{T\to\infty} T_i = Tp_i$. Thus, the total consumed energy in state *i*. is $E_{S_i} = Tp_i \times P_{S_i}$ ($i \in \{1,2,3,4,5\}$) and also transition energy from state *i* to *j* can be calculated by multiplying one transition cost by the average number of transitions from state *i*. to *j*. In state transition calculations, we will only use the wake up costs ($C_p(E_{S_{12}})$ ve $C_r(E_{S_{23}})$) because the sleep cost for both the CPU and the radio is negligible.

Thus, to summarize, the wireless node's long term lifetime estimation can be defined based on three distinct components:

$$A_1 = \{n | X_j = S_1, X_{j+1} = S_2, X_{j+2} = S_3\}$$

$$A_2 = \{n | X_j = S_1, X_{j+1} = S_2, X_{j+2} = S_3, X_{j+3} = S_4\}$$

$$A_3 = \{n | X_j = S_1, X_{j+1} = S_2, X_{j+2} = S_3, X_{j+3} = S_4, X_{j+4} = S_5\}$$

Furthermore, as can be seen from the model in Fig. 1 and Fig. 2, the CPU and the radio have to wake up for entering states S2 and S3. Since $\mu_{ii} = D_\alpha$, average number of cycles during T is $C_T = \frac{T}{D_\alpha}$

C_T : Total transition energy cost during T
k(T): average cycle number during T

By observing the limiting behavior of the function in (11),

$$\hat{P}_{td} = \lim_{T\to\infty} \frac{1}{T}[\sum_{1\le k\le 5} E_{S_k} + C_T] \quad (11)$$

the total amount of energy spent at each state, E_{S_i} and the transition energy, C_T over *T*, the average power consumption for an event-triggered node can be calculated as follows:

$$\hat{P}_{td} = [p_1 P_{S_1} + p_2 P_{S_2} + p_3 P_{S_3} + p_4 P_{S_4} + p_5 P_{S_5} + \frac{C_p + C_r}{D_\alpha}] \quad (12)$$

$$\hat{P}_{td} = [\frac{\hat{X}}{D_\alpha} P_{S_1} + \frac{\hat{Y}}{D_\alpha} P_{S_2} + \frac{\hat{Z}}{D_\alpha} P_{S_3} + \frac{\alpha\hat{V}}{D_\alpha} P_{S_4} + \frac{\alpha^2\hat{W}}{D_\alpha} P_{S_5} + \frac{(C_p+C_r)}{D_\alpha}])$$

$$\hat{P}_{td} = [\frac{\hat{X}P_{S_1} + \hat{Y}P_{S_2} + \hat{Z}P_{S_3} + \alpha\hat{V}P_{S_4} + \alpha^2\hat{W}P_{S_5} + (C_p+C_r)}{\hat{X}+\hat{Y}+\hat{Z}+\alpha\hat{V}+\alpha^2\hat{W}}] \quad (14)$$

From the initial model presented in Fig. 1 we can deduct the time dependent state/energy consumption representation given in Fig. 3 for the power profile of an event-triggered wireless node. The average processing and communication time is very small compared to the average sleep time (time when the module is waiting for an event), thus the average sleep time for one cycle is equal to the event inter-arrival time.

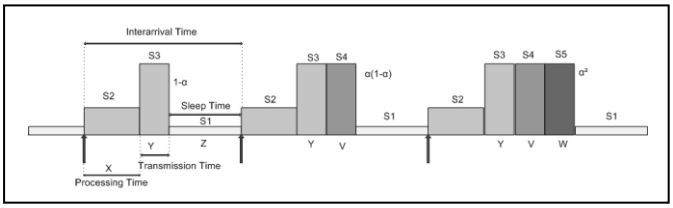


Figure 3.Basic power profile of the event triggered node

If λ denotes the arrival rate and $\hat{X} = 1/\lambda$ then:

$$\hat{P}_{td} = [\frac{P_{S_1} + \lambda(\hat{Y}P_{S_2} + \hat{Z}P_{S_3} + \alpha\hat{V}P_{S_4} + \alpha^2\hat{W}P_{S_5} + (C_p+C_r))}{1+\lambda(\hat{Y}+\hat{Z}+\alpha\hat{V}+\alpha^2\hat{W})}] \quad (13)$$

Thus (13) is the direct analytical expression corresponding to the basic model in Fig. 1. However, taking into account the more detailed model in Fig. 2 we can derive the time dependent representation given in Fig. 4. It has been extended with the inclusion of the embedded states S6, S7, S8 and S9 for the processing and communication tasks.

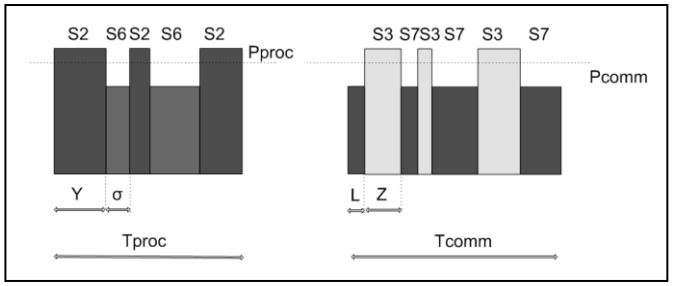


Figure 4.Embedded chains of Markov-chain model

So, the average processing power $\overline{P}_{proc}$ and the average processing time, $\overline{T}_{proc}$, respectively $\overline{P}_{comm}$ and $\overline{T}_{comm}$ are:

$$\overline{P}_{proc} = \frac{\overline{\sigma}P_{S_6} + \hat{Y}P_{S_2}}{\overline{\sigma}+\hat{Y}} \quad (15)$$

$$\overline{T}_{proc} = (\overline{\sigma} + \hat{Y})\,\overline{N}_\sigma \quad (16)$$

$$\overline{P}_{comm} = \frac{\overline{L}P_{S_7} + \hat{Z}P_{S_3}}{\overline{L}+\hat{Z}} \quad (17)$$

$$\overline{T}_{comm} = (\overline{L} + \hat{Z})\,\overline{N}_L \quad (18)$$

where $\overline{\sigma}$ is the average time the CPU spends in idle states, $\overline{L}$ is the average time the radio spends listening, $\overline{N}_\sigma$ is the average number of CPU's idle states, $\overline{N}_L$ is the average number of the radio's idle states. Thus for the average power consumption using (15), (16), (17) and (18) in (14), we get (19):

 ISBN: 978-1-61208-100-7

$$\widehat{P}_{td} = [\frac{P_{S_1}+\lambda((\bar{\sigma}P_{S_6}+\hat{Y}P_{S_2})\bar{N}_{\sigma}+(\bar{L}P_{S_7}+\hat{Z}P_{S_3})\bar{N}_L+\cdots}{1+\lambda((\bar{\sigma}+\hat{Y})\,\bar{N}_{\sigma}+(\bar{L}+\hat{Z})\,\bar{N}_L+\cdots}$$

$$\frac{\alpha(\bar{L}P_{S_8}+\hat{Z}P_{S_4})\bar{N}_L+\alpha^2(\bar{L}P_{S_9}+\hat{Z}P_{S_5})\bar{N}_L+(C_p+C_r))}{\alpha(\bar{L}+\hat{Z})\,\bar{N}_L+\alpha^2(\bar{L}+\hat{Z})\,\bar{N}_L)}] \qquad (19)$$

Dividing the total power by equation derived in (19), the average lifetime for an event-triggered node is calculated as:

$$\widehat{T}_l(\lambda) = \frac{E_{TOTAL}}{\widehat{P}_{td}} \qquad (20)$$

IV. MODEL VALIDATION

In this part, we present some real life measurements in support of the suggested model, taken using wireless node platforms CC2530 and MC13213. The experimental setup includes a serial 10Ω load. The resulting time-voltage graphs in response to an event are given in Fig. 5 and Fig. 6. When turning the radio on, the node performs a sequence of transitions between different states as shown below.

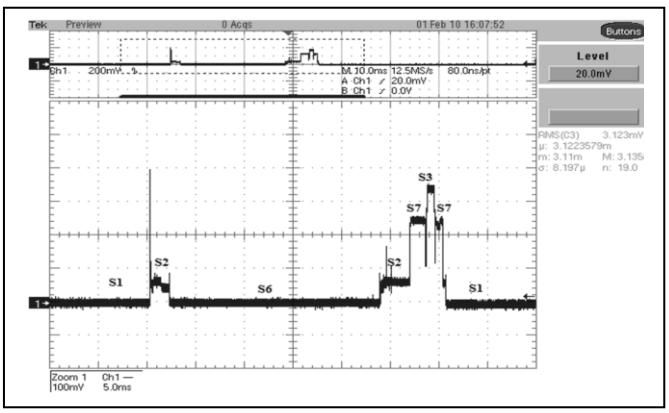


Figure 5.Oscilloscope output for a noise-free channel

Fig. 5 corresponds to the case when the channel is clear and the first time a packet is transmitted a positive acknowledgement is received. Fig. 6 presents the case when

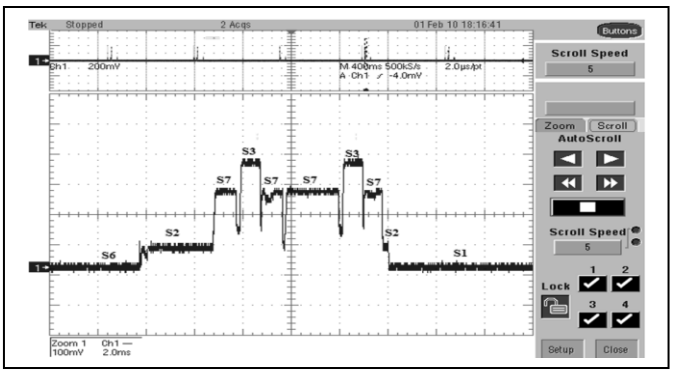


Figure 6.Oscilloscope output for a noisy channel

due to a noisy channel the first transmission was unsuccessful and following retransmissions were required. The states S1 to S7 are as defined earlier in Table 1. The measurements support our selection and definition of the states and the required transitions, given in Fig. 1 and Fig. 2. When the node is not in use it is in state S1, which is the lowest power state, in order to enable long battery life. This state is defined as the deepest possible sleep, using min power. However, coming out of that state is inherently associated with unwanted effects as input jitter. When an event occurs (a key is pressed), the device wakes up on a key press and sets the sleep timer to expire in 25 ms, entering back sleep mode. After this timer expires, the CPU wakes up, reads the pressed key, prepares the packet to be sent, the radio listens to the channel and transmits waiting for acknowledgement. Upon receiving the acknowledgement the CPU and the radio enter back into deep sleep state (S1). In case the first attempt to send the packet is unsuccessful (Fig. 6), the radio and CPU will return to state S7, attempt a second and third retransmission and if again unsuccessful will force return to sleep state. So, when retransmission is required, it is visible that the node does not return to sleep state S1 between the successive trials. It only returns to the initial, low power state (S1) after a successful transmission or if the maximum number of allowed retransmission attempts is reached, which proves the need to define additional states S7, S7 and S9.

V. SIMULATION RESULTS

In this section, the RF control module lifetime is investigated for two different scenarios for the two wireless platforms CC2530 and MC13213. Our main goal is to examine the dependence of the lifetime as a function of the event arrival rate and the channel conditions and then compare the two platforms.

Fig. 7 and Fig. 8 present the results as the number of arrival events per hour increases from 0 to 40 in case of transmission success ratios set to 0.2 and 0.8 respectively. We notice that the lifetime decreases for both platforms as the rate increases and the channel quality is decreased. While the CC2530 platform achieves a determinedly longer lifetime for low arrival rates (5 – 7 events per hour), the decrease in lifetime duration for high event arrival rate (35 events per hour) is higher (nearly 71% compared to 55% for the MC13213 platform). In the long run, the CC2530 platform is more sensitive to channel noise and event arrival rate than the MC13213 platform. Results are given in Table 2.

TABLE 2 PERCENTAGE LIFETIME COMPARISON

Platform	α		λ	
	0.2	**0.8**	**12**	**4**
CC2530	71%	78%	25%	17%
MC13213	60%	64%	20%	11%

 ISBN: 978-1-61208-100-7

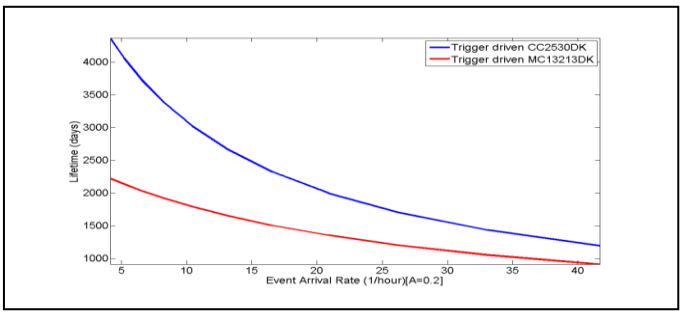


Figure 7.Lifetime prediction for α=0.2

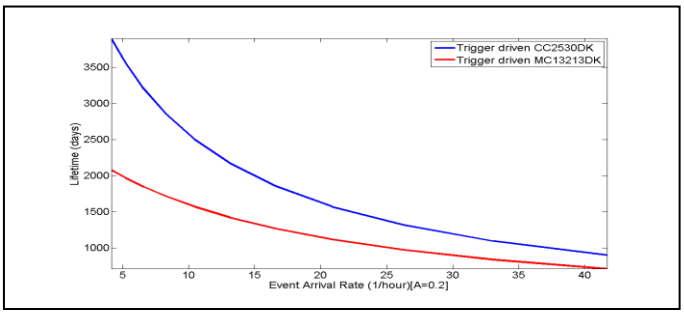


Figure 8.Lifetime prediction for α=0.8

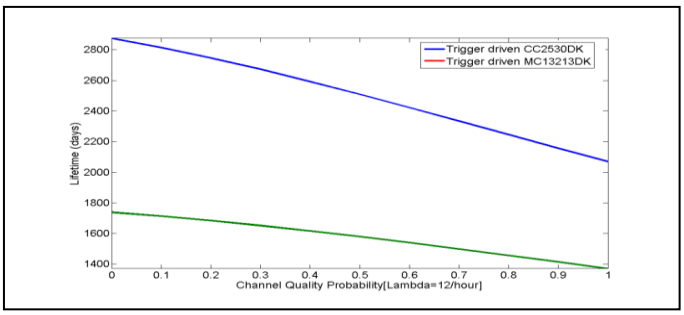


Figure 9.Lifetime prediction for λ=12

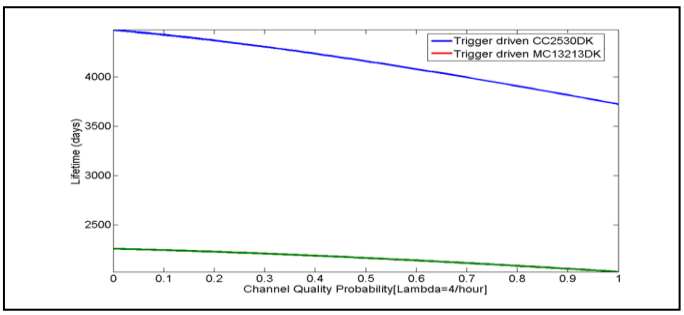


Figure 10.Lifetime prediction for λ=4

Fig. 9 and Fig. 10 give the lifetime prediction for the two platforms as a function of the channel conditions (transmission success ratio) for two fixed event arrival rates 4 and 12 events per hour. MC13213 platform shows less sensitivity to both varying arrival rate and channel conditions. However CC2530 has a longer lifetime for all conditions better performance for all cases. It is worth nothing that the arrival rate is more determinant parameter than channel noise.

VI. CONCLUSION

In this paper we have suggested a mathematical model, based on embedded Semi-Markov Chains, for the evaluation of the lifetime of an RF wireless module operating in an event-triggered mode. The model is validated using real time measurements for two hardware platforms compatible with the newly approved standard, RF4CE, for RF devices for consumer electronics. Furthermore, simulation results have been presented showing the dependence of the two platforms on the event arrival rate and the channel conditions with the assumption that CSMA with AQR is used.

VII. ACKNOWLEDGEMETS

The work presented in this paper was partially supported by SAN-TEZ funds, project no 00446.STZ.2009–2.

REFERENCES

[1] ZigBee RF4CE Specification, ZigBee Alliance document 094945r00ZB

[2] CC2530 Datasheet, SWRS081A–April 2009

[3] MC13213 Datasheet, MC1321x Rev 1.8 08.2009

[4] www.audioholics.com/news/editorials/rf-remote-controls-is-it-time-for-you-to-step-up (Accessed: 06.02.2010)

[5] http://techon.nikkeibp.co.jp/article/HONSHI/20071127/143101/ (Accessed 06.02.2010)

[6] S. Coleri, M.Ergen, and T. Koo, Lifetime analysis of a sensor network with hybrid automata modeling, 1st ACM International Workshop, 2002

[7] R. Mini, F.,Machado, M. V., Loureiro, F. A., and Nath, B. Prediction-based energy map for wireless sensor Networks, Elsevier Ad-hoc Networks Journal, 2005

[8] K. Negus, A. Stephens and J. Lansford, HomeRF: Wireless Networking for the Connected Home, IEEE Personal Communications, pp. 20 – 27, 2000

[9] K. Watanabe, M. Ise, T. Onoye, H. Niwamato and I. Keshi, An Energy Efficient Architecture of Wireless Home Network Based on MAC Broadcast and Transmission Power, IEEE Transactions on Consumer Electronics, Vol. 53, No. 1, 2007

[10] Bahareh Gholamzadeh and Hooman Nabovati, Concepts for Designing Low Power Wireless Sensor Network, Vol.35, 2008

[11] Shnayder, V., Hempstead, M., Chen, B., Werner-Allen, G., and Welsh, Simulating the power consumption of large-scale sensor network applications, ACM Conference, 2004

[12] D.Jung, T. Teixeira and A. Savvides, Sensor Node Lifetime Analysis: Models and Tools, ACM Transactions on Sensor Networks, Vol.5, No.1, 2009

[13] D. Jung and A. Savvides, An Energy Efficiency Evaluation for SN with Multiple Processors, Radios and Sensors, pp. 1112- 1120, In the Proc. of the IEEE INFOCOM, 2008

[14] B. Bougard, F. Cathhoor, D.C. Daly, A. Chandrakasan and W. Dehaene, Energy efficiency of the IEEE 802.15.4 standard in dense wireless microsensor networks: Modeling and improvement perspectives. In Proceedings of the Design, Automation, and Test in Europe, pp. 196–20, 2005.

[15] Introduction to Probability Models, Sheldon M. Ross, Elseveir, 9th Edition, 2007

 ISBN: 978-1-61208-100-7

DS-CDMA receiver in Software Defined Radio technology

Admission to the implementation of a RAKE receiver

Wojciech Siwicki
Gdansk University of Technology
Faculty of Electronics, Telecommunications and Informatics,
Department of Radiocommunications Systems and Networks
Gdansk, Poland
e-mail: wojciech.siwicki@eti.pg.gda.pl

Jacek Stefański
Gdansk University of Technology
Faculty of Electronics, Telecommunications and Informatics,
Department of Radiocommunications Systems and Networks
Gdansk, Poland
e-mail: jstef@eti.pg.gda.pl

***Abstract* — Programmable radio is one of the latest trends in the construction of multi-standard receivers. The technology, called Software Defined Radio (SDR), is also an ideal test platform that allows to try out different algorithms of signal receiving. This particular feature led to choose this platform to implement a DS-CDMA receiver (Direct Sequence Code Division Multiple Access). The use of SDR allows for a gradual upgrade of data processing algorithms in terms of correctness of the received signal and, what is equally important, its processing time. This paper mainly focuses on these two aspects. Subsequently, the plans are submitted for further development of the software receiver.**

Keywords - Software Defined Radio; SDR; RAKE; DS-CDMA; phase correction

I. INTRODUCTION

The variety of standards for radio systems make it necessary to construct a mobile terminal that has the technical possibilities of cooperation with different radio standards. This simple concept created an idea of a programmable radio called Software Defined Radio, based on a universal hardware layer, with only a layer of software determining its functionality [1-5].

Based on these assumptions a programmable receiver has been developed for DS-CDMA signals. It consists of a wideband receiver and a personal computer (PC) equipped with a acquisition card. PC with appropriate software is used to control both the receiver and the acquisition card; also it serves for signal processing.

In this article, a basic concept of SDR will be presented; also basics of the RAKE receiver and the purpose of the WMSA filter will be introduced. Next, the physical implementation of DS-CDMA signal receiver and its structure will be shown. Selected parameters of received signal and encountered problems with data processing (initial phase correction, frequency jitter, processing time) will be presented as well.

Finally, the direction of expansion of the programming layer multipath signal reception (RAKE receiver) will be presented.

II. SOFTWARE DEFINED RADIO

As was mentioned, the concept behind Software Defined Radio is to implement – to the greatest possible extent – signal processing blocks of a radio transceiver in software rather than in dedicated hardware. The differences between the classic "analogue" version of a receiver and the programmable one are illustrated, respectively, in Figures 1 and 2 [1], [2].

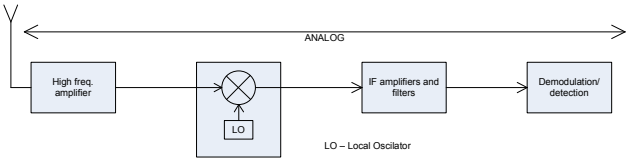


Figure 1. Block diagram of analogue receiver

Receiver shown in Figure 2 can be divided into two different parts of a system:

• hardware (analogue radio) in the form of a set of classic radio components,

• software (digital), whose main element is fast signal processor DSP (Digital Signal Processor).

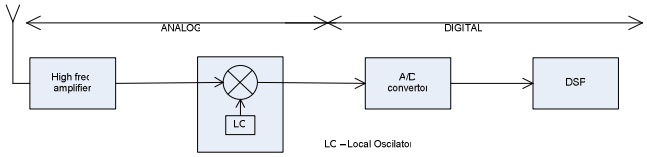


Figure 2. Block diagram of Software Defined Radio

Conceptually speaking, a SDR should have the following properties [4]:

- reconfigurable RX/TX architecture, controlled by software,
- most part of the radio functionality performed by software,

 ISBN: 978-1-61208-100-7

- system specification (bandwidth, bit rate, demodulation) can be updated whenever needed to do so.

The task of the analogue radio part is to strengthen appropriately and convert the received radio signal from the high-band radio frequency to lower frequency band. Then, in this band with a fast A/D converter (Analogue to Digital Converter) a received analogue signal is converted into its digital form. Processing is performed in a properly programmed digital signal processor.

In short, we will find that having a broadband receiver and analogue-to-digital high-frequency sampling with high dynamics, we can get a very comprehensive platform of the receiver, which can perform demodulation/detection of any signal only through the change of software.

III. THEORY OF A RAKE RECEIVER

The mobile radio communication systems, radio wave radiating from the transmitter to the receiver, encounter many obstacles on its way, which means that the signal reaches the receiver in many routes and many random time delays. DS-CDMA receiver is called in the literature as a RAKE receiver. It gives the opportunity of receiving signals in a multi-way propagation environment. In the presented method, due to the use of a spreading sequences and their good correlation properties, the receiver can extract the signal transmitted in several propagation multipath. Therefore the phenomenon of multipath can be used to improve the quality of the received signal.

Technique of direct spread spectrum DS-CDMA creates new opportunities for receiving such signals. If the maximum delay difference between different paths is larger than the duration of one chip, then the temporary reception will allow almost independent reception of each signal. Thanks to this property, the phenomenon of multipath can be used to improve the quality of the received signal.

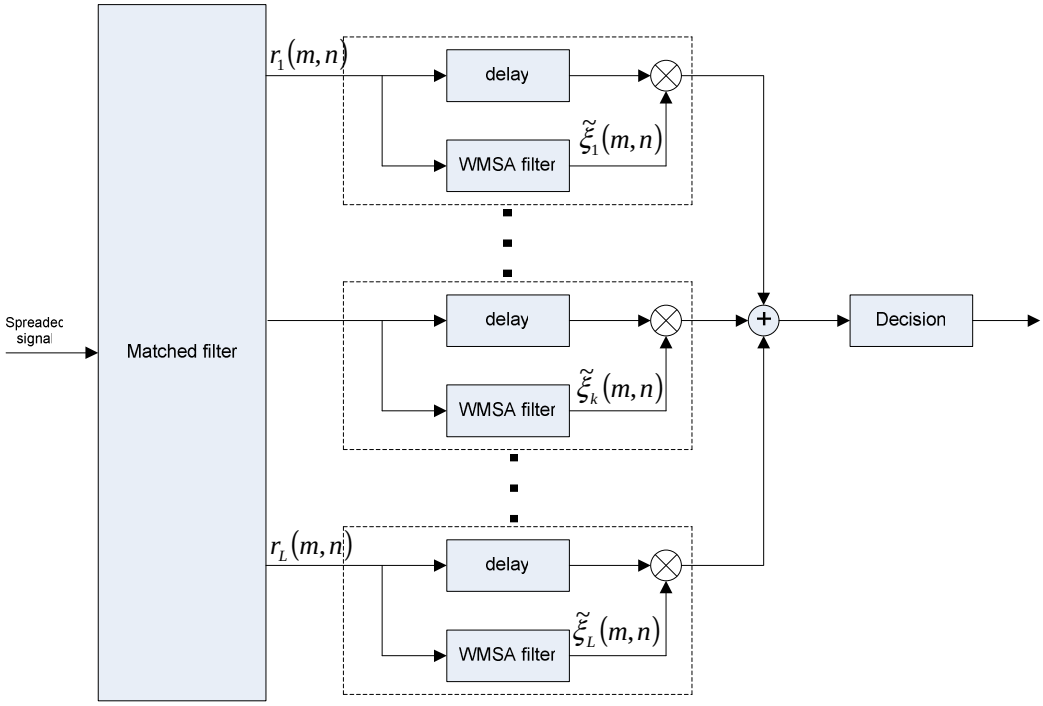


Figure 3. Block diagram of a RAKE receiver ($\xi_L(m,n)$ – channel estimation for l-th branch)

A. Receiver

Analysing the block diagram shown in Figure 3 we can conclude that the RAKE receiver is a set of correlation receivers, acting in parallel. In each branch signal is correlated with the spreading sequence delayed by the same time as the delay introduced by the different signal propagation paths. The signal after phase correction is sent to the decision-making system. Next, the results of decision are gathered in the system of accumulating the results of decisions in order to give them logical values: zero or one. The RAKE receiver is so named because it reminds the function of a garden rake: each branch collects symbol energy like tines of a rake collect leaves.

As known, multipath fading depends on the speed of a terminal [7]. With the increasing speed phase shift estimation is more difficult to estimate [6], [7]. Therefore a chosen method should estimate the phase shift for both slow and rapid fading of the signal. Popular technique filters are operating on linear interpolation, the Gaussian interpolation [8], [9] or Wiener filters [10], [11]. However, Wiener filters require information about the statistical properties of the

 ISBN: 978-1-61208-100-7

existing fading, which in practice is almost impossible. WMSA filter (*Weighted Multi-Slot Averaging*), presented in this paper, is based on the weight variables whose values are linearly interpolated basing on data derived from sequentially received elementary frames.

A simulation model was developed prior to the implementation of a receiver. This method was selected due to its potential low computational complexity which translates to speed of the program. A structure of transferred data, illustrated in Figure 4, was applied for the purpose of simulation. Frame containing the data begins with a known sequence of N_P pilot bits, followed by N_D data bits.

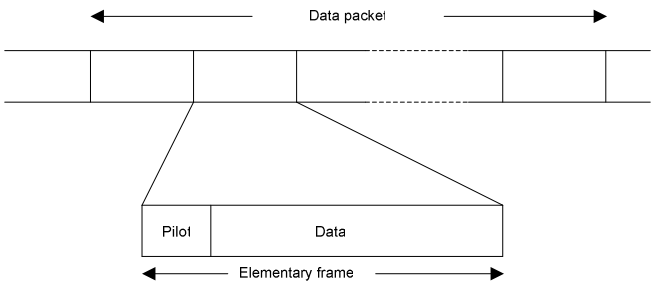


Figure 4. Data packet structure

B. WMSA filter

WMSA filter is based on the weight coefficients of the variables. Block diagram of WMSA filter is shown in Figure 5 [6], [7]. In comparison with linear interpolation filters, *the Gaussian* interpolation filters or Wiener filters, the WMSA filter allows better tracking of fading.

Due to the elimination of fading, the weight factors are changed from frame to frame. This eliminates the problem of proper tracking of the channel fading.

Sets of coefficients have been designated, basing on the simulation results presented in [7]. These factors have been optimized to measure frames and are independent of the position of the symbol. Based on these results, two sets of factors have been appointed. These sets are presented in [6]. The first set of designated factors is referred to as Type I and the second one as Type II. They differ only in the number of branches (respectively six and four).

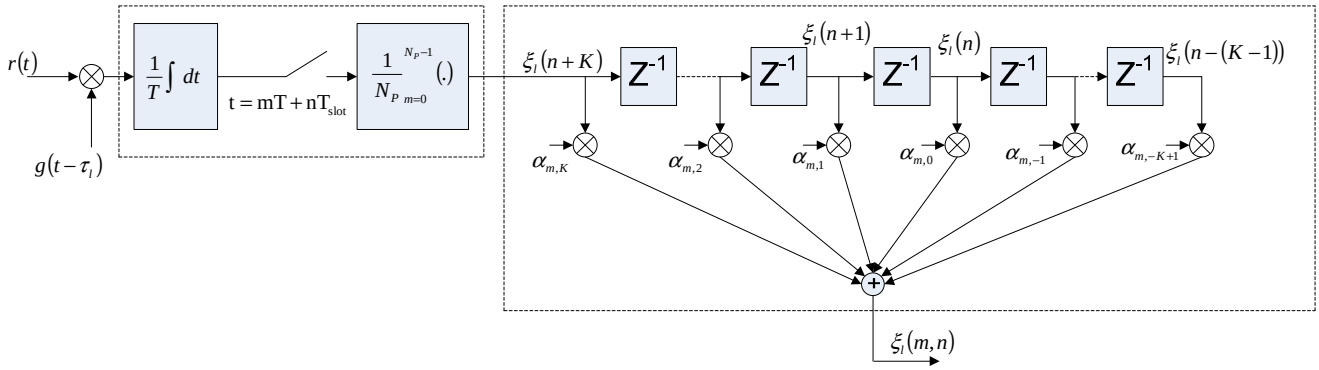


Figure 5. Block diagram of WMSA filter (m – received symbol, n – no of frame, T – time of a symbol, T_{slot} – time of elementary frame $T_{slot}=(N_P+N_D)T$, N_P – number of pilot bits, N_D – number of data bits, r(t) – received signal, g(t) – spreading sequence, τ_l – l chips delay, $\xi_l(n)$ – channel estimation, n-th frame, l-th path, Z^{-1} – 1 frame delay block, $\alpha_{m,K}$ – changing weighting factors, $\xi_l(m,n)$ – channel estimation for l-th branch)

 ISBN: 978-1-61208-100-7

C. *Simulation results*

As a master thesis both of WMSA filters were tested in a simulation environment [12]. The presented results refer to the following signal parameters:

- transmission speed 8kb/s,
- terminal speed 3km/h,
- propagation model Outdoor to Indoor & Pedestrian (according to ITU-R M 1034 recommendation),
- channel model with Rayleigh's fading

Sample results are shown in Figure 6.

Based on the simulations, several important conclusions can be extended [12]. Due to use of WMSA filters, lower ratio of of E_b/N_o is required to keep the same rate of errors. Estimator used in the receiver phase (WMSA) will improve from 1 to 3 dB (the ratio is less than the required E_b/N_o at the same error rate). WMSA filters Type I and Type II differ only in the number of branches. After comparing filter Type I (six branches) with filter Type II (four branches), an improvement of 0,1-0,4 dB is obtained. This is a small difference, and from a practical point of view it can be assumed that the filter Type II is computationally more efficient due to the smaller number of arms and thus a smaller number of operations to be performed.

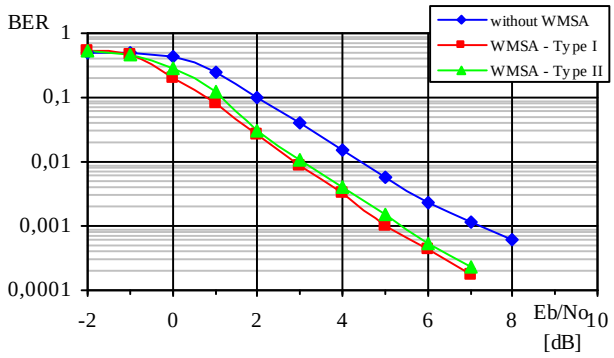


Figure 6. Results of simulation (speed of a receiver 3km/h, transmission speed 8kb/s)

IV. HARDWARE IMPLEMENTATION

Figure 7 shows the diagram of the receiver made in software defined radio technology.

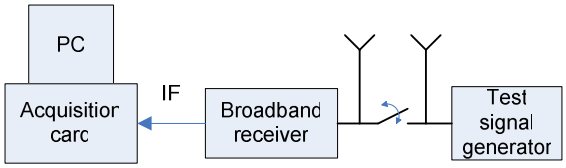


Figure 7. Block diagram of programmable receiver

It consists of a broadband receiver and a PC fitted with a data acquisition card (Figure 8), whereby the analogue radio signal is converted into a sequence of discrete data samples. Computer task is to control the operation of the receiver (the choice of frequency, width filters, etc.) and to process data received from data acquisition card.

Actual received signal parameters are as follows:

- The signal carrier frequency: 450MHz,
- Data rate: 1kHz,
- Bandwidth after spreading: 1MHz,
- Modulation: QPSK,
- In accordance with the structure of the frame (Figure 5): $N_P = 13$, $N_D = 26$,
- Inphase and quadrature components carry independent data

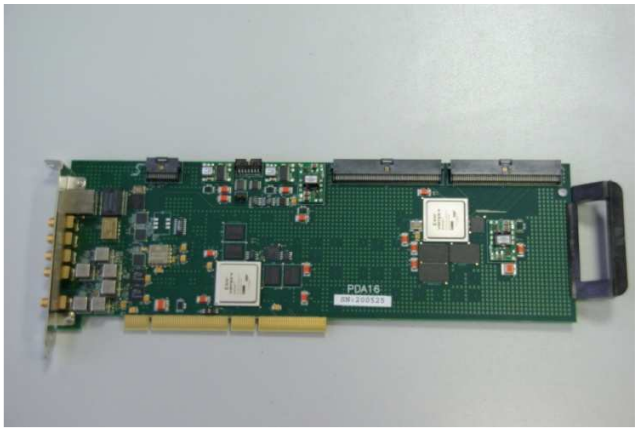


Figure 8. Acquisition card

A. *Algorithm*

The basic algorithm of operation is shown on Figure 9. The program begins with launching and configuring the data acquisition card and setting the parameters of the receiver. Then, the process of acquisition (writing to a binary file data samples) begins. After a certain time, acquisition is completed and acquisition card goes into standby mode. The next step is to process the collected samples. The values of the samples are in the range from 0 to 65536 (16 bit resolution) and they are changed into the range from - 32768 to 32768. Then, the data are divided by the maximum value of level in order to normalize the data (after this, samples are in the range from -1 to +1). Next, samples are multiplied by carrier (both the cosine and sine) and by the spreading sequences. The correlation function, obtained in this way, allows to determine the actual value of the *Realis* component (*Re*) and *Imaginary* component (*Im*). A correction phase is being fixed on their basis. When receiving a bit, sequence is compared with the sent sequence, and on this basis bit error rate (BER) is calculated.

 ISBN: 978-1-61208-100-7

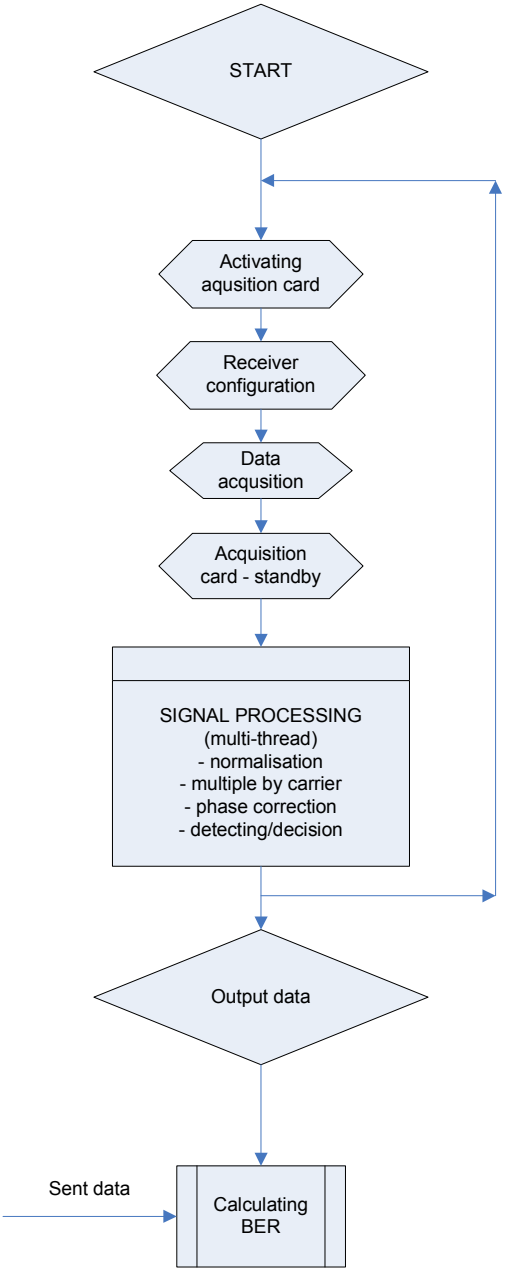


Figure 9. Algorithm

B. *Frequency jitter*

During the processing of data, we should consider errors resulting from inaccuracies in the internal clocks of the receiver and data acquisition card. Correction is being determined after receiving a few bits; next, it is added in the following stage of detection. This frequency drift and phase error can be illustrated as the constellation of the received signal. Due to the independent component in inphase and quadrature, received signal can be detected as two BPSK modulation.

Bearing in mind the above, received signal constellation may take the form shown in Figure 10. Bellow the following sequence of bits has been given for illustration:

bit no.	1	2	3	4	5	6	7
sequence	1	1	1	1	1	0	0

bit no.	8	9	10	11	12	13	14
sequence	1	1	0	…	…	…	…

bit no.	15	16	17	18	19		
sequence	…	0	1	1	0		

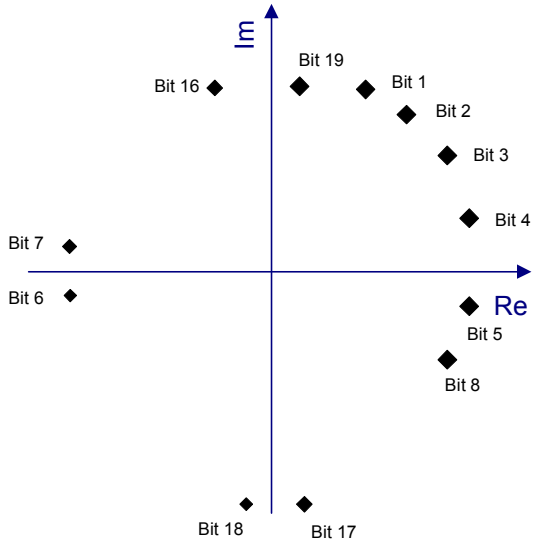


Figure 10. BPSK constellation of a received signal

Analyzing the presented sequence with constellation shown in Figure 10 it can be seen that bits 18 and 19 without phase correction have opposite values. After applying the correction phase (Figure 11), both points are on the side of their actual representation.

 ISBN: 978-1-61208-100-7

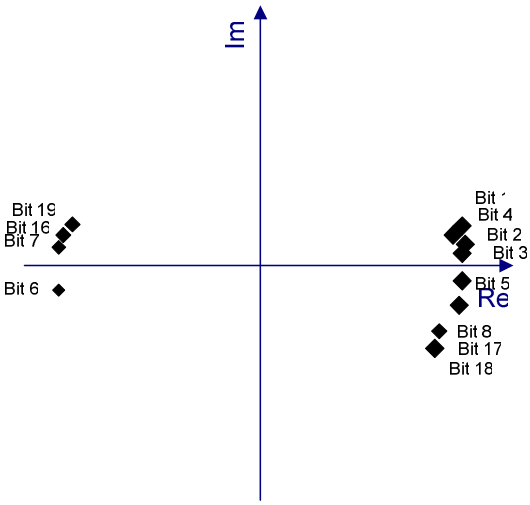


Figure 11. BPSK constellation after phase correction

In the first approach of implementation the phase is calculated on the basis of receiving the first few bits of a known pilot. Currently, tests are conducted on receipt of a number of generators at the same time (using different spreading sequences).

C. *Processing time*

Time of signal processing is one of the most important parameters of a programmable radio. It is recommended to obtain the strongest possible CPU for a PC. If possible, convert each calculating subprogram to the application, using several processor cores. Only the development of multithreaded applications can effectively cope with the complexity of computing with which we meet in implementing a programmable radio. Not only the hardware layer affects the processing speed, it is equally important to optimize the code.

The first attempts to receive transmitted signal took over an hour of processing time (for the acquisition of 40 seconds). At the moment, the software was significantly enhanced (algorithm were optimized, multithread environment was implemented). These treatments resulted in reducing the processing time to approximately 40-60 seconds.

V. CONCLUSION AND FUTURE WORK

Application of the DS-CDMA signal receiver in programmable radio technology allows a flexible approach to updating and verification of the implemented software receiver. It also allows to implement several methods of reception in order to compare their efficiency, correctness and processing time. Preliminary studies conducted in the lab confirms the versatility of the platform.

The next step will be implementation of RAKE receiver algorithms (multipath reception). In this case, properties of WMSA filters may be required for proper signal detection. The results are going to be compared with simulations presented in this article.

ACKNOWLEDGMENT

New solutions for the software implementation of the RAKE receiver techniques, hereby described; are funded by the Polish Ministry of Science and Higher Education as a part of research and development project No O R00 0049 06. The authors express their sincere thanks for allocated funds for this purpose.

REFERENCES

[1] Mitola III J.; *Software Radio Architecture*, A Wiley & Sons 2000.

[2] Buracchini E.; *The Software Radio Concept*, IEEE Commun. Mag., vol 38, No. 9, Sep. 2000, pp. 138-143.

[3] Chamberlain, M.W.; *A software defined HF radio*, Military Communications Conference, 2005. MILCOM 2005. IEEE, 17-20 Oct. 2005, vol. 4, pp. 2448 – 2453.

[4] Duan L.; The *Key Issues to Design Software Radio*, Radio Science Conference, 2004. Proceedings. 2004 Asia-Pacific, 24-27 Aug. 2004, pp. 119 – 122.

[5] Katulski R., Marczak. A., and Stefański J.; *Technika radia programowalnego*, Przegląd Telekomunikacyjny nr 10/2004r, pp. 402-406.

[6] Sadayuki A., Mamoru S., and Fumiyuki A.; *Performance Comparison between Time-Multiplexed Pilot Chanel and Parallel Pilot Channel for Coherent Rake Combining in DS CDMA Mobile Radio*, IEICE Trans. Commun., vol.E81-B, no.7, July 1998, pp. 1417-1425.

[7] Hidehiro A., Mamoru S., and Fumiyuki A.; *Channel Estimation Filter Using Time-Multiplexed Pilot Channel for Coherent RAKE Combining in DS-CDMA Mobile Radio*, IEICE TRANS. COMMUN., VOL.E81-B, NO.7 July 1998, pp. 1517-1526.

[8] Moher M.L. and Lodge J.H., *TCMP-a modulation and coding strategy for Rician fading channel*, IEEE J. Sel. Areas. Commun., vol. SAC-7, pp. 1347-1355, Dec. 1989,

[9] Cavers J.K., *An analysis of pilot symbol assisted modulation for Rayleigh fading channels*, IEEE Trans. Veh. Technol. vol. VT-40, pp. 689-693, Nov. 1991,

[10] Sampei S. and Sunaga T., *Rayleigh fading compensation for QAM in land mobile radio communication*, IEEE Trans. Tech. Technolo. vol. VT-42, pp. 137-147, May 1993,

[11] Ling F., Coherent detection with reference-symbol based estimation for direct sequence CDMA uplink communications, Proc. VTC'93, pp. 400-403, New Jersey, USA, May 1993

[12] Siwicki W.; Multipath RAKE receiver in WCDMA/FDD systems, Master's thesis, PG WETI, Gdańsk 2002

 ISBN: 978-1-61208-100-7

Resource Allocation of Adaptive Subcarrier Block with Frequency Symbol Spreading for OFDMA

Chang-Jun Ahn†, Tatsuya Omori, and Ken-ya Hashimoto
Graduate School of Engineering, Chiba University
1-33 Yayoi-cho, Inage-ku, Chiba-shi, Chiba, 263-8522 Japan
E-mail: {junny†, k.hashimoto, omori}@faculty.chiba-u.jp

***Abstract*—In a wireless network, the signals transmitted from one sender to different users have independent channel fluctuation characteristics. The diversity that exists between users is called multiuser diversity and can be exploited by the sender to enhance the capacity of wireless network. In multiuser diversity OFDMA system, exploiting channel fluctuation diversity is in essence done by selecting the user with the strong subcarrier channels. The individual subcarrier selection for each user can achieve the best system performance but high signaling overhead and high system complexity are required. On the other hand, the adaptive subcarrier block method achieves worse BER than that of individual subcarrier selection. This is because the selected block contains the poor channel subcarriers. To overcome this problem, in this paper, we propose an adaptive subcarrier block selection with frequency symbol spreading for an OFDMA system.**

***Keywords*-OFDMA; Frequency Symbol Spreading; Minimum Mean Square Error Combining (MMSEC)**

I. INTRODUCTION

Orthogonal frequency division multiplexing (OFDM) is one of the most promising physical layer technologies for high data rate wireless communications due to its robustness to frequency selective fading, high spectral efficiency, and low computational complexity [1]. Moreover, OFDM has been chosen for several broadband WLAN and broadcasting standards like IEEE802.11a, IEEE802.11g, European HIPERLAN/2, terrestrial digital audio broadcasting (DAB) and digital video broadcasting (DVB) [2]-[4].

OFDM allows only one user on the channel at any given time. To accommodate multiple users, a strictly OFDM system must employ time division multiple access (TDMA) or frequency division multiple access (FDMA). Neither of these techniques is time or frequency efficient. Orthogonal frequency division multiplexing access (OFDMA) is a multi-user OFDM that allows multiple access on the same channel. It has been adopted for both uplink and downlink air interfaces of WiMAX fixed and mobile standards, i.e IEEE802.16d and IEEE802.16e respectively [5]-[8]. It has also been adopted by third generation partnership project (3GPP) long-term evolution (LTE) downlink air interface [9].

In a wireless network, the signals transmitted from one sender to different users have independent channel fluctuation characteristics. The diversity that exists between users is called multiuser diversity and can be exploited by the sender to enhance the capacity of wireless network [10]. In multiuser diversity OFDMA systems, exploiting channel fluctuation diversity is in essence done by selecting the user with the strong subcarrier channels [11]. In this case, one subcarrier is selected at once, users do not share the same subcarriers. Therefore, multiuser diversity techniques promise dramatically increased system throughput and spectral efficiency. However, high signalling overhead and high system complexity are required to select the strong subcarrier for each user.

To reduce high signalling overhead and high system complexity, the adaptive subcarrier block (ASB) method has been proposed. Low signalling overhead and low system complexity are merits of an ASB method. It means that an amount of feedback information (FBI) can be reduced. However, an ASB method contains the deep faded subcarriers. Therefore, many errors occur in the deep faded subcarrier. If we can improve the system performance of an ASB method with reduced FBI, the resource allocation is a promising method. In this paper, we propose the resource allocation of an ASB method with frequency symbol spreading for multiuser diversity OFDMA for enhancement of the BER performance. This paper is organized as follows. The system model is described in Section II. In Section III, we describe the proposed multiuser diversity for OFDMA with frequency symbol spreading. In Section IV, we show the simulation results. Finally, the conclusion is given in Section V.

II. SYSTEM MODEL

A. Channel Model

We assume that a propagation channel consists of L discrete paths with different time delays. The impulse response $\hbar_k(\tau, t)$ for user k is represented as [12]

$$\hbar_k(\tau, t) = \sum_{l=0}^{L-1} \hbar_{k,l}(t)\delta(\tau - \tau_{k,l}), \tag{1}$$

where $\hbar_{k,l}$ and $\tau_{k,l}$ are the complex channel gain and time delay of the lth propagation path for user k, and $\sum_{l=0}^{L-1} E|\hbar_{k,l}^2| = 1$, where $E|\cdot|$ denotes the ensemble average operation, respectively. The channel transfer function

 ISBN: 978-1-61208-100-7

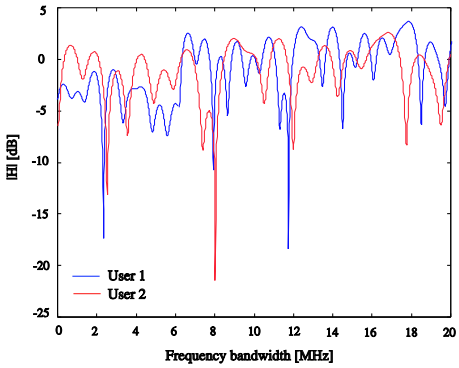


Figure 1. Magnitude of channel transfer function for a radio channel with multiptath.

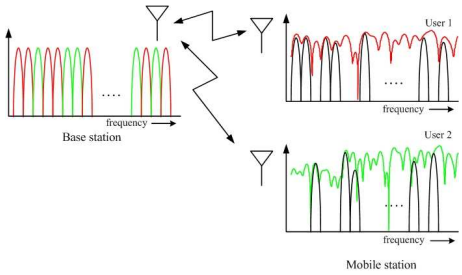


Figure 2. Resource allocation method of maximal sum capacity method for each user.

$h_k(f,t)$ is the Fourier transform of $\hbar_k(\tau,t)$ and is given by

$$h_k(f,t) = \int_0^{\infty} \hbar_k(\tau,t)\exp(-j2\pi f\tau)d\tau = \sum_{l=0}^{L-1} \hbar_{k,l}(t)\exp(-j2\pi f\tau_{k,l}). \quad (2)$$

B. Resource Allocation for Multiuser Diversity OFDMA

Figure 1 shows the magnitude of channel transfer function for different users in a single cell. From Figure 1, each subcarrier fades differently from user to user. Therefore, exploiting channel fluctuation diversity is in essence done by selecting the user with the strong subcarrier. The diversity that exists between users is called multiuser diversity and can be exploited by the transmitter to enhance the capacity of wireless network. Moreover, the combination of multiuser diversity and resource allocation makes great synergy to increase the capacity of wireless network. In multiuser diversity OFDMA systems, the resource allocation criteria for each user is given by

$$Z_{msc}=\sum_{k=0}^{K-1}\sum_{n=0}^{N_c-1}|h_k(n)|^2\alpha_{k,n},\ \alpha_{k,n}=\begin{cases}1 & \text{allocation}\\ 0 & \text{no allocation}\end{cases} \quad (3)$$

where $h_k(n)$ is the channel impulse response for k user and n subcarrier, $\alpha_{k,n}$ is the selection parameter, N_c is the number of subcarriers, and K is the number of users, respectively. This resource allocation method is called the maximal sum capacity (MSC) as shown in Figure 2 [12]. From Eq. (3), the subcarrier is selected for maximization of Z. In this case, one subcarrier is selected at once, users do not share the same subcarriers. Therefore, multiuser diversity techniques promise dramatically increased throughput and spectral efficiency. However, no data rate proportionality among users and high signalling overhead are serious matters [12]. To improve the data rate fairness, the maximal weighted sum capacity (MWSC) method has been proposed. The resource allocation criteria of the MWSC method is given by

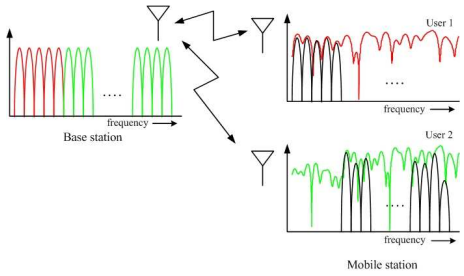


Figure 3. Resource allocation method with adaptive subcarrier block for each user.

$$Z_{mwsc}=\sum_{k=0}^{K-1}\sum_{n=0}^{N_c-1}\omega_{k,n}|h_k(n)|^2\alpha_{k,n}, \quad (4)$$

where $\omega_{k,n}$ is the weight. However, the MWSC method has no guarantee for meeting proportional user data rates, high signalling overhead and high system complexity are still remained [13]. In an OFDM, the channel response at a particular subcarrier frequency is not supposed to be totally different from its neighboring frequencies, and hence, they must have correlation which depends on the coherence bandwidth of the channel B_c. By using the concept of coherence bandwidth, resource allocation method with the adaptive subcarrier block (ASB) has been proposed for user data rate fairness with reduced complexity as shown in Figure 3 [14]. The resource allocation criteria of an ASB method is given by

$$Z_{asb}=\sum_{k=0}^{K-1}\sum_{\varsigma=0}^{N_c/N_{SF}-1}\Big\{\sum_{q=0}^{N_{SF}-1}|h_k(\varsigma\cdot N_{SF}+q)|^2\Big\}\cdot\alpha_{k,\varsigma}. \quad (5)$$

An ASB method achieves worse BER than that of a MWSC method. This is because an ASB method contains the deep faded subcarriers. Therefore, many errors occur in the deep faded subcarrier. However, low signalling overhead and low system complexity are merits of an ASB method. It means that an amount of FBI can be reduced. Although, the performance degradation of an ASB method is still considerable problem. If we can improve the system performance of an

ASB method with reduced FBI, the resource allocation is a promising method. In this paper, we propose the resource allocation of an ASB method with frequency symbol spreading for multiuser diversity OFDMA for enhancement of the BER performance.

III. Proposed System

A. Transmitter Structure

The transmitter block diagram of OFDMA with an ASM method using frequency symbol spreading is shown in Figure 4(a). The binary data sequence for user k is modulated, and N_p pilot symbols are appended at the beginning of the sequence. The symbol sequences consist of serial to parallel (S/P) converted to N_{SF} parallel sequences. The N_{SF} parallel sequences are assigned on the adequate frequency symbol spreading block by using Eq.(5), as illustrated in Figures 4(a) and 5(a), where N_{SF} is the spreading code length. Each parallel sequence(subcarrier bock) feds into the frequency symbol spreading block is copied by the same length of spreading code N_{SF}, as shown in Figure 5(a). These copied parallel sequences are spread by the spreading code with N_{SF} and combined. The transmitting signal waveform is obtained by applying an inverse Fourier transform (IFFT). The proposed transmitting signal for user k can be expressed in its equivalent baseband representation as [12]

$$s_k(t) = \sum_{i=0}^{N_p+N_d-1} g(t-iT) \cdot \Big\{ \sqrt{\frac{2S}{N_c}} \sum_{n=0}^{N_c-1} u_k(n,i) \cdot \exp\left[j2\pi(t-iT)n/T_s\right] \Big\}, \quad (6)$$

where N_d, N_p are the number of data and pilot symbols, T_s is the effective symbol length, S is the average transmitting power, T is the OFDM symbol length, and $u_k(n,i)$ is the nth subcarrier of the ith OFDM symbol for user k. The power spectra of the input and output signals of the frequency symbol spreading block are shown in Figure 5(b). As indicated by Figure 4, the parallel converted data $d_k(n,i)$ is fed into the ξ_kth frequency symbol spreading block. The input data $d_k(n,i)$ is copied N_{SF} times and multiplied as shown in Figure 5. In the same frequency symbol spreading block, the output spread signals are combined. Therefore, all data are superimposed over the frequency domain. As shown in Figure 5(b), the energy of each input data is divided over N_{SF} subcarriers by means of a spreading subcode, and each subcarrier conveys N_{SF} divided data. The frequency separation between adjacent orthogonal subcarriers is $1/T_s$ and can be expressed, by using the nth subcarrier of the ith modulated symbol $d_k(n,i)$ with $|d_k(n,i)| = 1$, as

$$u_k(n,i) = \begin{cases} \sum_{q=0}^{N_{SF}-1} c_q(n \bmod N_{SF}) \cdot d_k(\xi_k N_{SF}+q, i) & \text{for } \xi_k \le \lfloor \frac{n}{N_{SF}} \rfloor < \xi_k+1 \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

where $\xi_k \in \{0, 1, \cdots, N_c/N_{SF}-1\}$ is the assigned subcarrier block for user k, "mod" denotes modulus, and $\lfloor x \rfloor$ denotes the largest integer less than or equal to x. From Eq. (7), the subcarrier block is adaptively assigned for each user. The orthogonal spreading sequences $c_q(m)$ satisfy

$$\sum_{m=0}^{N_{SF}-1} c_q(m) c_w^*(m) = \begin{cases} N_{SF} & \text{for } q = w \\ 0 & \text{for } q \neq w \end{cases} \quad (8)$$

where $|c_q(m)| = 1$, where $*$ denotes the complex conjugate. In Eq. (6), $g(t)$ is the transmission pulse given by

$$g(t) = \begin{cases} 1 & -T_g \le t \le T_s \\ 0 & \text{otherwise} \end{cases}. \quad (9)$$

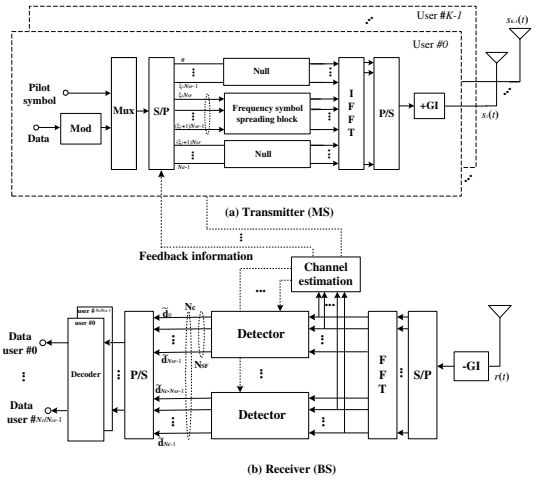


Figure 4. Proposed multiuser diversity OFDMA with frequency symbol spreading and adaptive subcarrier block selection.

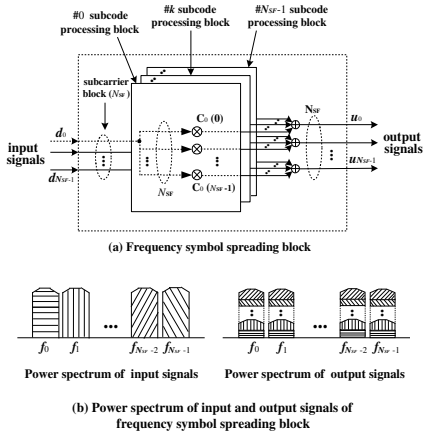


Figure 5. The structure of frequency symbol spreading block and power spectrum of input and output signals of frequency symbol spreading block.

B. Receiver Structure

The receiver structure is illustrated in Figure 4(b). By applying the FFT operation, the received signal $r(t)$ is resolved into N_c subcarriers. First, the received signal is frequency equalized to reduce the frequency distortion due

to the frequency-selective fading. The received signal $r(t)$ in the equivalent baseband representation can be expressed as

$$r(t) = \sum_{k=0}^{K-1} \int_{-\infty}^{\infty} h(\tau, t) s_k(t-\tau) d\tau + n(t), \qquad (10)$$

where $n(t)$ is additive white Gaussian noise (AWGN) with a single sided power spectral density of N_0. The nth subcarrier of the ith received signal $\tilde{r}(n,i)$ is given by

$$\begin{aligned} \tilde{r}(n,i) &= \frac{1}{T_s} \int_{iT}^{iT+T_s} r(t) \exp\left[-j2\pi(t-iT)n/T_s\right] dt \\ &= \sqrt{\frac{2S}{N_c}} \sum_{k=0}^{K-1} \sum_{e=0}^{N_c-1} u_k(e,i) \cdot \frac{1}{T_s} \int_0^{T_s} \exp[j2\pi \\ &\quad \cdot (e-n)t/T_s] \cdot \Big\{ \int_{-\infty}^{\infty} h(\tau, t+iT) g(t-\tau) \\ &\quad \cdot \exp(-2\pi e\tau/T_s) d\tau \Big\} dt + \hat{n}(n,i) \qquad (11) \end{aligned}$$

where $\hat{n}(n,i)$ is AWGN noise with zero-mean and a variance of $2N_0/T_s$. After abbreviating, Eq. (11) can be rewritten as

$$\begin{aligned} \tilde{r}(n,i) &\approx \frac{1}{T_s} \sqrt{\frac{2S}{N_c}} \sum_{k=0}^{K-1} \sum_{e=0}^{N_c-1} u_k(e,i) \\ &\quad \cdot \int_0^{T_s} \exp[j2\pi(e-n)t/T_s] + \hat{n}(n,i) \\ &= \sqrt{\frac{2S}{N_c}} \sum_{k=0}^{K-1} H(n/T_s, iT) u_k(n,i) + \hat{n}(n,i). \qquad (12) \end{aligned}$$

Observing Eq. (12), we can see that the received signals have frequency distortion arising from the frequency-selective fading. To reduce this frequency distortion, frequency equalization combining is necessary. Here, we explain a channel estimation scheme using N_p pilot symbols. The channel response of the nth subcarrier is given by

$$\tilde{H}(n/T_s) = \frac{1}{N_p \sqrt{2P/N_c}} \sum_{i=0}^{N_p-1} \tilde{r}(n,i) \cdot p^*(n,i), \qquad (13)$$

where $\{p(n,i),\ 0 \le i \le N_p - 1\}$ and P are the transmitted pilot symbol and the power, respectively. In frequency-selective fading, the orthogonality among spreading codes may be destroyed. To compensate for the possible broken orthogonality among the spreading codes, frequency equalization schemes such as orthogonal restoration combining (ORC) and minimum mean square error combining (MMSEC) are used for detection.

1) ORC: ORC uses a combining weight that is inversely proportional to the channel transfer function $H(n/T_s)$, in order to perfectly restore the orthogonality. The ORC weight $\omega_{orc}(n,i)$ is given by

$$\omega_{orc}(n,i) = \frac{1}{\tilde{H}(n/T_s)} \qquad (14)$$

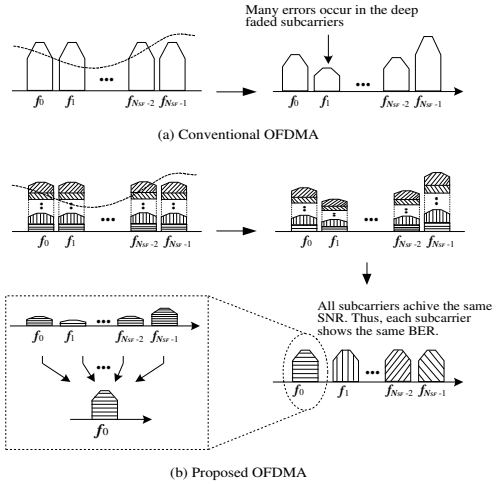


Figure 6. Power spectrum of the proposed OFDMA system.

and the weight $\{\hat{u}_k(n,i),\ n = 0, 1, \cdots, N_c - 1\}$ of the nth subcarrier becomes

$$\begin{aligned} \hat{u}_{k,orc}(n,i) &= \omega_{orc}(n,i)\tilde{r}(n,i) \\ &= \sqrt{\frac{2S}{N_c}} \eta(n,i) u_k(n,i) + \frac{\hat{n}(n,i)}{\tilde{H}(n/T_s)}, \\ &\quad \text{for } \xi_k \le \lfloor \tfrac{n}{N_{SF}} \rfloor < \xi_k + 1 \qquad (15) \end{aligned}$$

where

$$\eta(n,i) = \frac{H(n/T_s, iT)}{\tilde{H}(n/T_s)}. \qquad (16)$$

The decision variable $\tilde{d}_{k,orc}(n,i)$ of the nth subcarrier of ith data symbol for user k as

$$\begin{aligned} \tilde{d}_{k,orc}(n,i) &= \sum_{q=0}^{N_{SF}-1} \hat{u}_{k,orc}(\xi_k N_{SF} + q, i) c^*_{n \bmod N_{SF}}(q) \\ &= \sum_{q=0}^{N_{SF}-1} \Big(\sqrt{\frac{2S}{N_c}} \eta(\xi_k N_{SF} + q, i) \\ &\quad \cdot u_k(\xi_k N_{SF} + q, i) + \frac{\hat{n}(\xi_k N_{SF} + q, i)}{\tilde{H}((\xi_k N_{SF} + q)/T_s)} \Big) \\ &\quad \cdot c^*_{n \bmod N_{SF}}(q) \\ &= \sqrt{\frac{2S}{N_c}} \sum_{q=0}^{N_{SF}-1} \eta(\xi_k N_{SF} + q, i) \\ &\quad \cdot d_k(\xi_k N_{SF} + q, i) \\ &\quad + \sqrt{\frac{2S}{N_c}} \sum_{q=0}^{N_{SF}-1} \eta(\xi_k N_{SF} + q, i) \\ &\quad \cdot d_{intr}(\xi_k N_{SF} + q, i) c_w(q) c^*_{n \bmod N_{SF}}(q) \\ &\quad + \sum_{q=0}^{N_{SF}-1} \frac{\hat{n}(\xi_k N_{SF} + q, i) c^*_{n \bmod N_{SF}}(q)}{\tilde{H}(\xi_k N_{SF} + q/T_s)} \\ &\quad \text{for } \xi_k \le \lfloor \tfrac{n}{N_{SF}} \rfloor < \xi_k + 1 \ , \qquad (17) \end{aligned}$$

where d_{intr} is the interference term. From Eq. (17), we can observe that the first term is the desired signal, the second term is the interference, and the third term is a noise term. From the third term, ORC scheme can restore the orthogonality, but it enhances the noise term for a deep faded subcarrier.

2) MMSEC: MMSEC combining weight $\omega_{msc}(n,i)$ is given by

$$\omega_{msc}(n,i) = \frac{\sqrt{\frac{2S}{N_c}} \cdot \tilde{H}(n/T_s)}{\left|\sqrt{\frac{2S}{N_c}} \cdot \tilde{H}(n/T_s)\right|^2 + 2\tilde{\sigma}^2} \qquad (18)$$

where $\tilde{\sigma}^2$ is the estimated noise power per subcarrier, which is assumed identical for all subcarriers in this paper. The received modulated data symbol for user k can be written as

$$\tilde{d}_{k,msc}(n,i) = \begin{cases} \sum_{q=0}^{N_{SF}-1} \hat{u}_{k,msc}(\xi_k N_{SF}+q,i) \\ \cdot c^*_{n \bmod N_{SF}}(q) \quad \text{for } \xi_k \le \lfloor \frac{n}{N_{SF}} \rfloor < \xi_k+1 \\ 0 \qquad\qquad\qquad \text{otherwise} \end{cases} \qquad (19)$$

where $\{\hat{u}_{k,msc}(\xi_k N_{SF}+q,i), \quad q=0,1,\cdots,N_{SF}-1\}$ is the weighted component of the user k and is given by

$$\hat{u}_{k,msc}(n,i) = \begin{cases} \sqrt{\frac{2S}{N_c}} u_k(n,i) \\ \cdot w_{msc}(n,i) \\ +\hat{n}(n,i) w_{msc}(n,i) \\ \qquad\qquad \text{for } \xi_k \le \lfloor \frac{n}{N_{SF}} \rfloor < \xi_k+1 \\ 0 \qquad\qquad \text{otherwise} \end{cases} \qquad (20)$$

Eq. (20) is made by Eqs. (12) and (18). In OFDMA with an ASB method using frequency symbol spreading, each S/P transformed signal is spread by an orthogonal spreading code with length N_{SF} over N_{SF} subcarriers and combined. This means that each subcarrier holds several superimposed S/P transformed signals with the same power rate. In this case, frequency-selective faded subcarriers are obtained with the same power rate for each S/P-transformed signal. Therefore, the detected signals also have the same SINR. As a result, each subcarrier shows the same BER performance under the frequency selective fading. From this reason, the proposed method can improve the system performance with maintaining low signalling overhead and low system complexity. Moreover, each operation is carried out symbol by symbol. Therefore, the latency is necessary only one symbol time. Since the current systems such as LTE and WiMAX, use the subcarrier block method for each user, it is easy to embed the proposed method on the current systems.

IV. Computer Simulated Results

Figure 4 shows a simulation model of the proposed multiuser diversity OFDMA with $N_c = 64$ subcarriers and its bandwidth 20MHz. On the transmitter side, the bits are QPSK modulated and then serial-to-parallel transformed. The OFDM time signals are generated by an IFFT and transmitted over the frequency selective and time variant radio channel after cyclic extensions have been inserted. The transmitted signals are subject to broadband channel propagation. In this model, $L = 15$ path Rayleigh fadings have exponential shapes with path separation $T_{path} = 140nsec$ and the RMS delay spread is $\tau_{rms} = 0.65\mu s$. This situation causes severe frequency selective fading. The maximum Doppler frequency is assumed to be 10Hz. The packets consist of 64 subcarriers and 22 OFDM symbols (number of pilot signals: $N_p = 2$, number of data signals: $N_d = 20$). Table 1 shows the simulation parameters.

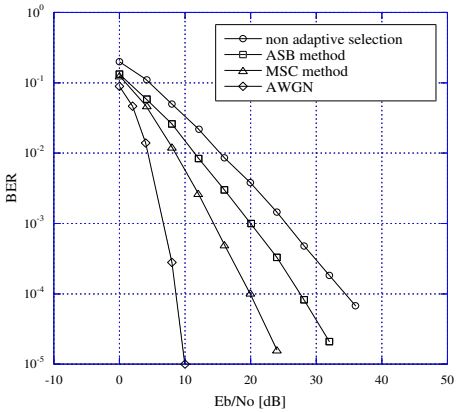


Figure 7. BER of non adaptive selection, the ASB method and the MSC method for multiuser diversity OFDMA system with 4 users and N_{SF}=16 for Doppler frequency of 10Hz.

Figure 7 shows the BER of non adaptive selection, the ASB method and the MSC method for multiuser diversity OFDMA system with 4 users and N_{SF}=16 for Doppler frequency of 10Hz. From the simulation results, the MSC method achieves the best BER performance compared with non adaptive selection and the ASB method. This is because the MSC method can select subcarriers with highest SNR for each user. On the other hand, the ASB method shows worse BER than that of the MSC method. This is because the selected block includes the poor subcarriers.

Figure 8 shows the BER of non adaptive selection, the

Table I
Simulation parameters.

Modulation	QPSK
Demodulation	Coherent detection
Effective data rate	20 Msymbol/s
Number of users	4
Number of carriers	$N_c = 64$
Bandwidth	20MHz
Guard interval	16 sample times
Frame size	22 symbols ($N_p = 2, N_d = 20$)
Fading	15 path Rayleigh fading
Doppler frequency	10 Hz
Subcarrier block size	$N_{SF} = 16$

 ISBN: 978-1-61208-100-7

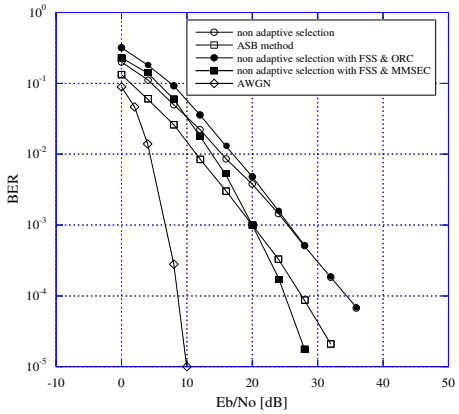


Figure 8. BER of non adaptive selection, the ASB method, non adaptive selection with frequency symbol spreading (FSS) and ORC and the ASB method with frequency symbol spreading and MMSEC with 4 users and N_{SF}=16 for Doppler frequency of 10Hz.

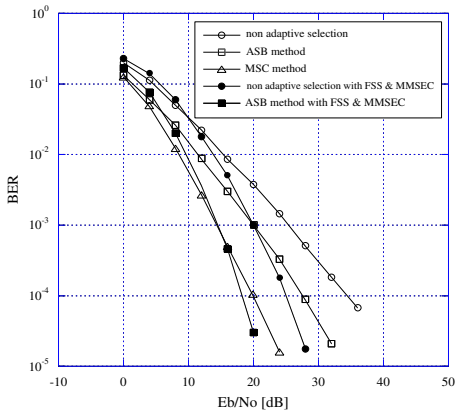


Figure 9. BER of non adaptive selection, the ASB method, the MSC method, non adaptive selection and the ASB method using frequency symbol spreading and MMSEC with 4 users and N_{SF}=16 for Doppler frequency of 10Hz.

ASB method, non adaptive selection with frequency symbol spreading and ORC and the ASB method with frequency symbol spreading and MMSEC with 4 users and N_{SF}=16 for Doppler frequency of 10Hz. Non adaptive selection with frequency symbol spreading and ORC shows worse BER than that of non adaptive selection in low E_b/N_o. This is because the ORC based non adaptive selection with frequency symbol spreading enhances the noise although no error floor is seen. The MMSEC provides the best BER performance, since the MMSEC minimize the power loss while suppressing the noise enhancement.

Figure 9 shows the BER of non adaptive selection, the ASB method, the MSC method, non adaptive selection and the ASB method using frequency symbol spreading and MMSEC with 4 users and N_{SF}=16 for Doppler frequency of 10Hz. From the simulation results, the proposed ASB method with frequency symbol spreading and MMSEC provides the best BER performance in high E_b/N_o, since the ASB method obtains the highest SNR subcarrier block and the frequency symbol spreading and MMSEC minimize the noise enhancement.

V. Conclusion

In this paper, we have proposed the ASB method with frequency symbol spreading for OFDMA system. From the simulation results, the proposed ASB method with frequency symbol spreading and MMSEC achieves 1dB gain compared with the conventional MSC method at BER of 10^{-4}. Moreover, the proposed method can reduce the quantities of FBI compared with the MSC method.

References

[1] J. A. C. Bingham, “Multicarrier modulation for data transmission: an idea whose time has come,” IEEE Comm. Mag., vol. 28, pp. 5-14, May 1990.

[2] “IEEE draft standard for local and metropolitan area network-part 16: Air interface for fixed broadband wireless access systems - medium access control modifications and additional physical layer specifications for 2-11GHz,” IEEE LAN MAN Standards Committee, 2002.

[3] ETSI ETS 301 958, “Digital Video Broadcasting (DVB); interaction channel for digital terrestrial television (RCT) incorporating multiple access OFDM,” ETSI, Tech. Rep., March 2002.

[4] W. Fischer, “Digital Video and Audio Broadcasting Technology,” Springer-Verlag, Berlin Heidelberg, 2008.

[5] I. Koffman and V. Roman, “Broadband wireless access solutions based on OFDM access in IEEE802.16,” IEEE Commun. Mag., vol. 40, pp. 96-103, April 2002.

[6] H. Yin and S. Alamouti, ”OFDMA: A Broadband Wireless Access Technology,” Proc. of IEEE Sarnoff Symposium, pp. 1-4, March 2006.

[7] IEEE standard for local and metropolitan area networks, “Part 16: Air interface for fixed broadband wireless access systems,” 1 October 2004.

[8] IEEE standard for local and metropolitan area networks, “Part 16: Air interface for fixed and mobile broadband wireless access systems,” 28 February 2006.

[9] 3GPP TSG-RAN, “3GPP TR 25.814, Physical Layer Aspects for Evolved UTRA (Release 7),” v1.3.1, May 2006.

[10] H. Koubaa, V. Hassel, and G. Ien, ”Contention-less feedback for multiuser diversity scheduling,” Proc. of IEEE VTC2005-Fall, vol.3, pp. 1574-1578, September 2005.

[11] T. Ban, W. Choi, B. Jung, and D. Sung, ”Multi-user diversity in a spectrum sharing system,” IEEE Transactions on Wireless Communications, Vol.8, no.1, pp. 102-106, January 2009.

[12] Chang-Jun Ahn, “Achievable Throughput Enhancement Based on Modified Carrier Interferometry for MIMO/OFDM,” Elsevire Digital Signal Processing, vol.20, no.5, pp. 1447-1457, September 2010.

 ISBN: 978-1-61208-100-7

IMS Signalling in LTE-based Femtocell Network

Melvi Ulvan, Robert Bestak
Department of Telecommunication Engineering
Czech Technical University in Prague
Prague, Czech Republic
[ulvanmel, robert.bestak]@fel.cvut.cz

Ardian Ulvan
Department of Electrical Engineering
University of Lampung
Bandar Lampung, Indonesia
ulvan@unila.ac.id

***Abstract*—The IP Multimedia Subsystem functionality is designed to work on various wireless access technologies and in all network coverage such as macrocell, microcell and femtocell. This paper describes the study of IP Multimedia Subsystem signalling in femtocell network. The 3GPP Long Term Evolution-based femtocell integrated with IP Multimedia Subsystem core network is considered as the system architecture. Session establishment signalling procedure is taken into account to analyse the IP Multimedia Subsystem signalling call flows in femtocell network. Signalling performance is analyzed by mean of Session Initiation Delay properties.**

Keywords - Femtocell, IMS Signalling, 3GPP LTE, SIP.

I. INTRODUCTION

The Internet protocol Multimedia Subsystem (IMS) is a framework that specified for the third generation (3G) mobile networks, to provide Internet Protocol (IP) telecommunication services. The IMS standard was specified by the Third Generation Partnership Project (3GPP) in the specification Release 5 [1] and Release 6 [2]. The standard supports multiple access types, e.g., Global System for Mobile (GSM), Wideband Code Division Multiple Access (WCDMA), CDMA2000, or IEEE802.11 Wireless Fidelity (WiFi) [3].

The present active standard describes the IMS functionalities for classical wireless network coverage such as macrocell and microcell. As a new emerging wireless network which has a very small coverage, the standardization for femtocell is still under development.

In a femtocell network, the Femto Access Point (FAP) is defined as a small cellular base station designed to be used in the house, residential or in the office. The FAP allows service providers to improve indoor coverage mainly where access is limited or even unavailable. In the 3GPP terminology, the FAP is called as Home Node B (HNB) [4]. Though the IMS can accommodate current and future services in wired and wireless networks, however, the IMS-femtocell interoperability is still a challenging issue.

In the IMS network, the communication between the IMS core network (IMS-CN) and its clients is performed on the basis of request/response signalling messages. At the first step, client sends a Session Initiation Protocol (SIP) message specific for the given request when requiring a service from the IMS-CN. The IMS-CN responds by sending the particular SIP message with regard to the received request.

The SIP is used as a signalling protocol in the IMS environment. It is defined in RFC 3261 [5] which have functionalities on registration, session establishment, session management and participant invocation (including creating, modifying, and terminating sessions with one or more participants). SIP signalling is the primary method for user registration and session control in the IMS architecture [6]. The Call Session Control Function (CSCF) is the core signalling server in the IMS networks architecture. It acts as both a SIP Registrar and a SIP proxy server [7].

Many research and scientific works consider either femtocell or IMS, but none of them concern on the interworking between IMS and femtocell. Accordingly, this paper is concern with this issue. The main focus is to investigate the effective and efficient IMS' SIP signalling mechanism when it works on femtocell environment. The all-IP connection integration between IMS-CN and FAP is described. In addition, another critical issue such as the delay properties of session establishment signalling of source and correspondent nodes is also taken into account. The IMS clients in femtocell environment are based on the 3GPP LTE.

The paper is organized as follow. Section II summarizes the related work regarding IMS and femtocell environment. Section III describes the integration architecture between IMS and FAP. Some literatures and technical aspects in IMS are portrayed, as well as the LTE-based femtocell. Section IV explains the proposed IMS signalling mechanism in femtocell network. The new messages flow is also depicted. End-to-end delay in the proposed messages flow is studied. Finally, Section V concludes our paper.

II. RELATED WORK

There are many published papers exposing the femtocells in terms of increasing the network capacity, saving energy, supporting high-speed data rate, and providing benefits from the social and economic side. Authors in reference [8] simulated the deployment of femtocells in residential scenario to study their effects on the service experienced by users connected to a macrocell. They found out no significant impact on the dropped call rate when auto-configuration is deployed in the femtocells. In [9], Chandrasekhar and Andrews addressed the reduced cost by deploying macrocell and femtocell users in a shared region of spectrum. They proposed a link quality protection algorithm to progressively reduce the target Signal-to-

 ISBN: 978-1-61208-100-7

Interference Noise Ratio (SINR). This two-tier distributed power control algorithm ensured minimal network overhead on femtocells. In addition to energy saving and coverage issues, Claussen *et al.* [10] proposed the mobility event based self-optimization and coverage adaptation method for femtocell deployment. As the result, the total number of mobility events caused by femtocells deployment is significantly reduced. Moreover, the femtocell's indoor coverage is improved as well. Other technical and business advantages of femtocell deployment as well as the technology state-of-the-art and its challenges have been overviewed and described in [11]. Generally, those papers discuss benefits and technical issues of femtocell deployment, however very few papers address the integration of femtocells into IMS architecture.

In addition to IMS research and technological development, several works have been addressed, mostly in terms of system performances of session establishment procedure. The works in [12] and [13] provided the SIP-based IMS signalling delay for IMS session establishment procedure. In [12], A. Munir analyzed the end-to-end delay where both source node and correspondent node are the combination of 3G/UMTS and WiMAX networks. The signalling delay is analyzed separately as transmission, processing and queue delays. However, the paper only shows the delay as a whole. It was not described which delay part that contributed the most significant delay.

More comprehensive analysis of session establishment procedure was carried out in [13] where the delay properties in each delay entities were investigated. The structure of session establishment signalling that is based on the standard was presented. The authors also examined which delay among the transmission, processing and queuing delays that contributes the most significant delay in the system.

The optimization of SIP session setup delay for voice over IP (VoIP) service in 3G wireless networks is studied in [14]. The authors evaluated SIP session setup performances with various underlying protocols transport control protocol (TCP), user datagram protocol (UDP), radio link protocols (RLPs)) as a function of the frame error rate (FER). The adaptive retransmission timer is proposed to optimizing the delay. In addition [15], the analysis of SIP-based mobility management in 4G wireless network was carried out. Despite they were not concern on session establishment procedure, but some delay issues, particularly the delay on radio link protocol (RLP) and non-RLP have been carried out.

The optimization of efficient route for femtocell-based all IP networks is carried out in [16]. The Session Initiation Protocol (SIP) signalling routes and packet route of FAP resided in IP domain was compared with the FAP connected to IMS-CN through a Radio Access Network. For this purpose, the authors created the testbed for each FAP scenario. The end-to-end system delays were measured and reported.

III. THE OVERVIEW OF IMS – FEMTOCELL INTEGRATION

A. *The Operational Concepts of IMS*

The IMS is not about services, but the concept providing access to all services regardless of the media type. It uses a common control architecture that works well for all media. Functionalities that are required for supporting multimedia real time services, such as service creation, registration, invocation and execution are incorporated in the service architecture of IMS. Among these entities, IMS contains multiple SIP proxies called Call Session Control Functions (CSCFs) with following roles:

- Proxy-CSCF (P-CSCF) which is the first contact point in IMS and interacts with GGSN (Gateway GPRS Support Node), i.e., the access point from UMTS to external networks, for policy control and resources allocation.
- Interrogating-CSCF (I-CSCF) which acts as a SIP Registrar and is responsible for routing sessions to appropriate Serving-CSCF (S-CSCF).
- S-CSCF which performs session control and service trigger.

The IMS provides an efficient service provisioning capability compare to circuit-switched (CS) and packet-switched (PS) networks. When a client registers into IMS network, a Subscriber Service Profile (SSP) is downloaded by the S-CSCF from the HSS.

The SSP contains service-related information and identifies the services that need to be provisioned. If multiple services need to be implemented, it determines the order in which they are provisioned. The SSP also includes the address of the servers that must execute the subscriber's request. This approach allows IMS to serve as a re-usable service infrastructure by letting providers control and manage the complexities involved in service filtering, triggering, and interaction.

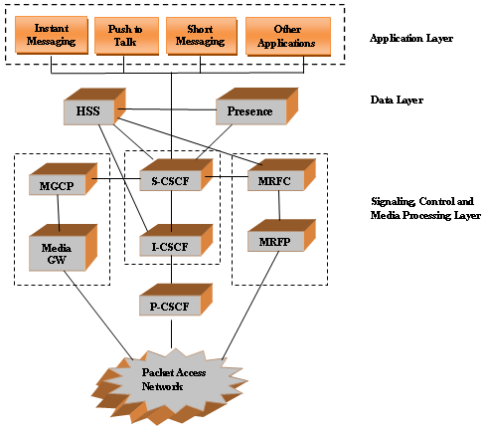


Figure 1. Layer architecture of IMS

Both PS and CS networks can be integrated into a single session by the network operator, accordingly, users can add multimedia services to existing services in real time. Fig. 1 represents how IMS logical function interacts to support a few sample applications.

B. *Femtocell Network*

1) *LTE-based Femtocell*

Based on 3GPP specification [17] the LTE-based FAP (HNB) provides the Radio Access Network (RAN) connectivity using Iuh interface, support the macrocell NodeB and most of the Radio Network Controller (RNC) functions, FAP authentication, Femto-Gateway (F-GW), FAP registration and Femto User Equipment (F-UE) registration over Iuh. Furthermore, the F-GW serves the purpose of a RNC and present to the mobile CN as a concentrator of FAP connections that provides concentration function for the control plane and user plane. Logical architecture of LTE-based FAP can be seen in Fig. 2.

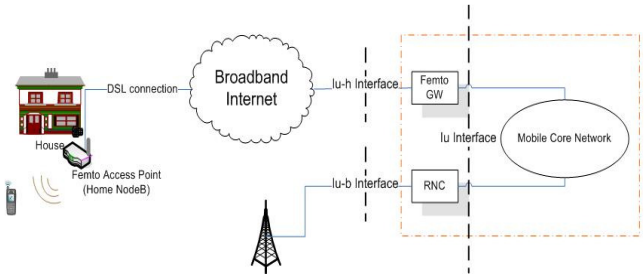


Figure 2. Logical architecture of LTE-based femtocell

2) *LTE Advanced-based Femtocell*

Additionally, in the basis of LTE-A [17], the evolved RAN (or Evolved UMTS Terrestrial Radio Access Network/E-UTRAN) is the key element since it provides all system functionalities included the Physical (PHY), Medium Access Control (MAC), Radio Link Control (RLC), and Packet Data Control Protocol (PDCP) [18]. It consists a single node, i.e., evolved Node B (eNodeB) or Home evolved Node B (HeNB)/FAP. It also provides radio resource control (RRC) functionality that corresponds to handover procedure.

E-UTRAN interacts with the Evolve Packet Core (EPC) system that consist the Mobility Management Entity (MME), Serving Gateway (SGW) and F-GW. The interaction between all functional elements of EUTRAN and EPC is depicted in Fig. 3.

The FAP supports the same functions as those supported by an eNodeB. The procedures between FAP and EPC run as same as those between macrocell eNodeB and the EPC. The F-GW serves as a concentrator for the control-plane, specifically the S1-MME interface. The F-GW may optionally terminate the user-plane towards the FAP and towards the SGW. Moreover, the F-GW provides a relay

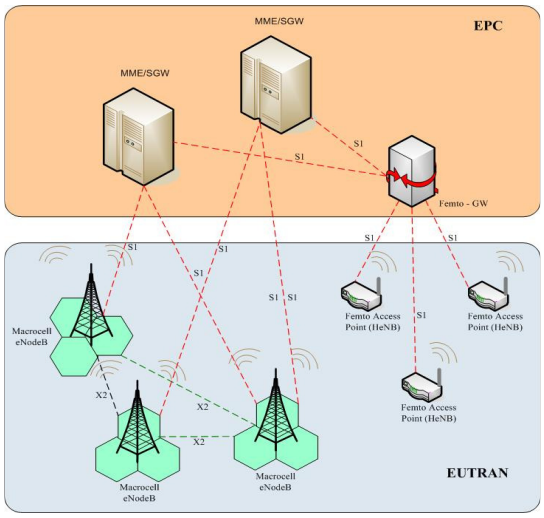


Figure 3. Logical architecture of LTE-A based femtocell

function for relaying user plane data between the HeNB and the SGW.

In addition, 3GPP also specified two standard interfaces, i.e., X2 and S1 interface, for the Evolved Packet System (EPS). The X2 interface provides capability to support radio interface mobility and shall support the exchange of signalling information between eNodeB macrocells. Therefore, for handover between eNodeB macrocells, the procedure is performed without EPC involvement. Preparation and exchange of signalling flows in the handover procedure are directly between eNodeB using X2 interface. On the other hand, the S1 interface supports many-to-many relations between EPC's elements (MME/SGW) and eNodeB. Moreover S1 is also used for the communication between FAP/HeNodeB with the MME/SGW through the F-GW. Specifically, the connection to MME is using S1 control plane (S1-C) interface and the connection to SGW is using S1 user plane (S1-U) interface.

C. *SIP/IMS-based Integration*

Introducing the femtocells obviously extends the wireless access directly into the homes. However, how to integrate thousands of FAPs with the existing mobile infrastructure is still challenged. There are several different ways of achieving this integration. In December 2007, Femto Forum proposed thirteen different network architectures [19] for the integration of femtocell with existing network infrastructures. Those proposals can be classified into three categories, i.e., cellular-based integration, SIP/IMS-based integration, and Unlicensed Mobile Access (UMA)-based integration. In this paper, the SIP/IMS-based integration has been considered to analyse the IMS signalling. The architecture of integrated system is depicted in Fig. 4.

In this architecture, the FAPs integrated directly with the IMS core through all-IP connectivity. Alternative architecture under this category includes softswitch-based implementations where the FAPs are integrated to a softswitch via SIP interface.

 ISBN: 978-1-61208-100-7

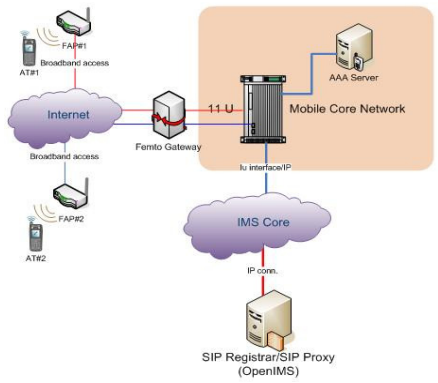


Figure 4. IMS-Femtocell integrated system architecture

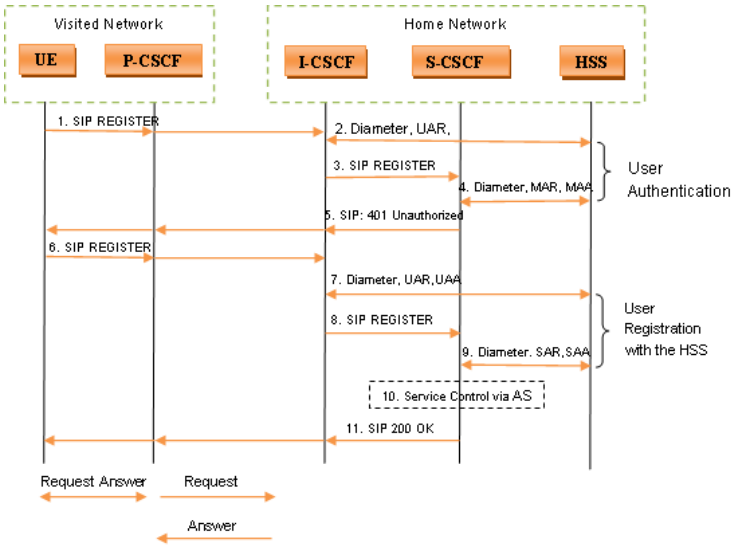


Figure 5. Signalling messages flow in registration procedure

This architecture leverages a SIP based VoIP network for cost-effective delivery since it is interworking with a cellular core to extend legacy circuit switched services. The signals (2G/3G) from legacy mobile stations (MSs) are converted into appropriate SIP messages over IP by FAP. Afterwards, they are interfaced to a SIP-MSC inter-working function (IWF) which connects to IMS-CN through internet.

Integrating femtocells directly with the IMS core offers definite advantages [20], i.e., there is no scalability issue since the mobile core network is completely offloaded, and the requirements for upgrading the mobile core network infrastucture can be avoided as the FAP has the functionality of Radio Access Network (RAN), it can scale the traffic. Traffic latency challenges are mitigated as the number of hops is minimal. In addition, the use of all-IP in the core network significantly reduces the operational expenditures for carriers.

Nevertheless, the SIP/IMS integration seems has some limitations, i.e., the lack of standards-based support for security, mobility, and supplementary services. The security procedures need to be concerned, particularly in the authorization, authentication and accounting (AAA) phases during the network entry process. Two types of authorization may be applied, i.e., the conventional network-based authorization and a novel scenario that allow the subscriber to perform the authorization procedure to another subscriber (as on the Bluetooth pairing procedure).

On the mobility side, several requirements have to be met in order to achieve a seamless FAP-to-FAP and FAP-to-macrocell handovers. At last, this architecture should also be interworked with the existing circuit switch network to allow the continuity implementation of existing supplementary services such as call barring, call waiting, call forwarding, voicemail, SMS, etc.

IV. IMS SIGNALLING IN FEMTOCELL NETWORK

A. *IMS Signalling Procedure*

In IMS system everything is controlled using one common signalling method. The signalling allows network entities to communicate with one another as well as application servers (ASs).

As the comparison, in the legacy Integrated Service Digital Network (ISDN), Signalling System #7 (SS7) is the signalling method used to connect facilities in the telephone network. SS7 connects different operators' networks with one another, and connects all of the switches within the network with one another. In IMS, SS7 is replaced with a new signalling method: SIP, which is used to control everything in the network.

SIP signalling is the primary method for user registration and session control in the IMS architecture. The CSCF is the core signalling server in the IMS networks. It acts as both a SIP Registrar and SIP proxy server. Fig. 5 depicts the signalling procedures for registration in the IMS Core.

The procedure starts with the user's SIP REGISTER request being sent to the visited P-CSCF. Due to air interface bandwidth limitation, messages are compressed before being sent out by the user and are decompressed at the P-CSCF. If multiple S-CSCFs exist in the user's home network, an I-CSCF needs to be deployed for selecting an S-CSCF for serving the user session. In this case, the P-CSCF resolves the address of the user's home I-CSCF using the user's home domain name and forwards the REGISTER to the I-CSCF. After the I-CSCF sends a User-Authorization-Request (UAR) to the HSS, which returns available S-CSCF addresses, the I-CSCF selects one S-CSCF and forwards the REGISTER message.

B. *Session Establishment Signalling in Femtocell*

In IMS, there are three forms of messaging, i.e., immediate messaging, session-based messaging and deferred delivery messaging. Session establishment is part of session-based messaging. The procedure involves an end-to-end signalling message exchange.

There are several signalling messages that take part in the session establishment procedure as shown in Table 1.

 ISBN: 978-1-61208-100-7

Originating F-UE1 generates a SIP INVITE request and sends it to the P-CSCF. The P-CSCF processes the request: for example, it decompresses the request and verifies the F-UE1's identity before forwarding the request to the S-CSCF. The S-CSCF processes the request, executes service control which may include interactions with ASs and eventually determines the entry point of the home operator of F-UE2 based on F-UE2's identity in the SIP INVITE request. I-CSCF receives the request and contacts the HSS to find the S-CSCF that is serving F-UE2. The request is passed to the S-CSCF. The S-CSCF takes charge of processing the terminating session, which may include interactions with ASs and eventually delivers the request to the P-CSCF. After further processing (e.g., compression and privacy checking), the P-CSCF delivers the SIP INVITE request to F-UE2.

TABLE I. SESSION ESTABLISHMENT MESSAGES

Session establishment message	Type of message	Compressed size (byte)
INVITE	Request	810
100 TRYING	Response	260
183 SESSION PROGRESS	Response	260
PRACK	Request	260
200 OK	Response	100
UPDATE	Request	260
180 RINGING	Response	260
ACK	Request	60

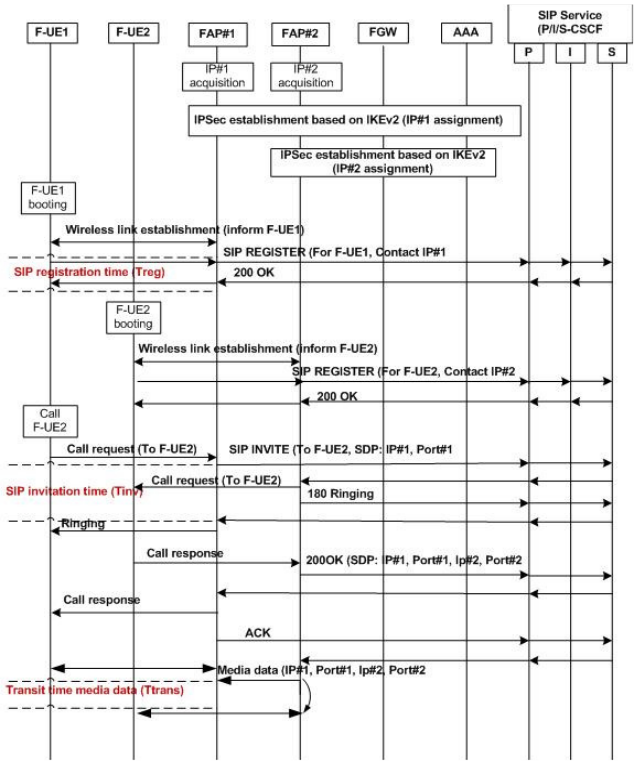


Figure 6. Session establishment signalling in femtocell

Correspondent F-UE2 generates a response – 183 Session Progress – which traverses back to F-UE1 following the route that was created on the way from F-UE1 (i.e., F-UE2 → FAP2 → F-GW → MME/SGW → P-CSCF → S-CSCF → I-CSCF → S-CSCF → P-CSCF → MME/SGW → F-GW → FAP1 → F-UE1). After a few more round trips, both sets of F-UE1 and F-UE2 complete session establishment and are able to start the actual application (e.g., voice calls).

Detail on IMS session establishment procedure can be seen in [13]. The complete signalling flow diagram of IMS session establishment procedure on femtocell network is depicted in Fig. 6.

C. Signalling Analysis and Performance

The Session Initiation Delay (SID) is often used to analyze the performance of session establishment signalling. SID is a user QoS parameter that can be defined as the period between the instant the F-UE1 triggers the initiate session command and the instant the F-UE1 receives the message that is alerted by F-UE2.

Obviously, there are many factors that influence the SID. In this paper we distinguish the delay into three categories, i.e., IMS connections delay, processing delay and queue delay. Thus, it can be approximated as:

$$SID = D_{IMS\_conct} + D_{proc} + D_{queue} \quad (1)$$

IMS connection delays are influenced mostly by the message propagation delay in the air interface (F-UE to FAP) and the transmission delay through the backhaul links and backbone network. In wireless transmission, the Radio Link Control (RLC) is used to improve the performance of frame error rate (FER) due to bandwidth availability and channel condition. The channel bandwidth (B/W) is assumed as 5 MHz, 10 MHz, 15 MHz and 20 MHz. Later, the channel rate (R) can be calculated by considering 3GPP TS. 36. 213 and TS 36. 306.

Furthermore, the number of frame a packet (k) is required to be calculated for every specified channel rates. The RLC frame duration or inter-frame time (τ) is assumed 10 ms for downlink access. Then number of bytes in each frame can be calculated as $\boldsymbol{R}$ x $\boldsymbol{\tau}$ x $\frac{1}{8}$. Subsequently, the value of $\boldsymbol{k}$ for particular signalling messages as shown in Table 1 can be calculated as [12]:

$$k = \frac{number\ of\ byte}{message\ size} \quad (2)$$

Since the detail operation of LTE-based femtocell radio link is not specified yet, the RLC delay analysis given by [15] is assumed to be used. The analysis of transport delay when transmitting a packet over the RLC is approximated as [15]:

$$D_{RLC} = D_{prop} + (k-1)\tau + \frac{k\left[P_f - (1-p)\right]}{P_f^2} \times \left| \sum_{j=1}^{n} \sum_{i=1}^{j} P(C_{ij}) \left[2jD_{prop} + \left(\frac{j(j+1)}{2} + i \right) \tau \right] \right| \quad (3)$$

The open-air operation radio access network is vulnerable to noise influenced that generate packet loss. In Eq. 3 above, the effective packet loss is noted by P_f, and can be calculated as follow [14] [15]:

$$P_f = 1 - p + \sum_{j=1}^{n} \sum_{i=1}^{j} P(C_{ij}) = 1 - p\left[p(2-p)\right]^{n(n+1)/2} \quad (4)$$

where k is number of frames, n is number of RLC retransmission trials, p the probability of a RLC frame being in error in the air link, D_{prop} is end-to-end propagation delay over the air interface, τ is the interframe time and $P(C_{ij})$ is the first frame received correctly at destination as the i^{th} retransmission frame at the j^{th} retransmission.

Based on Fig. 6, it can be assumed that there are A messages involved in the session establishment processes between F-UE1 and P-CSCF of the visited IMS network. In addition, there are also B messages involve between P-CSCF of the terminating IMS network and F-UE2. Therefore the IMS connection delay ($D_{IMSconct}$) is given as:

$$D_{IMS\_conct} = (A + B) messages \times D_{RLC} \quad (5)$$

Furthermore, the processing delay (D_{proc}) is determined at all entities in the IMS signalling path, i.e., P-CSCF, I-CSCF, S-CSCF in both home and visited networks, plus the HSS. It included the queue delay and address lookup table delay.

If number of messages processed in each entity is denoted as C, the delay in each entity (D_{proc}) can be approximated as:

$$D_{proc} = (C_{F-UE1} \times D_{F-UE1}) + (C_{PCSCF} \times D_{PCSCF}) + (C_{SCSCF} \times D_{SCSCF}) + (C_{ICSCF} \times D_{ICSCF}) + (D_{HSS}) + (C_{F-UE2} \times D_{F-UE2}) \quad (6)$$

Finally, the queuing delay for IMS session establishment signalling is determined in every network entities. The end-to-end packet delay from F-UE1 to F-UE2 depends on the number of the packets at each queue as well as the applied queue model (i.e., M/M/1, M/D/1, etc.).

By assuming that the transmission buffer at the network node is delay free, thus the queue delay is considered only at the receive buffer of F-UE1 and F-UE2. They can be approximated:

$$D[\omega_{F-UE1}] = \frac{\rho_{F-UE2}}{\mu_{F-UE2}(1 - \rho_{F-UE2})} \quad (7)$$

where $\rho_{F-UE2} = \frac{\lambda_{e-F-UE2}}{\mu_{F-UE2}}$ represents the utilization at originating terminal queue, μ_{F-UE2} denotes the service rate at originating terminal queue, λ_{e-CT} represents the effective arrival rate at originating terminal queue, which can be calculated as follow:

$$\lambda_{e-F-UE2} = \sum_{i \in N_{F-UE2}} \lambda_i \quad (8)$$

where N_{F-UE2} denotes the number of active sessions including the considered IMS session. Moreover, by determine the utilization at a network node, the effective arrival rate λ_e at that node can be obtained. In the same way to other network nodes, the λ_e at queue can be calculated. The expression can be determined for the expecting waiting time at other network entities.

Similar with the processing delay, if the number of messages processed in each entity is denoted as C, thus the queue delay for the IMS session establishment in femtocell network (D_{queue}) can be approximated as:

$$D_{queue} = (C_{F-UE1} \times D\omega_{F-UE1}) + (C_{PCSCF} \times D\omega_{PCSCF}) + (C_{SCSCF} \times D\omega_{SCSCF}) + (C_{ICSCF} \times D\omega_{ICSCF}) + (D\omega_{HSS}) + (C_{F-UE2} \times D\omega_{F-UE2}) \quad (9)$$

V. CONCLUSION

In this paper, the IMS signalling in femtocell network has been studied and analyzed. The LTE-based femtocell is considered for the integration with the IMS core network. The session establishment signalling procedure has is described and analyzed in order to show the IMS signalling functionality in femtocell network. We proposed the particular signalling call flows for session establishment procedure. In addition, the signalling performance is intended to be determined by mean of delay properties. The IMS connection delay, processing delay and queue delay are suggested to be considered as the delay properties of Session

 ISBN: 978-1-61208-100-7

Initiation Delay. The further work on simulation may corroborate the proposed mechanism and solution.

ACKNOWLEDGMENT

This work is part of the SGS project that was supported by the Grant Agency of the Czech Technical University in Prague, grant No. SGS10/274/OHK3/3T/13.

REFERENCES

[1] 3GPP Rel. 5. TS 23.228, "IP Multimedia Subsystem (IMS), Stage 2".

[2] 3GPP Rel. 6. TS 22.250, "IP Multimedia Subsystem (IMS), Group Management, Stage 1".

[3] 3GPP TS 29.163, "Interworking between the IP Multimedia (IM) Core Network (CN) Subsystem and Circuit Switched (CS) Networks".

[4] 3GPP TS 25.467, "UTRAN Architecture for 3G Home Node B".

[5] J. Rosenberg and H. Schulzrinne,"Reliability of Provisional Responses in the Session Initiation Protocol (SIP)". RFC 3262, June 2002

[6] 3GPP TS 24.229,"IP Multimedia Call Control Protocol Based on SIP and SDP".

[7] 3GPP TS 24.228 (v5.15.0), "Signalling flows for the IP multimedia call control based on Session Initiation Protocol (SIP) and Session Description Protocol (SDP); Stage 3".

[8] L. T. W. Ho and C. Holger,"Effect of User-deployed, Co-channel Femtocells on The Call Drop Probability in a Residential Scenario". Proc. the 18th Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC'07), 2007, pp. 1-5, 3-7 Sept. 2007.

[9] V. Chandrasekhar and J. G. Andrews,"Uplink Capacity and Interference Avoidance for Two-Tier Cellular Networks", Proc. Global Telecommunications Conference, 2007. IEEE GLOBECOM '07, pp. 3322-3326, 26-30 Nov. 2007.

[10] H. Claussen, L. T. W. Ho, and L.G. Samuel,"Self-optimization of coverage for femtocell deployments", Proc. Wireless Telecommunications Symposium, 2008. WTS 2008, pp. 278-285, 24-26 April 2008.

[11] V. Chandrasekhar, J. G. Andrews, and A. Gatherer,"Femtocell Networks: a Survey", IEEE Communications Magazine, vol. 46, no.9, pp. 59-67, Sept. 2008.

[12] A. Munir, "Analysis of SIP Based IMS Session Establishment Signalling for WiMAX-3G Networks". Proc. 4th International Conference on Networking and Services, pp. 282-287, 2008.

[13] M. Ulvan and R. Bestak,"Analysis of Session Establishment Signalling Delay in IP Multimedia Subsystem". Wireless and Mobile Networking, Berlin: Springer, 2009, pp. 44-55. ISBN 978-3-642-03840-2.

[14] H. Fathi, S. S. Chakraborty, and R. Prasad,"Optimization of SIP Session Setup Delay for VoIP in 3G Wireless Networks". IEEE Trans. Mob. Comput. 5(9): pp. 1121-1132, 2006.

[15] N. Banerjee, W. Wu, K. Basu, and S. K. Das."Analysis of SIP-based mobility management in 4G wireless networks". Computer Communications 27, pp. 697–707, 2004.

[16] T. Chiba and H. Yokota, "Efficient Route Optimization Methods for Femtocell-based All IP Networks", Proc. IEEE Int. Conference on Wireless and Mobile Computing, Networking and Communications, 2009, pp. 221-226, 12-14 Oct. 2009.

[17] 3GPP TR 23.830, "Architecture Aspect of Home Node B (HNB)/Home enhanced Node B (HeNB)".

[18] Motorola, "Long Term Evolution (LTE): Overview of LTE Air-Interface ". Technical White Paper. Can be accessed at http://business.motorola.com/experiencelte/pdf/LTEAirInterfaceWhitePaper.pdf . Last accessed on May 2010.

[19] W. Franks,"IMS femtocells: the basis for a new generation of IP-PBX?" http://femtocellpioneer.blogspot.com/2009/09/ims-femtocells-basis-for-new-generation.html. Last accessed on 18 April 2010.

[20] Singh, M., "Integrate Femtocells with Existing Wireless Infrastructure", Feature Article, Continuous Computing, www.ccpu.com. Last accessed on April 2010.

 ISBN: 978-1-61208-100-7

A Solution for Seamless Video Delivery in WLAN/3G Networks

Claudio de Castro Monteiro
Computing Department
Federal Institute of Education of Tocantins - IFTO
Palmas, Brazil
ccm.monteiro@ieee.org

Paulo Roberto de Lira Gondim
Department of Electrical Engineering
University of Brasilia
Brasilia, Brazil
pgondim@unb.br

***Abstract*—In this paper, we introduce a Session Proxy (SP) into a scenario in which wireless local network (WLAN) and 3th generation of networks cellular (3G networks) are integrated using peer-to-peer architecture, with mobile IP (MIP) as the mobility manager. This solution tries to preserve the quality of the streaming video upon each handover, recovering the user video session damaged by the high delay caused by MIP.**

Keywords-Video, 3G, WLAN, Handover, Proxy

I. INTRODUCTION

Currently, the demand for high performance services offered for mobile 3G connections has increased. In order to meet this demand, cellular providers have deployed 3G networks using the Universal Mobile Telecommunication System (UMTS) [4] and the Code Division Multiple Access 2000 (CDMA2000) [7].

However, due to the high cost of deployment and low transmission rates, cellular providers have not been successful in attracting customers.

Although cellular networks can have high transmission rates, the cost for mounting or adapting the necessary infrastructure is very high. On the other hand, the infrastructure of WLAN has been deployed in various countries and offers data rates much higher than 3G networks, with lower deployment and maintenance costs.

These two networks, therefore, complement each other and, if properly integrated, can provide the user with the appropriate conditions to access services regardless of the access networks being used.

The integration of these networks has been the subject of several studies in recent years. In general, these studies can be divided into three main areas, according to their chosen solution: those who use systems of engagement, those with mobility management and those with direct applications of IMS (IP Multimedia Subsystem) standards and MIH (Media Independent Handover).

The focus of this paper is to demonstrate the possibility of integration between WLAN and 3G using an architecture without coupling, with the peer-to-peer protocol MIP as the mobility manager and a proxy session to ensure continuity of the connection sessions for the user in order to maintain quality when receiving video streams.

The effectiveness of our proposal is shown through experiments conducted in real life with a WLAN network, over which we have total managerial control, and a 3G network from a local carrier in Brazil, over which we have no management control. The technical difficulty of access to the local cellular operator has motivated us to create this solution, which, in principle, is not dependent on the resource deployment of any hardware or software in the core of mobile carriers.

This environment has a multi-mode mobile terminal, containing two network interfaces (WLAN and 3G). In addition, an access point based on Linux was implemented to generate the WLAN coverage. An implemented MIP Home Agent (HA) is at the core of a WLAN, while a MIP Foreign Agent (FA) is implemented and available on the Internet.

Thus, we show that the measured delay during the handover between the mobile WLAN and 3G is reduced by 40% as compared to results in the literature, and we propose that this integration uses MIP.

However, this reduction of delay is not sufficient for sessions that are connecting video stream applications to have continuity.

Therefore, we insert a proxy session at the core of a WLAN that is able to cache the video frame after receiving the unit, signalling the start of a handover and delivering the frames to the mobile once the handover is completed, using a tunnel made between the SP and the mobile and duly recording between the FA and the HA.

The article is divided as follows: In Section II, we present some related works. In Section III, we present our proposal including the algorithms used to implement the SP. In Section IV, we present the testbed used to test and validate the proposed solution. In Section V, we present the conclusions of the work.

II. RELATED WORKS

The problem of reduced video quality caused by the handover in WLAN/3G integrated networks has been treated to some extent.

However, this problem has been divided into two parts: one part studies ways to integrate WLANs with 3G networks in order to reduce delays in the exchange of control messages between the two networks, in an attempt to make the user feel as though he is working with a single network; and the other part studies methods to maintain a video session when a handover process occurs, by adapting the transmission.

We analyze works that address solutions for this problem, focusing on these two categories.

In [1], the authors present a scheme for dynamic negotiation of QoS parameters between access points. The paper introduces a concept of BAG (Bandwidth Aggregation), which allows the user to dynamically

 ISBN: 978-1-61208-100-7

negotiate, using the access points involved, the path to be used to request a service.

Thus, users can move between access points, keeping the best service condition possible. Despite showing the efficiency of the solution, the authors only considered the bandwidth dependence of the QoS parameters, without emphasizing the reflections in the delay, jitter and packet loss that the solution generates. However, the proposal shown for QoS negotiation allocates the resources after the handover.

Nothing is reported about the handover period, still leaving open questions that are very important with regard to streaming video.

The work presented in [2] proposes a comparative study of video-encoding tools and storage techniques used for the delivery of video on demand to users on a next generation network.

The paper analyzes the performance of these tools and techniques in a tightly coupled environment using the IP Multimedia Subsystem (IMS) connected to a live operator network.

All results are presented at the Quality of Experience (QoE) parameter level and indicate that some of these techniques can be used to help solve the problem treated in this paper.

However, the techniques and tools were evaluated in a tightly coupled environment, where it was assumed that the cellular operator had control over the technical solution.

This is not the case with our situation, since we are proposing a structure in which the cellular operator does not need to have control over the solution and does not have to install any hardware or software components in the core.

In [3], the authors presented and compared three strategies for checking the quality of user experience for the reception of video streams.

Although the authors presented three strategies, the paper focused more on the results of a hybrid strategy, called Pseudo Subjective Quality Assessment (PSQA), which, according to the authors, presents advantages over the traditional PSNR.

This work demonstrates the applicability of the objective QoE metric, known as PSNR, which does not have temporal characteristics of the delays inherent in the transmission of video streams over the network, but the excessive loss of frames only.

This has led us to choose PSNR over the PSQA, whereas we measure the quality of perception of the user in the continuing video session after the vertical handover of the mobile.

Some studies found in the literature present coupling architectures for 3G-WLAN interworking. One of these studies is shown in [4], where the authors propose a loosely coupled architecture called SHARE.

In this proposal, each AP has a WLAN card to connect to a 3G network, which is used for the AP and can be connected simultaneously to the cell WLAN and cellular 3G.

The result is that 3G base stations can share their control channels with the APs, facilitating the detection of available APs in the region for mobiles, without, according to the authors, generating the characteristic delay of discovery networks in the handover process.

The proposal of the paper is good, and we found that the delay caused by network discovery activity in the handover process is small enough, considering the cost of inserting a 3G connection hardware into each AP and software procedures into both APs and 3G base stations.

Thus, although the solution of the authors has inspired us, we consider a scenario in which 3G networks and WLAN are overlapping and in which the mobile already has both 3G and WLAN interfaces available and is connected to each network, so that there is no need to detect the available networks.

In addition, we are concerned with reducing the handover latency at the level of reconnection IP, which is mandatory after the handover process and is decisive in solving our main problem.

In [5], a framework was presented to assist with admission control and resource reservation in next generation mobile networks.

The proposal presents a distributed call admission control with resource reservation, using the position information of the mobile to predict the movement of the user, determining the acceptance of the connection requests that can be made to that cell, which is based on the probability of movement of the user and the resources required by the connection.

However, this solution requires the use of a Global Position System (GPS) for each base station to find the location of the mobile and provide its movement in order to determine whether that cell has the resources required by the mobile.

This solution has a high cost and requires hardware and software elements to be installed into the 3G network core, with regard to admission control based on resource reservation.

Nothing is needed to maintain continuity of the session of the mobile connection after the handover, since we can predict the movement and allocate resources to be used after the handover of the mobile, but if a session is already open, the handover latency can be decisive for continuity of the session.

In an attempt to maintain multimedia services during a vertical handover between WLAN and 3G, we find in [6] an assessment of the support mobility mechanisms proposed by IEEE 802.21.

The authors presented the results of application studies of these mechanisms to assist with the handover in both directions, using an implementation of an MIH Link Going Down event for IEEE 802.11 networks.

The authors validated their results using a simulated environment with the NS-2 software. In the proposal, it is unclear how the multimedia session connection will be maintained, since the Link Going Down event informs the mobile that the link is suffering a fading.

Thus, other actions should be taken from this point forward to ensure the sessions, depending on the multimedia traffic type.

In our study, although MIH could have been implemented into the mobile and WLAN core, we chose to implement (only in the mobile) software elements that verify the operation of the device interfaces.

If there is any change in the active interface (the one used as the standard output), here represented by being powered

 ISBN: 978-1-61208-100-7

off, the hard handover process is then characterized, and all other procedures related to re-establishing the IP connection and caching the video frames are started.

A framework for interworking between a WLAN and the UMTS network is presented in [7]. This proposal uses the IP multimedia subsystem suggested by the 3GPP for mediating network coupling in order to manage real-time sessions.

The work uses the IMS in order to provide for a user from a cellular network and QoS (Quality of Service) features available in a WLAN, while also providing a common and unified platform for control session user connection.

This architecture provides service control for a mobility terminal in an environment of heterogeneous wireless networks. The work validates the proposed model to ensure continuity of services during and after the handover by simulation.

The work ensured that a better continuity of service was possible. The model does not explain how to characterize the traffic during the simulation and focused only on the coupling of the WLAN and 3G networks using IMS, assuming that the operation of IMS elements can reduce the latency caused by the exchange of messages during the handover.

When proposing an unified architecture with mediation by the IMS, the handover is to be transparent to the user with the viewpoint of exchanging signalling messages.

The solution is interesting but requires that significant changes be made to the core of the networks involved, which may be difficult in a real scenario.

Our solution uses an architecture without coupling and therefore does not require that changes be made to the core 3G network, while requiring only some simple implementations in the core WLAN network.

The authors in [8] proposed a practical design of a proxy agent – SPONGE (Stream Pooler Over a Network Graded Environment) - localized between the wireless user equipment and the streaming video to facilitate the adaptation of delivering video stream service by wireless networks.

The architecture proposes the storage of a video, encoded at different qualities, permitting the user equipment to receive videos according to the characteristics of their network.

The authors focused their contribution to ensuring the video delivery, considering the conditions of the network QoS and the quality of the streams stored. The results did not lead to a guarantee of video stream delivery during or after the handover of the user.

Our work, despite having a similar direction, presents a Session Proxy (SP) without considering multiple video qualities with the aim of adapting the delivery.

The SP is presented as an alternative to associating a connection session of a mobile with a central element, designed to lead the intermediate cache frames of video transmission during the vertical handover of the mobile, in an attempt to deliver them in the same sequence after the handover.

The work of a SP is supported by our strategy for MIP implementation, which contributes to the reduction of handover latency in the re-establishment of IP connections.

A proposal of a handover study in mobile IP networks and mobile IP protocol extensions for handover latency minimization was given in [9], indicating that native mobile IP has a very high handover latency, and the proposal improved the performance of handover latency by 15%. Our proposal reduces this latency by 40%.

In [10], a proxy-based multimedia scheme is proposed for control Real Time Streaming Protocol (RTSP) to support fast signalling in the home network. The testbed implementation showed that the proposed scheme improved performance as compared with the RTSP in terms of the latency time, but it did not resolve the RTSP session continuity problem. The proposal reduced latency time, but the loss rate was large enough that the RTSP session could not continue.

A framework is proposed in [11] for multimedia delivery and adaptation in mobile environments. This work introduced the concept of a Personal Address (PA), which is a network address associated with the user instead of a network interface.

The proposed framework works at the network layer, and it moves the PA among networks and devices to deliver media in a seamless and transparent way.

The authors claim that the location's transparency sponsored by the PA allows the user to receive multimedia data independent of the IP network.

However, the solution presented used a mobile IP and did not show how it impacted the transmission multimedia session continuity, influenced by the implementation of the entities that manage the PAs.

All of the related works studied attempt to resolve problems in wireless network integration for service delivery.

In special cases, solutions have been reported for the integration and interoperation between WLAN and 3G networks in order to ensure the video delivery to users of these networks.

The problems discussed include the handover. Many studies have adopted ways to reduce the handover latency using MIH, IMS protocols, mobility, coupled architectures, algorithms for network discovery, or even an integrated set of techniques.

Our work attempts to solve the problem in two stages. First, we suggest an architecture without engagement and the implementation of a MIP to manage the mobility of the user, considering an architecture in which no change is required in the core 3G network.

In the second stage, we suggest a session proxy, implemented into this architecture, that is able to provide the user with session continuity after the video handover, which is granted only with the MIP.

In general, video stream sessions have a synchronization time that does not support the handover latency. The studies found in the literature handle problems with enlace retransmission techniques, with different frame types that deliver only the necessary or usual application technologies of mobile IP.

Thus, our solution is based on a set that uses a mobile IP and a session proxy (SP). The proposal attempts to resolve the session continuity problem after handover, ensuring the

 ISBN: 978-1-61208-100-7

receiver quality (PSNR – Peak Signal Noise Ratio) of the video transmitted.

III. PROPOSAL

In our proposal, we suggest inserting two components in the WLAN core operator: a MIP HA implementation and a Session Proxy (SP).

In addition, a MIP FA implementation is proposed for working through the Internet, in order to register the Care of Address (CoA) of the mobile. In Figure 1, these components and their links can be seen.

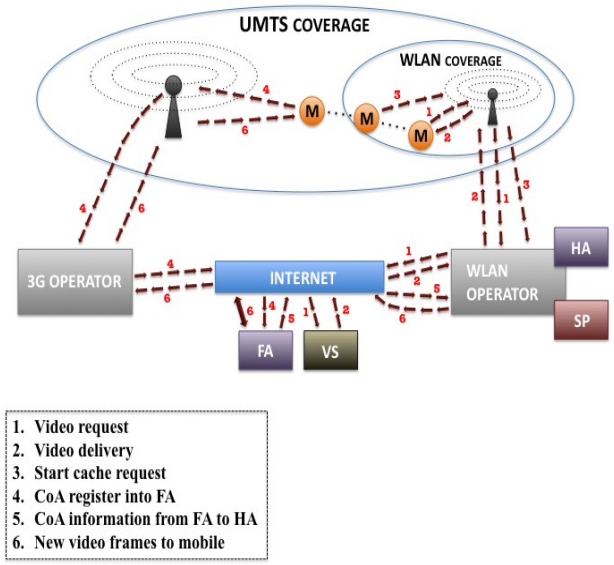


Figure 1. Proposed Architecture

The main idea is to maintain the continuity of the user video session after handover. To accomplish this, we set up a testbed containing native Linux MIP implementation only, as shown in Figure 2.

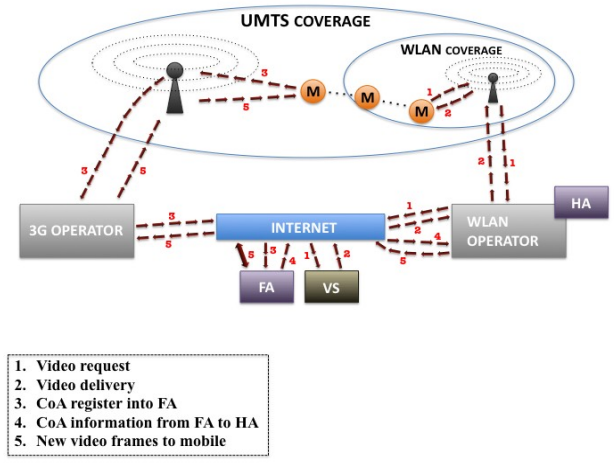


Figure 2. Proposed Architecture with MIP implementation only

A. *MIP utilization and implementation*

We considered 200 unicasts of video highway, requested by the mobile. When the video stream reaches frame 400, we turn off the WLAN interface of the mobile, make the client portion of the Linux MIP native implementation identify the fact and change the default route of the mobile to the 3G network gateway, telling the FA the new IP and triggering the registration process between the FA, the HA and the mobile, described in the specification MIP.

The results, as expected, showed a high delay during handover, reaching an average of 5 seconds, whereas this implementation presented difficulties in establishing the tunnel between the HA and the mobile, plus some additional delay in the necessary registration between the mobile and the FA and between the FA and the HA. For non-real-time applications, this result is satisfactory, but not for real-time applications.

To try to reduce the delay caused by the MIP handover, we decided to implement a HA and a FA, based on raw sockets, and to restrict the MIP signalling to recording the new IP address in the FA and the registration of the new address in the HA, made by the FA.

After that, the HA adds a new route to reach the mobile, using it for the establishment of an IP tunnel, with the help of openvpn software.

This implementation was more efficient than that available natively on Linux; it reduced the delay by 40%, as shown in Figure 3. However, 3 seconds is still a long delay for video session connections to be maintained.

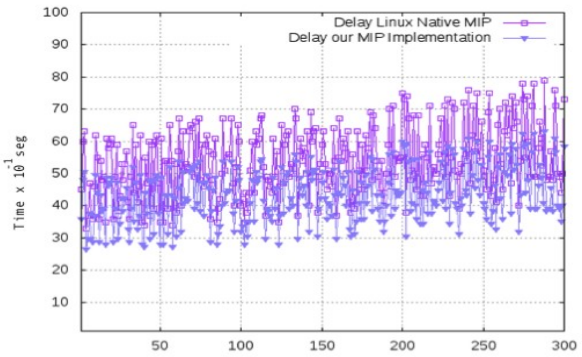


Figure 3. Native MIP vs proposed MIP delay

We observed that the objective QoE (Quality of Experience) metric, known as the PSNR, does not suffer a direct influence of the delay in receiving frames, rather an excessive loss of frames at the reception of the video, which can cause a significant reduction in the PSNR values obtained, as shown in Figure 4.

Thus, the loss of the video session during the reception generates, from the viewpoint of the player, an excessive loss of frames, directly affecting the PSNR of the received video.

 ISBN: 978-1-61208-100-7

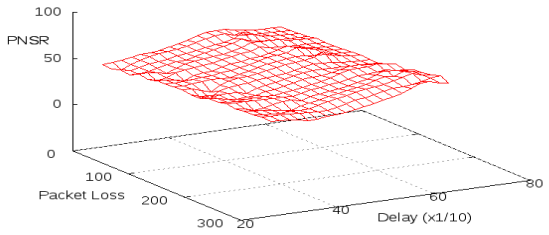


Figure 4. PSNR vs Frame Delay and Loss

B. *Session Proxy*

To address the problem of a continuous session during the transmission of video streams, we included a software entity in the proposed architecture, called a Session Proxy (SP) [12], as shown in Figure 1.

The SP ensures the continuity of the session even after long periods of link discontinuity. Although the SP can use the prediction of a handover of the mobile, through the thresholds defined after the extensive experiments detailed and displayed in Table 1 [12], for simplicity, we assume that the handover will be determined by the number of frames received from the video, using a setting of 200 frames; a hard handover will occur if the active interface of the mobile is turned off.

Before that, upon receiving frame 150, the mobile informs the SP that it is about to start the handover and that it should start the cache frames.

Thus, the mobile requests an open session RTSP with the video server. This request will be received by the SP, registered with the structure shown in Table 1 and then forwarded to the video server, according to the algorithm shown below.

```
receive_socket(socket, RTSP_request);
registered_session(session_ID, RTSP, IP_MAC, 0);
open_socket(socket1, IP_server);
send_socket(socket1, RTSP_request);
receive_socket(socket1, RTSP_response);
send_socket(socket, RTSP_response);

while(Session_ID <> 0)
{
        receive_socket(socket, RTSP_packets);
        receive_socket(socket2, status, MAC_AP);
        if(status==1)
        {
                FrameID=frame_ID;
                start_cache(Session_ID);
        }
        if(status<>1)
        {
                send_socket(socket1, RTSP_packets);
        }
        sendcache_socket(socket1, RTSP_packets);
}
```

The video server then opens an RTSP session with the SP, which will begin to receive the frames and transfer them to the AP, which delivers it to the mobile node. This process will continue until the mobile receives video frame 100, indicating that the SP should start the frame cache.

At this point, the SP-cached frames are transmitted to the mobile, using the data structure shown in Table 2. When the mobile receives frame 150, the hard handover will start.

When the mobile receives video frame 100 and informs the SP, it records the identifier of the last frame received in the session registration cache and continues with the video server session open, receiving frames, inserting in the cache and transmitting to the mobile.

When the active interface of the mobile is switched off, the MIP is employed in order to achieve the necessary records for rebuilding the user's IP connection, as described in the previous section.

This informs the SP that it must start the transmission of frames in its cache following the last frame that was received by the mobile, now using the new path created by our MIP implementation.

TABLE I. SESSION REGISTRATION CACHE STRUCTURE

Session ID	Service ID	IP association	Frame ID

With the use of a SP, we continue to deliver the video after the handover of the mobile, without the loss of frames influencing the quality of the video received. The display shown to the user stops for 2.23 seconds, with the next frame playing after the last frame received.

In Figure 5, can see the PSNR of the video received with and without the use of a SP. Note that when we use the SP, the values do not have variations that could impair the quality of the video received. This is not so when we do not use this software component, demonstrating the efficiency of the proposed mechanism.

With a SP, the mobile receives frames after the handover, from the last frame received before the handover, without affecting the quality of perception of the user.

Moreover, without the SP after the handover, the mobile no longer receives the remaining frames of the video, which, after a certain number of frames not received, reduces the PSNR of the received video. In practice, the user realizes that the session connection has been interrupted and that the video has stopped being displayed.

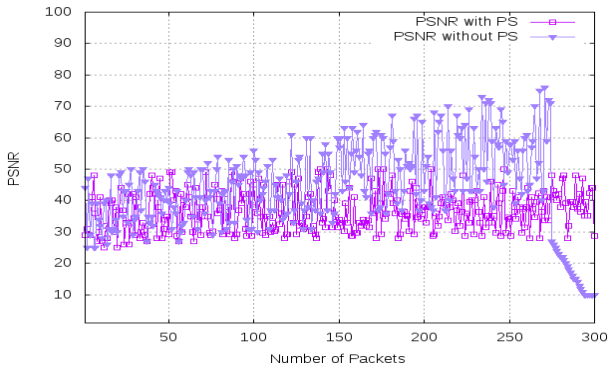


Figure 5. PSNR video with and without SP

 ISBN: 978-1-61208-100-7

C. Mobility Considerations

For testing purposes, we first considered the user mobility of WLAN → 3G. However, as we are not considering a soft handover, the methodology used to study the behaviour of the proposed architecture in the case of the user mobility of 3G → WLAN was exactly the same, and the results have similar values, since the components that are involved are the latency of re-establishing IP connectivity (in implementing MIP) and session user continuity (with a session proxy).

IV. Testbed Scenario

For testing, we mounted a scenario composed of three netbooks with the Linux 2.6.31 operating system.

The first netbook was used as a multi-mode device, with a WLAN interface (atheros) and a 3G interface.

The second netbook was used as a HA and network core of WLAN, providing Internet Gateway, NAT and DHCP for WLAN users.

This netbook also employed our implementation of a SP and a HA. The third netbook was used as a FA and was installed within the IP network of the Laboratory of Interactive Digital TV, UnB.

In this machine, our implementation of a FA was installed as well. Moreover, it also used an access point CISCO / LINKSYS, installed in an IFTO IP network, and a video server, with VLC software, installed in the LabTVDI/UnB IP network.

Importantly, our solution shows results based on an environment in which nothing needs to be inserted into the core of the 3G operators, signalling a possible offer of transmission services, continuous video streaming, free of charge.

V. Conclusions and future works

This study provided evidence supporting our two hypotheses: i) It is possible to reduce the delay caused by the implementation of MIP. This was achieved with the methodology of MIP implementation presented. ii) The use of a SP is also effective in environments in which heterogeneous wireless networks are integrated.

Our main goal was to reduce the delay caused by a vertical handover to minimize its impacts on the receipt of videos.

Using only the techniques of mobility management, some studies have proposed solutions that reduce the latency of the handover in different layers of the OSI model.

Although we also attempted (successfully) to reduce the latency in layer 3, the reduction was not sufficient to maintain the continuity of video sessions after a vertical handover.

Other studies have presented solutions to work around this problem generated by the delay in video connections subjected to a handover. These works have tried to adapt the application to network conditions, often using techniques of video compression and selective transmission.

Our work presents the concept and implementation of a proxy session that mediates the process of accessing the video, performing an activity together between the network and the mobile, which results in effectively maintaining the quality of video received at that session connection.

As a continuation of this work, we are developing a mechanism of Admission Control (CAC), based on distributed brokers that implement a primitive MIH and that can help the mobile to identify and determine the best network to perform the handover, in order to control the entry of the mobile into networks in accordance with their demands for QoS/QoE. We intend to do this by considering brokers independently of the networks available.

References

[1] Fernandez, J.C., Taleb, T., Guizani, M. and Kato, N. (2009). Bandwidth aggregation-aware dynamic QoS negotiation for real-time video streaming in next-generation wireless networks. IEEE Transactions on Multimedia, 11(6), 1082-1093.

[2] *Kuwadekar, A. and Al-Begain, K. (2009). User Centric Quality of Experience Testing for Video on Demand over IMS. First International Conference on Computational Intelligence, Communication Systems and Networks (pp. 497-504).*

[3] Piamrat, K., Viho, C., Bonnin, J. M. and Ksentini, A. (2009). Quality of Experience Measurements for Video Streaming over Wireless Networks. Sixth International Conference on Information Technology: New Generations. (pp. 1184-1189).

[4] *Lim, C., Kim, D.Y., Song, O. and Choi, C.H.. (2009). SHARE: seamless handover architecture for 3G-WLAN roaming environment. Wireless Networks, 15(3), 353-363.*

[5] *Judith, F.L. .(2008). Approaches to resource reservation for migrating real-time sessions in future mobile wireless networks, Wireless Networks, 13(2), p39-56.*

[6] *Machan, P., Serwin, S. and Wozniak, J.". (2008). Performance of mobility support mechanisms in a heterogeneous UMTS and IEEE 802.11 network offered under the IEEE 802.21 standard. 1st International Conference on Information Technology (pp. 1-4).*

[7] *Munasinghe, K. and Jamalipour, A. (2008). Interworking of WLAN-UMTS Networks : An IMS-Based Platform for Session Mobility. IEEE Communications Magazine, 46(9), 184-191.*

[8] Leu, J. S. and Tsai, C. W. (2009). Practical design of a proxy agent to facilitate adaptive video streaming service across wired/wireless networks. Journal of Systems and Software, 19(5), 1-12.

[9] Malekian, R. (2008). The Study of Handover in Mobile IP Networks. Third International Conference on Broadband Communications, Information Technology and Biomedical Applications (pp. 181-185).

[10] Lee, J. M., Yu, M., Choi, S. G., Lee, Y. R. and Kang, S. M. (2008). Proxy-based multimedia signaling scheme using RTSP for seamless service mobility in home network. International Conference on Consumer Electronics (pp. 1-2).

[11] Bolla, R., Mangialardi, S., Rapuzzi, R. and Repetto, M. (2009). Streaming multimedia contents to nomadic users in ubiquitous computing environments. 28th IEEE international conference on Computer Communications Workshops (pp. 218-223).

[12] Monteiro, C. C., Gondim, P. R. L. and Rios, V. M. (2010). Seamless Video Session Handoff between WLANs. HINDAWI International Journal on Computer System and Networking, 2010(2010), 36-46.

 ISBN: 978-1-61208-100-7

REST-based Meta Web Services in Mobile Application Frameworks

Daniel Sonntag
German Research Center for Artificial Intelligence (DFKI)
Stuhlsatzenhausweg 3,
66123 Saarbruecken, Germany
daniel.sonntag@dfki.de

Daniel Porta
German Research Center for Artificial Intelligence (DFKI)
Stuhlsatzenhausweg 3,
66123 Saarbruecken, Germany
daniel.porta@dfki.de

Jochen Setz
German Research Center for Artificial Intelligence (DFKI)
Stuhlsatzenhausweg 3,
66123 Saarbruecken, Germany
jochen.setz@dfki.de

***Abstract*—This paper describes how a multimodal dialogue application framework can be used to implement specific mobile applications and dynamic HTTP-based REST services. REST services are already publicly available and provide useful location-based information for the user on the go. We use a distributed, ontology-based dialogue system architecture where every major component can be run on a different host, thereby increasing the scalability of the overall system with a mobile user interface. The dialogue system provides customised access to the *Google Maps Local Search* and two REST services provided by GeoNames (i.e., the *findNearbyWikipedia* search and the *findNearbyWeather* search).**

***Keywords*-Multimodal Dialogue, Application Backend, REST Services**

I. INTRODUCTION

Over the last several years, the market for speech technology has seen significant developments [1] as well as powerful, commercial off-the-shelf solutions for speech recognition (ASR) or speech synthesis (TTS). However, these infrastructures have had only moderate success so far in the entertainment or industrial sector, especially in the mobile user interface (UI) context. This is the case because a dialogue system cannot easily be constructed for a mobile application. Only distributed systems, where the speech input processing is done on a server, allow for real-time dialogue reactions. Additionally, the dialogue engineering task requires many customisations to specific end user applications.

A modern multimodal dialogue system can act as the middleware between the mobile clients and the backend services, which hides complexity from the user. It should present only aggregated information to the user, thereby customising the presentation rules to the specifics and requirements of various output devices. This is possible because these architectures often encapsulate the dialogue proper from the rest of the application. These architectural decisions are often based on usability issues that arise when dealing with end-to-end dialogue-based interaction systems for industrial dissemination. Prominent examples of integration platforms include TRIPS [2], Galaxy Communicator [3], and SmartWeb [4]; these infrastructures mainly address the interconnection of heterogeneous software components. Earlier projects [5], [6] have integrated different sub-components into multimodal interaction systems. Thereby, hub-and-spoke dialogue frameworks played a major role.

We use the encapsulation characteristic for our own benefit. In multiple tier architectures, multiple user interfaces can be used more easily, and the application backend may comprise of several information sources. Accordingly, the multimodal dialogue application framework can be used to connect specific mobile interfaces at the frontend with dynamic HTTP-based REST [7] services at the backend. Works in mobile application frameworks are often concerned with physical issues, e.g., supporting wayfinding with tactile cues [8] or interactive experiences for cyclists [9]. Other work concerns mobile search scenarios and incidental information [10] where contexts such as location and time play major roles in information discovery [11] or the design of Web-based mobile services [12]. Our work addresses available services through an application backend in a three tier architecture and focused on multimodal dialogue for the user interaction with locations on a map.

In what could be a typical application scenario, the user is visiting historic sites in Berlin and wants to get location-based (restaurant) information. In previous projects [13], we used a Web service infrastructure for the backend access. We also developed a semantic representation formalism based on OWL-S and a service composition component to interpret an ontological user query. Although the composition module was able to dynamically compose different Web services for hotel, restaurant, and theater information, the maintenance of private/public WSDL Web services was more difficult and produced a high degree of dependence on external service providers. By contrast, the interaction with HTTP/REST services is much simpler, and these services provide useful location-based information, too. For combined searches with geographical coordinates, we selected the *Google Maps Local Search* REST service and the two services for reverse geocoding provided by GeoNames for inclusion, namely the *findNearbyWikipedia* search, and the *findNearbyWeather* search.

This paper is structured as follows: in Section II, we will

 ISBN: 978-1-61208-100-7

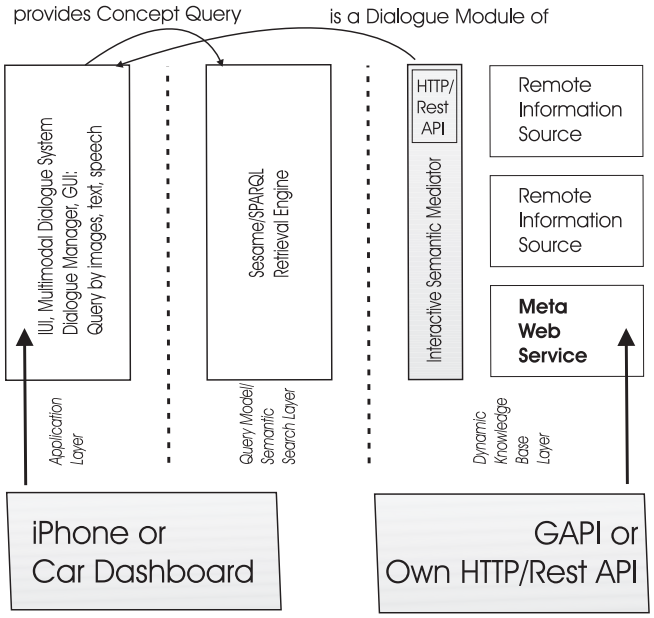


Figure 1. Three Tier Querying Architecture

present the application architecture, followed by the detailed application scenario (Section III). Section IV provides a conclusion and future work.

II. Application Architecture

Figure 1 outlines the three tier architecture. It consists of an application layer (the dialogue system), a query model and semantic search layer (which, e.g., extracts keywords from the result of the natural language understanding, NLU), and a dynamic knowledge base layer which addresses information sources in general. The knowledge layer hosts the interactive semantic mediator (providing information-source-specific queries) and remote information sources. The most important characteristic is the three tier distribution [4]. Multiple input/output devices and multiple heterogeneous information sources can be integrated into the respective technical layer. We already evaluated the architecture for a customised iPhone application [14] and a car dashboard [4] while addressing SOAP-based Web services as information sources. In this paper, we present a completely new technical implementation: the REST API in combination with a generic graphical UI concept for the iPhone.

A. Meta Web Services

Many Web services (that can be found on the Web) can be directly used as key components in a Web service composition process. For example, we have been experimenting with services that provide keyword-to-SPARQL query functionality (for the basic idea, see [15]).

Apart from services which help the composition process, other standard Web services such as *Amazon Web Services* exist that are often part of such composition chains. Preferably, SOAP services are considered for composition due to their WSDL [16]-formalised technical interface. We also consider services based on REST, JSON, XML-RPC, and the like.

Whenever we provide a custom interface to a compound of those services, we speak of *Meta Web Services*. A Meta Web Service provides a complex processing service. For example, one such Meta Web Service maps the GAPI [17] results on a ranked result table containing several YouTube videos with metadata (title, list of comments, etc.). The ranking of the videos is done by computing a string-based distance between the (keyword) query terms and the textual description of videos.

One prominent graphical framework for building Meta Web Services by aggregating different HTTP/REST-based information sources is Yahoo Pipes [18]. However, the aggregation processes implemented in such frameworks are hard-wired and do not allow the conditional execution of sub-processes. Additionally, information extraction from unstructured textual information sources is (apart from the extraction of location information) not supported in the framework. In contrast, the Business Process Execution Language (BPEL) [19] allows for more degrees of freedom. It defines a standard and complete declarative language for the composition of Web services. Graphical editors for modelling such processes are also available. Unfortunately, services in the composition chain have to be formally described in terms of WSDL (which is often not the case for HTTP/REST-based services). The Posr framework [20] provides a holistic approach for Web service composition not only on the technical but also on the UI layer. Although Posr is able to transform browser-based Web forms into a Meta Web Service, they only consider WSDL-formalised SOAP Web services and leave the inclusion of HTTP/REST-based services as future work. Hence, one of our main goals is the inclusion of HTTP/REST-based services into a mobile dialogue architecture. Our solution is a custom HTTP/REST Meta Web Service for mobile location-based scenarios.

B. Custom HTTP/REST Meta Web Service

As shown in Figure 2, our own HTTP/REST API service comprises of three modules: the query builder, the query/retrieval module and a set of presenters as the result aggregation module. The query builder takes keywords as input and creates the HTTP/REST URL query with respective arguments. The query/retrieval module handles the information retrieval step. An incoming result triggers a state change in the result aggregation module. The module informs all subscribed presenters of the new result. The presenters can serialise the XML results into files for logging purposes, or parse the results using a SAX parser to provide data tables or platform/ontology-specific exchange objects (TFS).

 ISBN: 978-1-61208-100-7

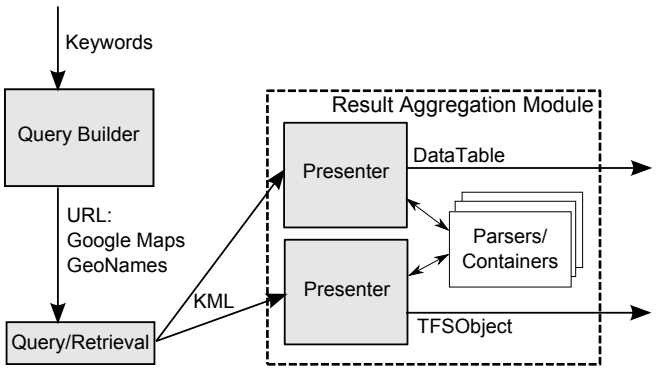


Figure 2. Own HTTP/REST API

Name	Rating
VAU	4.0
Fernsehturm	4.1
Letzte Instanz	3,4
Pergammonkeller	
Brecht	3,8

Name	Fernsehturm
Rating	4.1
Street	Panormastaße 1a
Phone	030 24757537
WWW	http://tv-turm.de
Pic	http://...

http://maps.google.com?
q=Restaurant&near=Berlin+M
&output=kml&mrt=all
&oe=UTF8&v=2.0

Figure 3. Results of a Google Maps Local Search

1) Google Maps Local Search: Currently, Google only provides a public JavaScript API for Google Maps; therefore, we had to rely on publicly available REST API specifications as, e.g., provided by Mapki [21]. Mapki provides an almost complete parameter description for accessing Google Maps functionalities. We selected a suitable subset of these parameters (Table I) for our mobile scenario. The `near` parameter is of particular importance as it parametrises a local search with a point-of-interest (POI) as starting point. The `output`, `oe`, and `v` parameters specify the result format. In our case, an *UFT-8* encoded *KML* [22] file is requested. The `mrt` can be used to narrow down the

Parameter	Value	Description
q	category	Keywords for the search
near	a location	Starts the local search around the POI location
mrt	all	Get as much information as possible
output	kml	Get the results as a kml file
oe	UTF8	Encoding of the output
v	2.0	Version of the KML schema

Table I
PARAMETERS USED FOR QUERYING GOOGLE MAPS

Name	Distance
Fernsehturm	0.009
Alexanderplatz	0.13
Karl-Liebknecht-Straße	0.1982
St. Mary's Church	0.239
Neptunbrunnen	0.2758
Rotes Rathaus	0.2973

Title	Rotes Rathaus
Summary	The Rotes Rathaus is the town hall of ...
Distance	0.2973
Wikipedia Article	http://en.wikipedia.org/...

Figure 4. Results of a Wikipedia NearBy Search

results at query time according to categories, e.g., locations, businesses, places annotated by users, and the links to Wikipedia. For example, search results in the HTTP response for the query “Restaurant near Berlin Mitte” are depicted in Figure 3.

Google Maps’ answers contain detailed pieces of information about restaurants or hotels including the names, user ratings, address data, the URL, a link to a thumbnail, and the geographical coordinates. The vast bulk of this information is well-structured in the KML output. Some details, however, e.g., phone numbers and user ratings, are only available as HTML fragments within CDATA text blocks in the KML structure. Hence, Java classes for regular expression String matches are also necessary. Similar customisation requirements are expected also for new Meta Web Services. This is clearly a drawback for the general scalability of HTTP/REST service integration. As Google Maps does not provide any Wikipedia information in the desired KML output (Google Maps only links to Wikipedia’s HTML) and since we are also interested in additional POIs and/or Wikipedia results, we access additional services. In the context of our mobile scenario, we decided to use the GeoNames [23] REST API, i.e., the *findNearbyWikipedia* search.

2) GeoNames: GeoNames provides two Web services for reverse geocoding which fit in our application scenario, namely the *findNearbyWikipedia* and *findNearbyWeather* services. (Both services take latitude, longitude, and radius as parameters.) The Wikipedia search provides links to articles about memorials, landmarks, or other interesting places. The weather search provides the current weather conditions. Wikipedia results contain summaries, distance, and the link to the full article, and other information (Figure 4). As the distance is just the beeline between the starting point and the place, we used Google Maps again to get the route planning

and actual walking distance as an additional result for the mobile user. The *findNearbyWeather* service obtains, apart from temperature, details about the weather conditions such as wind speed or humidity. Certainly, weather conditions are not really meaningful when planning a city walking tour (San Francisco's micro climate might be an exception) but can become very important when going hiking [24].

C. Mobile Client Application

Nowadays, many different mobile device platforms exist. The majority is equipped with a full-fledged Web browser that enables us to provide platform-independent graphical user interfaces (GUIs) by means of DHTML-based Rich Internet Applications (RIA). On the code basis (in our case a declarative XML-based language), we make use of the OpenLaszlo Rich Internet Framework [25], which turned out to be very suitable (in contrast to the findings presented in [26]) for Web applications on modern mobile devices without the need for a Flash player. Since we need to send and receive optional audio data for speech-based interaction, we implemented a lightweight native application that embeds a full-screen Web browser and additionally provides a platform-dependent audio streaming functionality (a similar approach is pursued by [27]). The communication with the dialogue system is implemented by a uniform client-side JavaScript API using long polling AJAX (Asynchronous JavaScript and XML) requests to a server-side HTTP/REST-based endpoint. Currently, client applications exist for the iPhone and the Android platform. Desktop Web browsers can (in combination with an optional Java applet for audio streaming) also render the DHTML-based GUIs, which eases the development process of such multimodal UIs.

III. Application Scenario

First, Google Maps supplies the geographical coordinates of a POI such as a hotel or restaurant. We use the coordinates to get a list of other interesting places close by. Then, we display further POIs on the current map. Second, the user can ask for additional POIs in the vicinity according to a manual POI selection on the map which initiates a GeoNames search. The geographical coordinates from Google Maps or GeoNames can, third, be used to obtain weather conditions or Wikipedia hits around the resulting POI.

The following dialogue between a user on the go and our mobile interface (which is connected to the dialogue system proper via UMTS or WiFi) is an example of a typical user interaction sequence, which combines the *Google Maps Local Search*, GeoNames *findNearbyWikipedia* search, and GeoNames *findNearbyWeather* search.

1 **U:** *"Where can I find good restaurants (in Berlin Mitte)?"* (If no named entity for location is mentioned, we use the built-in GPS locator. Good restaurants are selected according to a heuristics that takes user rankings into account.)
2 **S:** Shows a map and list of restaurants. See Figure 3, left.
3 **U:** *"Give me more information about the second one / the Fernsehturm."*
4 **S:** Shows corresponding detailed pieces of information according to referral restaurant data record. See Figure 3, right.
5 **U:** *"What else can I visit here?"*
6 **S:** Provides Wikipedia results of nearby POIs and the walking route. See Figure 4, left.
7 **U:** *"What do you know about this POI?"*
8 **S:** Provides the Wikipedia article (and current weather information for the respective district). See Figure 4, right.

Please note that the user input is paraphrased for illustration purposes. In a multimodal dialogue system, the user should be able to switch between spoken, written, or clicked input. This also means that a summary of the Wikipedia article could be synthesised. However, the speech option is not necessary in this scenario and does not seem to provide a good trade-off between system complexity and user experience since additional third-party components for ASR, NLU, and TTS would become necessary.

Figure 5 depicts the described dialogue and interaction sequence. In (1), the user searches for restaurants by typing an appropriate query into the text field and pressing the search button. Then the result list (2) pops up. The locations of the restaurants (the respective POIs) are shown on the map in the background. The user selects the entry about the restaurant "Fernsehturm" and presses the information button (indicated by *i-icon*) at the top of the screen. The resulting detail screen is shown in (3). After reading the information, the user intends to have dinner there, but he wants to combine it with a little sightseeing nearby. So, the user presses the "What's nearby?" button (indicated by the *circle-icon*). An appropriate result list sorted by the distance from the restaurant appears in which the user selects in (4) the last entry ("Rotes Rathaus"). By pressing the information button again, details to the selected sight including current weather information and a short Wikipedia summary are presented (5). Furthermore, the route from the restaurant to the sight is displayed on the map. The user can minimise the previous view in order to inspect the proposed route (6).

IV. Conclusions and Future Work

Multimodal dialogue application frameworks can be used to implement specific mobile applications and dynamic HTTP-based REST services. These infrastructures can overcome the technical limitations imposed by current mobile device hardware and software. In addition, new services of independent providers can be added easily.

We created a custom HTTP/REST Meta Web Service for mobile location-based scenarios and explained how this

 ISBN: 978-1-61208-100-7

Figure 5. Mobile GUI Screenshots

service integrates into a multimodal dialogue framework. In addition, we provided a real-world application scenario and explained our generic dialogue framework and a specific implementation of a new Meta Web Service.

With the new Meta Web Service, we can provide meaningful information for travellers or tourists. Currently, we are restricted by the information from the Google Maps and GeoNames services. We think that it will soon be possible to find new REST services to be integrated.

The keyword/term based input possibilities of Google Maps is quite suitable for dynamic POI searches. However, the POI classes (e.g., hotel, pharmacy) are not documented so that a user has to search for those in a trial-and-error fashion. This relevance-feedback interaction style could bring speech-based interaction into the fore if the speech grammar accepts the relevant portion of open-domain POI classes.

A distributed dialogue infrastructure overcomes the technical limitations imposed by current mobile device hardware and software (which severely hinders the potential benefits of mobile speech-based applications and, e.g., augmented reality applications such as 3D egocentric views of the user's surrounding [28]). We hypothesise that these new interaction and visualisation techniques, which demand for distributed applications, will become very beneficial in the future.

 ISBN: 978-1-61208-100-7

V. ACKNOWLEDGMENTS

This research has been supported by the THESEUS Research Programme in the Core Technology Cluster WP4 and the TEXO use case, which was funded by the German Federal Ministry of Economy and Technology under the promotional reference “01MQ07012“. The authors take the responsibility for the contents.

REFERENCES

[1] R. Pieraccini and J. Huerta, “Where do we go from here? research and commercial spoken dialog systems.” in *Proc. 6th SIGDdial Workshop on Discourse and Dialogue*, 2005, pp. 1–10.

[2] J. Allen, D. Byron, M. Dzikovska, G. Ferguson, L. Galescu, and A. Stent, “An Architecture for a Generic Dialogue Shell,” *Natural Language Engineering*, vol. 6, no. 3, pp. 1–16, 2000.

[3] S. Seneff, R. Lau, and J. Polifroni, “Organization, Communication, and Control in the Galaxy-II Conversational System,” in *Proc. of Eurospeech*, 1999, pp. 1271–1274.

[4] D. Sonntag, *Ontologies and Adaptivity in Dialogue for Question Answering*. AKA and IOS Press, Heidelberg, 2010.

[5] W. Wahlster, “SmartKom: Symmetric Multimodality in an Adaptive and Reusable Dialogue Shell,” in *Proc. Human Computer Interaction Status Conf.*, R. Krahl and D. Günther, Eds. DLR, 2003, pp. 47–62.

[6] N. Reithinger, D. Fedeler, A. Kumar, C. Lauer, E. Pecourt, and L. Romary, “MIAMM - A Multimodal Dialogue System Using Haptics,” in *Advances in Natural Multimodal Dialogue Systems*. Springer, 2005.

[7] R. T. Fielding, “Architectural Styles and the Design of Network-based Software Architectures,” Ph.D. dissertation, University of California, Irvine, 2000.

[8] M. Pielot, N. Henze, and S. Boll, “Supporting map-based wayfinding with tactile cues,” in *Proc. 11th Int’l Conf. Human-Computer Interaction with Mobile Devices and Services (MobileHCI)*. ACM, 2009, pp. 170–179.

[9] D. Rowland, M. Flintham, L. Oppermann, J. Marshall, A. Chamberlain, B. Koleva, S. Benford, and C. Perez, “Ubikequitous computing: designing interactive experiences for cyclists,” in *Proc. 11th Int’l Conf. Human-Computer Interaction with Mobile Devices and Services (MobileHCI)*. ACM, 2009, pp. 151–159.

[10] D. Arter, G. Buchanan, M. Jones, and R. Harper, “Incidental information and mobile search,” in *Proc. 9th Int’l Conf. Human-Computer Interaction with Mobile Devices and Services (MobileHCI)*. ACM, 2007, pp. 129–144.

[11] K. Church, J. Neumann, M. Cherubini, and N. Oliver, “SocialSearchBrowser: a novel mobile search and information discovery tool,” in *Proc. 14th Int’l Conf. Intelligent User Interfaces (IUI)*. ACM, 2010, pp. 101–110.

[12] C. Riva and M. Laitkorpi, “Designing Web-Based Mobile Services with REST,” in *Service-Oriented Computing - ICSOC 2007 Int’l Workshops, Revised Selected Papers*. Springer, 2009, pp. 439–450.

[13] D. Sonntag, R. Engel, G. Herzog, A. Pfalzgraf, N. Pfleger, M. Romanelli, and N. Reithinger, *Artifical Intelligence for Human Computing*. Springer, 2007, ch. SmartWeb Handheld—Multimodal Interaction with Ontological Knowledge Bases and Semantic Web Services, pp. 272–295.

[14] D. Porta, D. Sonntag, and R. Neßelrath, “A Multimodal Mobile B2B Dialogue Interface on the iPhone,” in *Proc. 4th Workshop on Speech in Mobile and Pervasive Environments (SiMPE)*, 2009.

[15] S. R. Duc Thanh Tran, Haofen Wang and P. Cimiano, “Top-k exploration of query graph candidates for efficient keyword search on rdf,” in *Proc. 25th Int’l Conf. Data Engineering (ICDE)*, 2009.

[16] W3C. (2010, Jul.) Web Service Description Language. [Online]. Available: http://www.w3.org/TR/wsdl20/

[17] Google. (2010, Jul.) Google APIs. [Online]. Available: http://code.google.com/intl/en-EN/more/

[18] Yahoo. (2010, Jul.) Yahoo Pipes. [Online]. Available: http://pipes.yahoo.com/

[19] OASIS. (2010, Jul.) Business Process Execution Language. [Online]. Available: http://docs.oasis-open.org/wsbpel/2.0/OS/wsbpel-v2.0-OS.html

[20] M. AbuJarour, M. Craculeac, F. Menge, T. Vogel, and J.-F. Schwarz, “Posr: A Comprehensive System for Aggregating and Using Web Services,” *IEEE Congress on Services*, vol. 0, pp. 139–146, 2009.

[21] Mapki. (2010, Jul.). [Online]. Available: http://mapki.com/

[22] OGC. (2010, Jul.) KML. [Online]. Available: http://www.opengeospatial.org/standards/kml/

[23] GeoNames. (2010, Jul.) WebServices overview. [Online]. Available: http://www.geonames.org/export/ws-overview.html

[24] D. Porta, “A Novel, Community-Enabled Mobile Information System for Hikers,” in *Proc. 2nd Int’l Conf. Mobile Ubiquitous Computing, Systems, Services and Technologies (UBICOMM)*. IEEE Computer Society, 2008, pp. 438–444.

[25] LaszloSystems. (2010, Jul.) OpenLaszlo. [Online]. Available: http://www.openlaszlo.org/

[26] M. Annett and E. Stroulia, “Building highly-interactive, data-intensive, REST applications: the Invenio experience,” in *Proc. Conf. Advanced Studies on Collaborative Research (CASCON)*. ACM, 2008, pp. 192–206.

[27] A. Gruenstein, I. McGraw, and I. Badr, “The WAMI toolkit for developing, deploying, and evaluating web-accessible multimodal interfaces,” in *Proc. 10th Int’l Conf. Multimodal Interfaces (IMCI)*. ACM, 2008, pp. 141–148.

[28] Y. Tokusho and S. Feiner, “Prototyping an Outdoor Mobile Augmented Reality Street View Application,” in *Int’l Symp. Mixed and Augmented Reality (ISMAR)*, 2009.

 ISBN: 978-1-61208-100-7

Evaluation of the Wireless Network used by a Tour Guide in a Cultural Environment

Ricardo Tesoriero, José A. Gallud, María D. Lozano, Víctor M. R. Penichet, Habib M. Fardoun
Department of Information Systems
University of Castilla-La Mancha
Albacete, Spain
[ricardo.tesoriero, jose.gallud, maria.lozano, vpenichet, habib.moussa]@uclm.es

***Abstract*—Wireless tour guides are mobile applications running on a PDA that are very popular among cultural environments, such as museums and art exhibitions. The Human-Computer Interaction quality derived from the use of these guides heavily depends on the wireless network used to access the information of the artefacts in the cultural environment. This work presents a procedure to evaluate the Wi-Fi network performance used by a tour guide system. The procedure is applied to a real museum in Spain. The procedure is based on a set of observation and measurements performed in measurement points distributed across the cultural environment. It also presents how to analyse the coverage of the wireless network, the set of environmental factors that may degrade the performance of the network and the effectiveness of the access point selection algorithm used by the mobile device. As result of the evaluation, we have identified a set of weak points in the network development. However, they were not relevant enough to alter the actual network deployment.**

Wi-Fi;Human-Computer Interaction; Mobile applications

I. INTRODUCTION

Museum guides have evolved through the time. Traditional approaches, such as text panels next to art pieces and triptychs, have serious limitations. For instance, the physical space that is available to support multi-language information. Besides, the maintenance costs on this media are high due to information reprinting in time-limited exhibitions.

In order to solve these problems, alternatives such as audio guides, supported by cassette players or solid memory devices, or radio frequency technologies emerged providing visitors with individual access to information. Main disadvantage of these approaches is the limitation to provide audio information only.

Although Web technologies offered new opportunities to open up the walls of the museum to the world [1], the information was not offered in-situ.

Thus, the Personal Digital Assistant (PDA) emerged as a powerful tool to provide multimedia information in-situ, providing applications that served as multimedia guides for cultural environments visitors. A comparative evaluation of different platforms for augmenting museums and art galleries where the PDA proved to be a successful information provider for visitors is presented in [2]. Additionally, according to [3], the PDA has also been accepted as a good approach to face this problem.

Although there are tour guides for indoor and outdoor environments, such as the exposed in [4], we will focus on guides for indoor environments.

The infrastructure to develop tour guides for indoor environments varies according to how the information is retrieved.

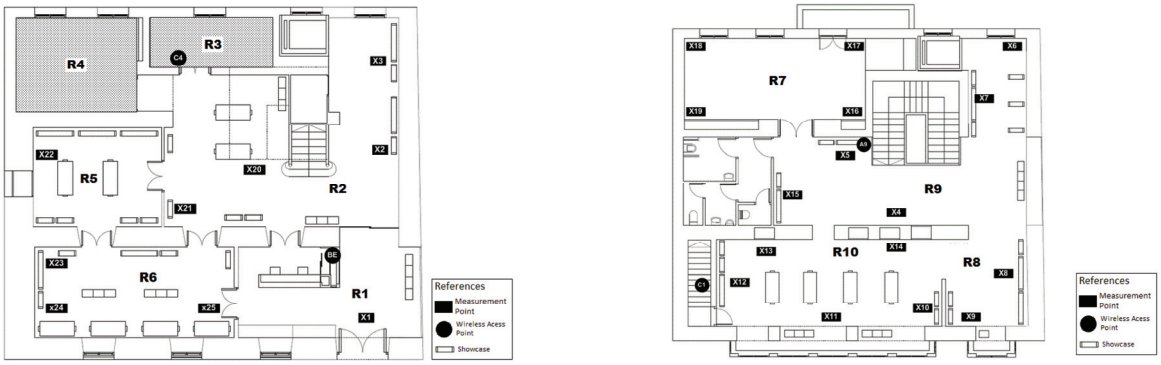
On the one hand, the standalone approach proposes the information retrieval from the PDA device memory [5, 6]. On the other hand, the client-server approach proposes storing information on a central server and the retrieval is achieved from the clients (PDAs) through a wireless network connection [7].

The standalone approach has some advantages such as: (a) the lack of networking infrastructure, and (b) the server storing the information to be exposed to visitors. However, the client-server approach has also some valuable advantages compared to the standalone approach. In a client-server approach, the cultural environment information is centralized; therefore when it changes, it is automatically propagated to the clients. Besides, the client-server does not restrict the use of the guide to those devices that are provided by the cultural environment because it can be easily deployed on visitor devices (it only requires a tiny the client application to be downloaded). Under this scenario, the communication network becomes a key factor of the tour guide application because from the human-computer interaction perspective, the user experience depends on the capabilities of the wireless network to retrieve multimedia information.

This paper proposes an evaluation procedure for wireless communication networks employed to deliver information from the tour guide content server to the tour guide clients within the cultural environment.

The evaluation procedure is performed on a tour guide deployed in a real museum the Cutlery Museum of Albacete (MCA) in Albacete, Spain. The goal of the system is to provide visitors with multimedia information about the knives exposed in the building through a PDA device running the tour guide application.

 ISBN: 978-1-61208-100-7



(a) (b)

Figure 1. The physical environment: (a) the ground floor (b) the first floor.

The evaluation procedure is explained as follows. In Section II, we expose how this work is related to other works, and where the evaluation procedure can be applied. In Section III, we describe the museum physical environment. Section IV presents the software and hardware architectures of the system. In Section V, we expose the measurement process and results. In Section VI, we analyse the measurements, and finally, in Section VII, we expose the conclusions and future work.

II. RELATED WORK

The evaluation of the wireless network is close related to the environment in which it is deployed. In this case, we expose an indoor environment represented by a cultural environment in which a wireless guide is deployed.

As we have mentioned in Section I, the wireless network that was evaluated was the Cutlery Museum of Albacete (MCA) in Albacete, Spain [7].

However, it is not the only wireless guide for cultural environments. There are some other interesting approaches that exploit this technology; for instance, mi-Guide [8] or the location aware system exposed in [9].

Thus, this paper proposes a guideline that can be used to evaluate the deployment of wireless networks following a defined procedure that takes into account the physical environment, the measurement process and the data analysis.

III. THE PHYSICAL ENVIRONMENT

The first step of the evaluation procedure is the analysis of the physical environment.

In the MCA, the physical environment where the system was deployed is defined by a two floor building. The ground floor and the first floor plans are depicted in Fig. 1 (a) and Fig. 1 (b).

The surface of each floor is about 200 m^2, external walls are 40 cm width, internal walls are about 25 cm width and the floor is tiled.

Floor rooms are identified by the "R" prefix. Rooms R1 to R6 belong to the ground floor, and rooms R7 to R11 belong to the first floor. The walls of the rooms are bare, except for R7 (in the first floor) that has wood-panelled walls.

Rooms also contain showcases where pieces are exposed. They are all wooden-made, but those that are in R10 that are made of stainless steel.

The Fig. 1 (a) and the Fig. 1 (b) show the distribution of showcases, wireless access points (C1, C4, A9 and B3) and measurement points (X1 - X25) through the building.

Note that the ground floor is connected to the first floor by a stainless steel elevator and a marble stair.

IV. THE WIRELESS TOUR GUIDE SYSTEM

Before introducing the measurement process, we expose in this section the hardware and software architecture to the system. This information is relevant to the evaluation process in order to specify the type of systems we are dealing with.

A. System architecture

The system is based on the traditional client-server architecture depicted in Fig. 2.

The server side of the application is composed by the Web server and the Database server. Each server is defined as a two-node cluster to improve system reliability.

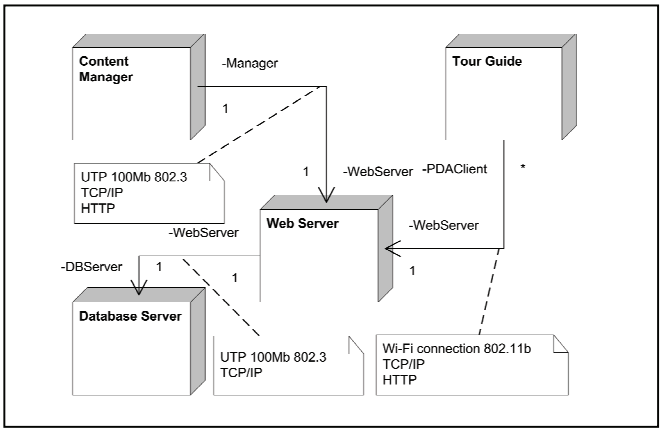
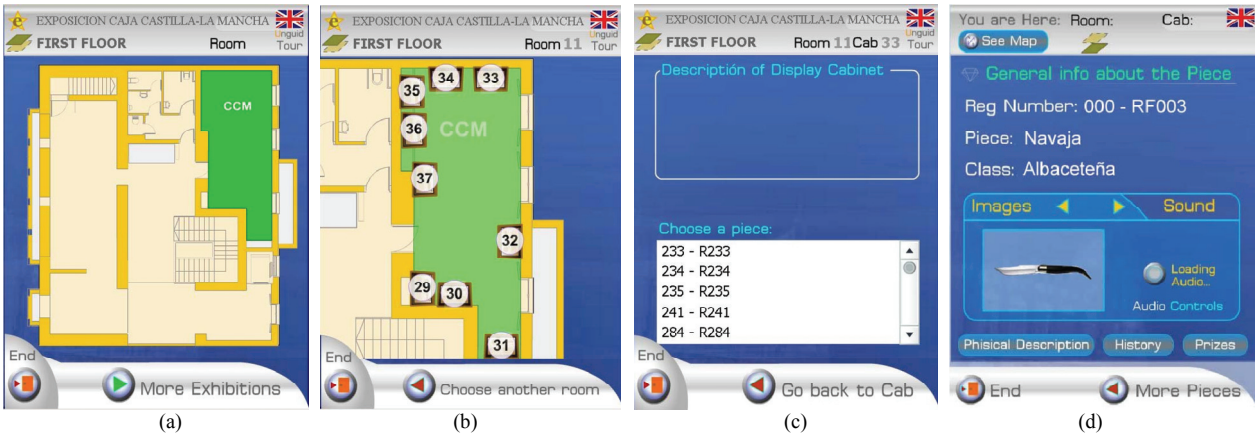


Figure 2. The wireless tour guide architecture



(a) (b) (c) (d)

Figure 3. The client user interface (UI): (a) the floor UI (b) the room UI (c) the showcase UI (d) the piece UI.

Server machines are equipped with Pentium IV 2GHz processors with 512 MB of RAM Memory and two 100 Mbps Ethernet interfaces.

One of them is used for data synchronization between cluster nodes, and the other one is used to connect the cluster to the information system.

While the communication between database and Web servers is wired with UTP cable for 100Mb, the communication between the Web server and mobile clients is conducted by a wireless network (802.11b). Transport and communication protocols are TCP/IP in both cases.

In order to cover the whole building, we have installed five wireless access points that supports 802.11b at 2.4 GHz. Besides, to reduce the network traffic and the processing overhead in the PDA client, no encryption was enabled in the system.

The client side of the application is composed by two types of clients: the Content Manager and the Tour Guide. The Content Manager is designed as a "What You See Is What You Get" (WYSWYG) application to manage the information of to be provided to visitors. On the other hand, the Tour Guide is designed as a PDA based application that runs on a PDA.

Each PDA is equipped with a 400 MHz processor, 64 Mb of SDRAM, 64 Mb of ROM flash memory, the QVGA TFT colour screen is 3.5 inches long and has a 320 x 240 resolution on 65,536 colours. Communication protocol is limited to 802.11b.

B. *The Client Application*

This section exposes the client application to provide readers with a conceptual idea of the wireless network load.

A full description of the application is exposed in [7] and a usability evaluation following the CIF (Common Industry Format for Usability Reports) standard defined by the ISO/IEC DTR 9126-4 is presented in [10].

The Adobe Flash technology was used to provide visitors with a high quality user interface. This interface provides visitors with the ability to explore the museum in several ways: (a) guided tour, (b) recommended routes, (c) through the finder and (d) unguided tour.

Independently of the museum exploration mode, the application provides information of floors, rooms, showcases and pieces. The Fig. 3 shows tour guide user interface.

All screens are stored locally by the device. However, the content showcase and piece related resource are retrieved from the Web server. The key screen, from the network information load point of view is the piece resource viewer. Through this interface the user is able to retrieve text, images, audio and video information.

C. *The Characteristics of theInformation*

The image file format used by the application resurces is JPG, the resolution is 800 px x 600 px and the file size varies from 5.72 KB up to 81.1KB (avg. 43.41 KB).

The audio file format used is MP3 at 22 KHz and 32 Kbps of bit rate; file size varies from 129 KB up to 370 KB (avg. 249.5 KB). The video video file format used is WMV, the video resolution is 352 x 288 at 368 kbps; the video and audio codec are WMV3 and two channel WMA2 at 22Khz and 32 kbps of bit rate respectively; file size is about 5 Mb.

V. THE MEASUREMENT PROCEDURE

Once, we have analysed the environment and system characteristics, we proceed with the measurement procedure. Thus, this section exposes the measurement procedure that was employed to carry out the evaluation of the MCA wireless network to be used as a guide for further procedures.

The software employed was the PDA's default software to show access point information. It provides the signal strength (S), the signal noise (N) and the S/N ratio for each access point that is at range. It also provides the access point the device is connected to.

The measurement process was carried out in two phases achieving two independent measurements following the same route. The route is defined as a sequence of points where the measurement is performed. The order of the sequence is defined by the order of the X nodes (see Fig. 1 and Fig. 2). Measurement results are exposed in Table I and Table II.

 ISBN: 978-1-61208-100-7

TABLE I. AVERANGE ACCESS POINT COVERAGE

M. Point	C4		A9		B3		C1	
	S^a	N^a	S^a	N^a	S^a	N^a	S^a	N^a
X1	-69	-93	-76	-92	-50	-88	-77	-92
X2	-69	-91	-56	-92	-61	-91	-84	-93
X3	-76	-93	-60	-90	-75	-90	-83	-91
X4	-80	-92	-46	-92	-69	-91	-53	-92
X5	-74	-92	-41	-91	-80	-91	-59	-92
X6	-88	-92	-60	-92	-79	-91	-70	-90
X7	-78	-92	-59	-92	-76	-88	-65	-92
X8	-89	-93	-50	-91	-62	-91	-55	-93
X9	-87	-91	-51	-92	-61	-91	-58	-92
X10	-86	-92	-56	-92	-62	-92	-53	-93
X11	-90	-92	-58	-92	-67	-91	-51	-91
X12	-88	-92	-53	-91	-70	-91	-44	-92
X13	0	0	-57	-88	-72	-91	-52	-90
X14	-85	-92	-50	-92	-63	-91	-53	-91
X16	-79	-91	-55	-90	-82	-91	-68	-91
X17	-75	-92	-56	-91	-86	-90	-72	-93
X18	-82	-93	-61	-88	-85	-87	-74	-92
X19	-78	-91	-63	-90	-88	-91	-74	-92
X20	-56	-92	-61	-91	-68	-92	-80	-92
X21	-64	-92	-56	-89	-77	-90	-84	-92
X22	-69	-92	-59	-92	-78	-92	-82	-92
X23	-78	-92	-83	-93	-53	-91	-73	-91
X24	-80	-92	-85	-90	-62	-92	-72	-91
X25	-84	-92	-76	-91	-48	-91	-78	-90

a. dB unit used

TABLE II. MOBILE DEVICE CONNECTION QUALITY

M. Point	Phase I			Phase II			Best
	A.P.	S^a	N^a	*A.P.*	S^a	N^a	S / N^b
X1	BE	-50	-88	BE	-50	-88	BE
X2	C4	-69	-91	BE	-61	-91	A9
X3	A9	-60	-0	A9	-60	-90	A9
X4	A9	-46	-92	A9	-46	-92	A9
X5	A9	-41	-91	C1	-59	-92	A9
X6	A9	-60	-92	C1	-70	-90	A9
X7	A9	-59	-92	C1	-65	-92	A9
X8	A9	-50	-91	C1	-55	-93	A9
X9	A9	-51	-9	C1	-58	-92	A9
X10	A9	-56	-92	C1	-53	-93	C1
X11	A9	-58	-92	C1	-51	-91	C1
X12	A9	-53	-91	C1	-44	-92	C1
X13	A9	-57	-88	C1	-52	-90	C1
X14	A9	-50	-92	C1	-53	-91	A9
X16	A9	-55	-90	C1	-68	-91	A9
X17	A9	-56	-91	C1	-72	-93	A9
X18	A9	-61	-88	C1	-74	-92	A9
X19	A9	-63	-90	A9	-63	-90	A9
X20	A9	-61	-91	A9	-61	-91	C4
X21	A9	-56	-89	A9	-56	-89	A9
X22	A9	-59	-92	A9	-59	-92	A9
X23	A9	-83	-93	BE	-53	-91	BE
X24	A9	-58	-90	BE	-62	-92	BE
X25	A9	-76	-91	BE	-48	-91	BE

a. dB unit used, b. dBW unit used

On the one hand, Table I shows the average access point coverage defined by the signal and noise strength received by the mobile device from all access points. On the other hand, Table II shows mobile device connection quality exposing the access point selected by the mobile device on each phase and the best option according to the measurements on Table I.

VI. THE MEASUREMENT ANALYSIS

This section analyses the measurements collected from Section V They are used to perform the analysis of the following set of features of the wireless network: (a) the signal coverage in the building, (b) the algorithm employed to choose an access point and (c) the environmental factors that affect the signal propagation.

A. The Signal Coverage in the Building

In order to evaluate the signal coverage in the building we use the signal-noise ratio as the reference parameter.

Thus, the equation (1) defines the f_r function that takes a measurement point x_i and an access point p_j as parameters to return the signal/noise ratio from p_j at x_i. The M constant defines the total amount of measurement points and the A constant defines the total amount of access points.

$$f_r (x_i, p_j), i \in [1 \ldots M] \wedge j \in [1 \ldots A] \quad (1)$$

The R_{MAX} and R_{MIN} values, exposed by the equations defined in (2), represent the maximum and the minimum signal/noise ratio that have been measured during both measurement phases.

 ISBN: 978-1-61208-100-7

$$R_{MAX} = f_r(x_k, p_l), \forall i, j\ (f_r(x_k, p_l) \geq f_r(x_i, p_j))$$

$$R_{MIN} = f_r(x_k, p_l), \forall i, j\ (f_r(x_k, p_l) \leq f_r(x_i, p_j)) \quad (2)$$

$$\forall i, k \in [1\ldots M] \wedge \forall j, l \in [1\ldots A]$$

Finally, (3) defines $g_r(f_r(x_i, p_j))$ that returns 1 if p_j covers x_i, or 0 otherwise.

$$r = f_r(x_i, p_j);\ RLIM = (R_{MAX} - R_{MIN})/2$$

$$g_r(r) = 1 \Leftrightarrow r \in [R_{LIM}\ldots R_{MAX}] \quad (3)$$

$$g_r(r) = 0 \Leftrightarrow r \notin [R_{LIM}\ldots R_{MAX}]$$

The application of the equations on measurement results is represented by the Fig. 4 (a). It shows the signal/noise ratio from the "best" access point on each measurement point. It also depicts the R_{MAX}, R_{MIN} and the R_{LIM} values. The main reason for being below R_{LIM} may be related to: (a) interference because of environmental conditions or (b) the algorithm used to select the access point the device connects to. To solve this question, we analyse these parameters in next subsections.

From Fig. 4 (a), we may also conclude that although in both phases the coverage was not complete, the wireless network is able to fulfil this requirement because if the best access points would have been selected the 100 % of measurements points would have been covered.

B. Environmental factors

This section analyses the environmental factors, such as interference generators, that may degrade the link quality. The first parameter to analyse is the noise as reference parameter.

The problem will be discussed at two levels. On the one hand, the first level studies the environmental noise in the building. On the other hand, the second level studies the set of particular situations that we have revealed on previous section.

1) The Environmental Noise. In order to study the environmental noise, we base our analysis in the noise level through the building. These levels are depicted in Fig. 4 (b). Thus, the equation (4) defines the function that takes a measurement point x_i and an access point p_j as parameters to return the signal noise at x_i from the p_j. Once noise function was defined, we apply (5) to obtain the X_m= -91.32 dB and the S_x = 1.19 dB. These values show that noise levels are stable.

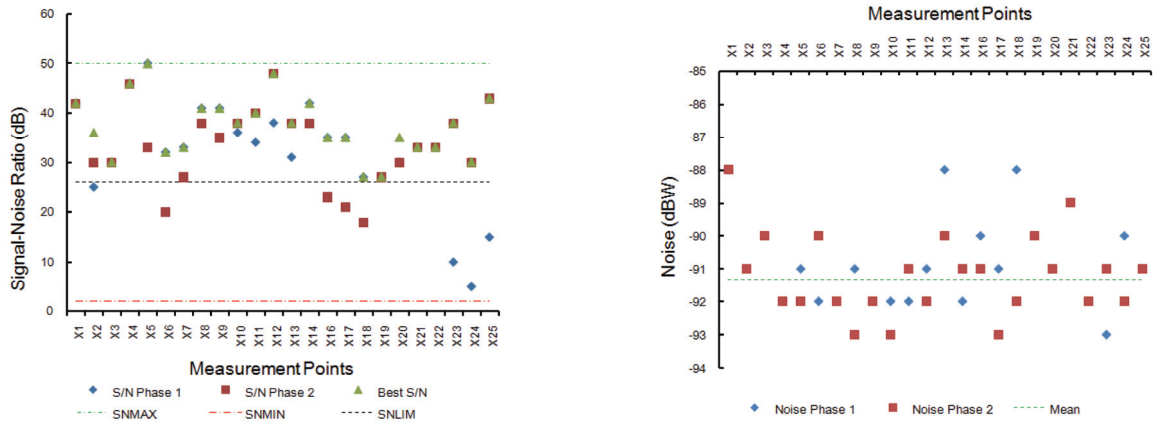
$$f_n(x_i, p_j), i \in [1\ldots M] \wedge j \in [1\ldots A] \quad (4)$$

$$X_m = (\Sigma f_n(x_i, p_j))/N \wedge S_x = (\Sigma f_n(x_i, p_j)) - X_m)^2/N;\ N=i \cdot j \quad (5)$$

Some Particular situations. Once we have analysed the environment as a whole, we continue with the analysis on problematic measurement points. First, we will analyse X2, X23, X24 and X25 from the first phase. Focusing on X2 in Fig. 1 (a) we can see that it is in the middle of two access points (C4 and BE). The most suitable is BE (see Table II). However, the C4 was chosen, even when BE was previously chosen (X1). The most reasonable explanation is the visitor orientation, coming from X1 the device may have pointed to C4 instead of BE. However, there is an argument to be discussed before closing this case. The question is related to Phase II access point selection (BE) for the same measurement point. The explanation is related to the access point selection algorithm that according to our observations follows this procedure: an access point is selected only if no access point was selected or the connection to actual access point is lost. Thus, on the one hand, in the first phase the connection with BE was lost while the visitor was pointing to C4. On the other hand, in the second phase, the connection was not lost, so no selection was made at all. The group composed by X23, X24 and X25 measurement points is attached to A9 instead of BE (C1 is discarded because these measurement points are in the ground floor and C1 is between the ceiling and the roof at first floor). In contrast to the previous case, it seems not to be the orientation of the visitor that affects the selection of the optimal access point (BE). Instead, it may be related to the access point selection algorithm described on previous paragraph. Once first phase analysis was performed, we analyse the X6, X16, X17 and X18 measurement points of second phase. The measurement point X6 is affected by the elevator that is made up stainless steel. The elevator was up when the measurement was taken, fading the signal. Another factor that has affected X6 is the change of the access point used by the mobile device from A9 to C4. It should have been provoked by the way up from ground floor to first floor through the stairs. Finally, another consequence of the change of access point mobile device from A9 to C4 is the low measurement at X16, X17 and X18.

2) The Access Point Selection Algorithm. From the analysis performed on both phase measurements, we have identified the following problems: (a) visitor orientation, (b) physical interference (i.e. elevator) and (c) the access point selection algorithm (change of access point connection). It is really difficult to control the first two problems, because they are not directly related to hardware or software. However, the access point selection algorithm deserves an extra analysis to propose future improvements on it. The results of the analisys expose that on the first phase, the 54 % of the access point selected by the device is coincident to the best possible. On the second phase of measurements, the 62 % of the access point selected by the device is coincident to the best possible. It proves that the access point selection algorithm should be improved in order to leverage the wireless network performance.

 ISBN: 978-1-61208-100-7



(a) (b)

Figure 4. Example of a TWO-COLUMN figure caption: (a) this is the format for referencing parts of a figure.

According to the study of building coverage we have performed, a new algorithm may leverage the signal-noise ratio to reach an optimal 100 %. Therefore, the performance may reach an average improvement of 41.67 %.

VII. CONCLUSIONS AND FUTURE WORK

This article proposes a procedure to evaluate the capabilities of wireless networks used by tour guides in cultural environments. It also applies the procedure to a real product deployed at the MCA in Albacete, Spain. It starts evaluating the signal coverage in the building and the environmental interference. As result of the analysis of these two parameters, the mobile device access point selection algorithm is evaluated conducting to the conclusion that an average improvement of 41.67 % may be achieved in the access point selection if the selection algorithm properly modified.

The future work is focused on three main research lines. The first one is related to how to improve the algorithm. We are actually exploring different ways to do it that goes from pooling to the application of prediction algorithms. On the second research line we are trying to use new mobile device features such as accelerometers and gyroscopes that could be used to acquire information from visitor movements. And finally, we are doing some research on how new technologies affect the wireless network capabilities. An example of new technology applied to this field is the use of RFID technology in both variants: passive and active, as presented in [11].

ACKNOWLEDGMENT

This research paper has been funded by the CDTI, project CENIT-2008-1019 and project CICYT TIN2008-06596-C02-01.

REFERENCES

[1] R. Jackson, M. Bazley, D. Patten, and M. King, “Using the web to change the relation between a museum and its users,” in Proc. of the 2nd International Conference Museums and the Web, D. Bearman and J. Trant, Eds., 1998.

[2] C. Baber, H. Bristow, S. lee Cheng, A. Hedley, Y. Kuriyama, M. Lien, J. Pollard, and P. Sorrell, “Augmenting museums and art galleries,” in Proc. of the 8th Conference on human-computer interaction. INTERACT’01, 2001, pp. 439–447.

[3] L. Ciolfi and L. J. Bannon, “Designing interactive museum exhibits: Enhancing visitor curiosity through augmented artifacts,” in Proc. of the 11th European Conference on Cognitive Ergonomics, 2003, pp. 311–317.

[4] H.W. Bristow, C. Baber, J. Cross, S. I.Woolley, and M. Jones, “Minimal interaction for mobile tourism computers,” in Proc. of the 4th International Symposium on Human Computer Interaction with Mobile Devices. MobileHCI. The Workshop “Mobile Tourism Support”, 2002.

[5] C. Ciavarella and F. Paternò, “Design criteria for locationaware, indoor, PDA applications,” in Human-Computer Interaction with Mobile Devices and Services, ser. Lecture Notes in Computer Science, L. Chittaro, Ed., vol. 2795. Springer, 2003, pp. 131–144.

[6] C. Ciavarella and F. Paternò, “The design of a handheld, location-aware guide for indoor environments,” Personal and Ubiquitous Computing, vol. 8, no. 2, 2004, pp. 82–91.

[7] J. A. Gallud, V. M. R. Penichet, L. Argandoña, P. González, and J. A. García, “Digital museums: a multi-technological approach,” in Proc. of the HCI-International Conference 2005. Las Vegas, USA: Lawrence Erlbaum Associates, 2005.

[8] N. Linge, D. Parsons, D. Bates, R. Holgate, P. Webb, D. Hay, and D. Ward, “mi-Guide: A wireless context driven information system for museum visitors,” in The 6th International Workshop on Wireless Information Systems, 2007, pp. 43–53.

[9] C.-Y. Tsai, S.-Y. Chou, and S.-W. Lin, “Location-aware tour guide systems in museums,” Scientific Research and Essays, no. 8, April 2010, pp. 714–720.

[10] R. Tesoriero, M. D. Lozano, J. A. Gallud, and V. M. R. Penichet, “Evaluating the users’ experience of a pda-based software applied in art museums,” in Proc. of the 3rd International Conference on Web Information Systems and Technologies. Barcelona, Spain: INSTICC, March 2007, pp. 351–358.

[11] R. Tesoriero, M. D. Lozano, J. A. Gallud, and V. M. R. Penichet, “Using active and passive rfid technology to support indoor location-aware systems,” IEEE Trans. Consum. Electron., no. 2, May 2008, pp. 578–583.

 ISBN: 978-1-61208-100-7

Mobile Services and Applications: Towards a Balanced Adoption Model

Krassie Petrova, Stephen G. MacDonell
School of Computing and Mathematical Sciences
Auckland University of Technology
New Zealand
kpetrova@aut.ac.nz, smacdone@aut.ac.nz

Abstract— **This paper synthesizes prior research to develop a novel model for the study of the adoption of mobile business services and applications incorporating a demand and supply perspective. The model complements and extends existing models while also leveraging data from industry reports; in particular, it focuses on the interrelationships between participants in the mobile services value chain and the impact of these interrelationships on the adoption of new services in a competitive and technology-saturated service market. There has been to date limited research reported that has considered the dynamics of the interrelationships between customers and (layers of) multiple service providers as a factor in the adoption and acceptance process; the proposed model addresses this gap and advocates the use of a combination of design science and service science methodologies. It is concluded that not mobility *per se* but the way mobility is used to create value plays a significant role as an adoption driver, and that the quality of the service and its relevance to personal or business lifestyle are the most important decision making factors. It is also asserted that while innovative mobile services (i.e., services that are not already offered using a different technology) may be compelling if they meet lifestyle needs, mobile services replacing or complementing existing ones will be favored by customers only if their quality is exceptional and motivates 'switching' to the mobile service.**

Keywords-mobile services; adoption; mobile commerce; quality of service expectations; lifestyle requirements; mCommerce.

I. INTRODUCTION

Business transactions between participants (e.g., customers, businesses) enabled by mobile data networks are commonly referred to as mobile commerce (mCommerce) via a range of related mobile services and applications [1-2]. A specific characteristic of mCommerce is its potential to support customer mobility by offering services that dynamically adjust to be available at the location in which the mobile customer operates [3]. In addition, mCommerce transactions may be facilitated by a specific form of payment known as mobile payment (mPayment) [4]. The definition of mCommerce adopted in this study is derived from [5]: 'A value-added service that enables mobile customers to conduct reliable and secure transactions through specifically-designed mobile applications'. Further, mobile business (mBusiness) services expand mCommerce to include not only transactions between participants but activities such as servicing customers, and collaborating and conducting mobile transactions with business partners based on an appropriate business model (adapted from [6], p. 685). Finally, with respect to the type of interaction between participants, most mCommerce transactions can be classified applying the categories used to classify electronic commerce (eCommerce) transactions; however at present B2C (business-to-customer) mobile transactions prevail [1].

A number of general frameworks and models for the study of mobile services and their adoption have been proposed in prior work drawing on eCommerce adoption studies and often including variables such as usefulness, ease of use, and usability [7-11]. Additional specific constructs such as customer mobility [2]; location awareness [12-13], trust [14], service cost [15], and perceived value proposition [16-17] have also been considered. A range of country-specific adoption barriers have been identified from customer, technology, company and business perspectives [18-20]. A global view of the effect of the legislative environment (government intervention) on location-aware services has also been provided in [21].

While empirical studies have been able to identify some of the factors affecting customer decisions, the dynamics of the processes of meeting customer needs and preferences (i.e., mobile business service demand) by the gamut of industry players (i.e., mobile business service supply) has not been studied in depth. With customers becoming both better informed and more experienced as technology users, it can be expected that additional factors may emerge from a study of the adoption processes from multiple perspectives and in a contemporary context including customer perceptions about mobile business service value [22-23].

It has previously been suggested that in order to explain mCommerce adoption processes both the supply and the demand side may need to be included in a comprehensive adoption framework [23], and that the relationship between customer and service supplier is one of the four main aspects of a mobile service [24]. The objective of the study presented here is to derive and extend an explanatory model capturing the relationships between supply and demand factors that can be used further to investigate how customer lifestyle requirements and expectations about the quality of a mobile business service affect market demand for these services and contribute to actual mobile service use.

 ISBN: 978-1-61208-100-7

The rest of the paper is organized as follows: The next section reviews the relevant literature and provides background information. The section following describes the proposed model including its variables and the relationships between them. The last section discusses the model from the perspective of further research and provides a conclusion.

II. MOBILE BUSINESS SERVICES AND APPLICATIONS

General mobile business-to-customer services (mobile business services) and enabling mobile services (e.g., mPayment) are delivered to customers as a result of the business interactions within the mobile business value chain [16][25]. Stakeholder interactions occur across multiple networks: the public Internet, the wireless networks provided by mobile operators, and the private networks, which may be operated by intermediaries such as enabling service providers. The adoption of a mobile business service therefore may be dependent on factors related to the role and contribution of each stakeholder group, on the relationships across the value chains in which the stakeholders participate, and on the regulatory and socio-economic environment within a single country or region, or across regions.

The relevant mobile business service supply stakeholder groups can be classified as: 1. Mobile network operators (MNOs); 2. Mobile device developers/vendors (MDDVs); 3. Mobile network services providers (MNSPs); 4. Mobile application developers (MADs); 5. Mobile service content developers (MSCDs); 6. Mobile business services providers (MBSPs); 7. Enabling mobile service providers (EMSPs); 8. Mobile business service aggregators (MBSAs); and 9. Legislators/regulators (LRs) [26]. All stakeholders contribute to the creation of customer service value and may affect customer demand. Service provision and value may also be affected by relevant regulatory and legislative context (Fig. 1).

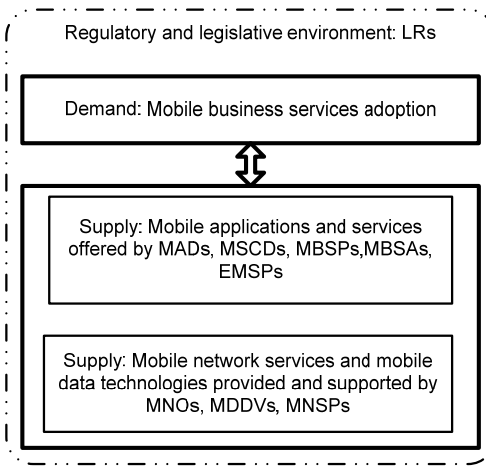


Figure 1. Mobile business service supply and demand framework (adapted from [27]).

At present SMS (Short Messaging Service) and MI (mobile Internet) are the mobile data technologies providing mCommerce platforms. A service may require an application developed for mobile devices e.g., downloadable mobile game software. Finally, services may be developed for multiple platforms: SMS banking provides some of the functionality of a banking online site designed for MI access, however different authentication mechanisms may be used.

While some mobile business services may extend or enhance existing services (e.g., SMS banking) others may also incorporate innovative features such as mPayment and personalization based on location awareness (location-based services – LBS). mPayment and LBS can be viewed both as independent business services, and as enablers of other mobile business services [26]. As the providers of these enabling services participate in the mCommerce value chain, factors related to location and payment can be expected to play a role in user adoption [26][28-32].

A number of classifications of mobile services and applications have been proposed in the literature including a seven-dimensional taxonomy [1]. The results indicate first that most contemporary mobile business services use a B2C model and involve significant personalization as the customer is normally 'known' to the service. Second, approximately half of the services include transaction options and similarly half of them are location-based. While at present the number of synchronous and asynchronous services is approximately the same, the trend in the temporal dimension is to develop synchronous services able to switch to an asynchronous mode when the network is overloaded.

SMS mobile learning (SMS learning), mobile banking (mBanking) and mobile gaming (mGaming) are three typical examples of interactive, transactional B2C mobile business services, which allow (or require) personalization, may run in real time or asynchronously, and may be location-based. All three have been studied empirically in both global and national contexts with respect to their adoption, acceptance and usage. Results indicate that customer requirements and expectations about the functionality of the service, about its design, and about the support of enabling services may play a critical role in the adoption and use of mobile business services.

It is envisaged, for example, that the mass adoption of mBanking would depend on the provision of secure, reliable and easy-to-customize user interfaces that can be implemented on a multi-standard, multi-functional mobile device designed for a long life and a 'rugged' service; customer requirements and the specific socio-cultural context may play a significant and critical role [33-38].

With respect to SMS mLearning, results indicate that it may be difficult to use due to the screen size and text message limitations for SMS learning, may be too costly to be afforded by a student on a limited budget, and may also be perceived as intrusive if large message traffic is generated by the service. It can be argued therefore that to overcome these adoption barriers, a successful mLearning service needs to be affordable, and to provide additional value by being accessible while the learner is travelling (i.e., compatible with the learner's daily

 ISBN: 978-1-61208-100-7

routine), flexible and allowing customization, and available on demand [39-41].

In the case of mGaming, findings from the literature highlight the critical role of both customer perceptions/attitudes and supply chain factors as determinants of adoption and use [42-44]. Customer perceptions about the value of playing a mobile game in the context of their lifestyle may be significant motivators, for example, 'expressiveness' [45] and 'socialization' [46]. On the other hand, enabling services such as payment, and game and device design, may play an important role as contributors to the quality of the mobile gaming experience [47-49].

We draw on these collective findings to suggest that first it is not customer mobility *per se*, but the way service support for customer mobility is used to create customer value, that plays a role as an adoption driver. Second, the quality of the service and its relevance to personal or business lifestyle are the most important decision making factors in the adoption process. There is also evidence to indicate that while innovative mobile services (i.e., services that are not already offered using a different technology) may be compelling if they meet lifestyle needs, mobile services replacing or complementing existing ones will be favored by customers only if their quality is exceptional and motivates 'switching' to the mobile service; here, appropriately designed enabling mobile services may have a significant motivational impact.

The above analysis has enabled us to derive a comprehensive model representing the factors that influence the adoption of mobile services. In particular, we have determined the critical customer benefit-related success factors for the adoption of a mobile business service, and the role of enabling services in the adoption process. The explanatory model that sets out these factors and their inter-relationships is presented in the next section.

III. AN EXPLANATORY ADOPTION MODEL

The Technology Acceptance Model (TAM), the Theory of Planned Behavior (TPB) and the diffusion of innovation theory are among the models used to inform the research design of empirical studies investigating the factors influencing intention to use a mobile business service, and actual use. However these models have some limitations; for example it was recently reported that while the two TAM variables *perceived usefulness* (PU) *and perceived ease of use* (PEU) may be predictors of the *intention* to use a technology, these variables have not been found to be good predictors of *actual usage* [50]; however *intention to use* was found to be a good predictor of *actual usage* [50] and *continuous usage* [11].

It has been suggested in prior work to include in adoption models variables measuring the benefits of the technology to the customer, as adoption models measuring technology do not measure the customer value of the technology [51], and to investigate perceived service value as an adoption factor [11]; [23]. Building on prior work and from the perspective of how a service may provide value and benefit the customer, the factors influencing the adoption of mobile business services can be grouped as shown below [26-27][38][41][52].

- Customer quality of service expectations: Technology factors that relate primarily to the infrastructure and the service architecture (e.g., interoperability of devices and protocols, bandwidth availability, device features and functions, connectivity). The customer may benefit from the advancement of technology, which makes it possible to deliver a mobile business service.
- Customer lifestyle requirements: Consumer factors that relate to how useful and value–adding a mobile business service is perceived to be. For example, in an investigation of how mobile services could help the elderly it was found both PEU and an actual need of the service were important as acceptance criteria [53]. Other factors may include content personalization and localization, service ubiquity, timeliness, convenience, cost, privacy, trust.

Customer quality of service expectations and customer lifestyle requirements are included in the synthesized model as mobile business service adoption antecedents. The model (Fig. 2) is described in more detail next.

A. *Variables and Relationships*

Variable 1 (customer quality of service expectations) refers to the customer in the capacity of a mobile technology user, to follow the terminology in [23]. The variable represents customer expectations about service quality. Possible measures include mobile data service interface PEU (and perceived ease of learning how to use it), and expectations about mobile network performance parameters such as network delay (e.g., synchronicity, a service working in real-time), 24/7 access to the network, seamless handover when the customer is mobile, service availability across different subscriber mobile networks, and affordability (data service cost).

Variable 2 (customer lifestyle requirements) refers to the customer as a consumer of the mobile business service [23]. The variable represents customer requirements with respect to the value of the service and the benefits it may bring. Possible measures include PU and PEU of the mobile business service, perceived service functionality, perceived compatibility with the customer's daily routine, perceived benefits of access to the service 'on the go' compared to other similar services, awareness of the service, perceived added value through customer mobility support, perceived service 'persistence' (sustainability), and affordability (service cost).

Variables 3 and 4 represent constructs used extensively in adoption studies, for example, empirical investigations based on TAM, or on TPB. Variable 3 (intention to use) signifies customer attitude towards using a mobile business service in the future. Variable 4 (actual use) can be measured both through customer self-evaluation (subjective, or perception-based), and through data obtained from service providers (objective, or fact-based) [11][50].

Variable 5 (perceived customer demand) is a new variable, reflecting [multiple] service provider perceptions about customer behavior with respect to the adoption of a mobile business service.

 ISBN: 978-1-61208-100-7

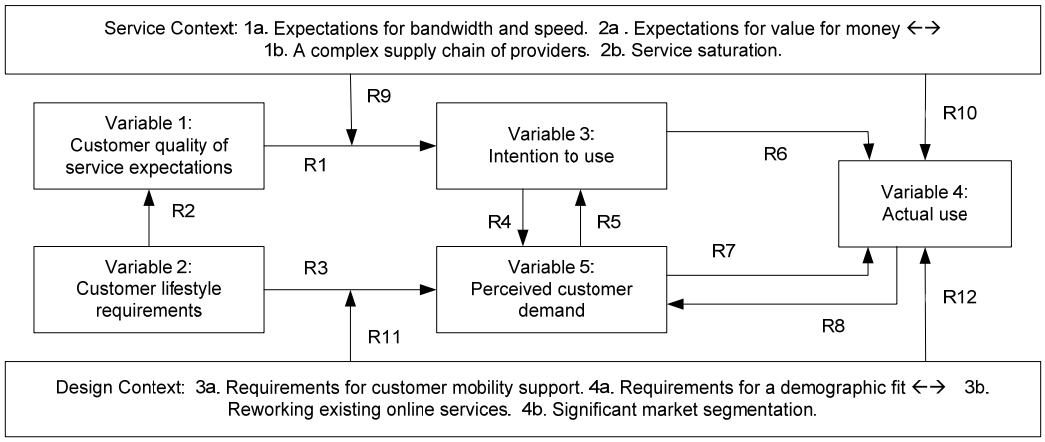


Figure 2. An explanatory model for mobile busines service adoption

Some findings suggest that there may be differences between the intended value proposition and customer acceptance of the proposition [17] due in part to the difficulties involved in communicating it [54]. Possible variable measures include service provider intentions to invest in developing and maintaining a service based on projections about market demand.

Multiple relationships exist between and among the variables. The model depicts a set of twelve relationships that explain the adoption process on both the demand and the supply sides of the value chain, based on the concept of creating value [11][23][25].

Relationships R1-R8 are internal; they capture the interrelationships among the five model variables and relate to the demand side of the value chain - customer perceptions, attitudes and behaviors (Table 1), or perceptions regarding these demand factors. The external relationships R9-R12 refer to the perceptions, attitudes and beliefs of the industry stakeholders. They link the context variables to the rest of the model and make explicit the role of the application design of the mobile business service and the design of the related mobile application(s) as factors influencing adoption and affecting actual use.

The mobile business service supply chain players identified earlier (Fig. 1) can be viewed as the co-creators of mobile business service value [55] and are represented in the explanatory model (Fig. 2) through the two context variables: 'service context' representing the business view of the service and 'design context' representing the technology and services supporting or leveraging customer mobility.

The viability of the business model depends on the perceived service value. For example, SMS based 'test revision' scenarios designed to be used by commuting learners offer a study aid available wherever the learner goes. The convenience of such a learning service may contribute to forming a positive perception as it fits in with the lifestyle of a learner frequently changing locations and spending time in traffic. However additional service value may be added by appropriate service design meeting the learner's quality of service expectations: for example, options such as micro-payment (for just one question/answer pair), a discount for a bundle of questions/answers, and supporting service availability across different subscriber networks, may all provide motivation for actual use [41].

B. The Design Context and the Service Context

The role of the design context may be especially significant where an application needs to be developed in order to deliver or support a service. For example, the perceived value of customer mobility support may also depend on the application design: an application, which requires high mental concentration while being used may be unlikely to be convenient to customers on the move [53].

The service context focuses on service design and on the business model used to deliver new and flexible mobile services [56]. The two contexts complement each other. An example of a partnership model for application design and development aligned with service design and provision, which includes all stakeholders, involved in creating value and bringing benefit to the customer, is provided by the SmartTouch project. SmartTouch is a service based on NFC (near field communication) technology and includes an mPayment and mobile ticketing application. The project development and the commercialization phase are supported by the partners in the international SmartTouch consortium, which includes developers, hardware manufacturers, and data service and business service providers [57].

IV. DISCUSSION AND CONCLUDING REMARKS

The synthesized model presented here in Figure 2 provides a framework for the investigation of the process of adoption of mobile business services, including a multiple stakeholder perspective, which provides a balanced view of both the supply and demand side of mobile business service provision.

 ISBN: 978-1-61208-100-7

TABLE I. MODEL RELATIONSHIPS

Relation-ship	Type	Descriptive Definition
R1	Internal	*Customer expectations* regarding the *quality* of the mobile data and mobile business service as a motivator and *intention to use* decision making factor.
R2	Internal	*Customer requirements* about the particular mobile services and the *value added* to a service by user mobility support as a factor in determining the *service quality requirements* and indirectly as a factor in decision making about *use*.
R3	Internal	*Customer requirements* about particular mobile services and the *value added* to a service by user mobility support as a driver of service *demand*.
R4, R5	Internal	*Intention to use* as a predictor of projected *demand*; service awareness emerging as a result of higher *perceived demand* as motivator for *intention to use*.
R6	Internal	*Intention to use* as a predictor of the *actual use* of a mobile service.
R7, R8	Internal	*Perceived demand* for mobile services as a predictor of *actual use*; *actual use* as a predictor of *perceived demand*.
R9, R10	External	*Mobile service design* as a mediator of mobile business service *quality expectations*; *mobile service design* as a determinant of *actual use*.
R11, R12	External	*Mobile artifact design* as a mediator of mobile service *lifestyle requirements*; *mobile artifact design* as a determinant of *actual use*.

The model may be used to investigate how the service adds value through mobility support (possibly competing with other similar services, which do not support user mobility) and the perceived benefits of the service.

While empirical results obtained in adoption studies have been used to define the constructs of the model, two other theories have been utilized to contextualize it: design science provides a perspective on the relationship between customer requirements and expectations and the design of the mobile application, while service science provides a perspective on how the respective mobile service may generate demand and become viable.

The model is based on the assumption that understanding the motivators of customer decision making about using a mobile business service may provide useful feedback both to developers of mobile applications and to mobile business service providers. Thus it should contribute to the development of viable and valuable mobile service scenarios in an environment characterized by the emergence of new services and technologies, with a significantly increased spectrum of customer choices and business investment opportunities.

In the next stage of this research we will gather and analyze qualitative data in order to operationalize the variable *perceived customer demand* and the two context variables. It is proposed to investigate the role of the service context as a mediator in the relationship between customer quality of service expectations and intention to use/actual use, applying a service science perspective [58] assuming that: i) the value provided through a mobile service, which is able to support customer mobility ('mobility value') has a dual customer/provider nature and needs to be investigated from both customer and service provider perspectives [59], and ii) mobile services are innovative and therefore it is important for their uptake to identify the critical features, which may positively influence use and demand [60].

It is proposed to study the mediating role of the design of the application underlying the service in the relationships between customer lifestyle requirements and perceived customer demand and between perceived customer demand and actual use from a design science perspective. The design science cycle starts with identifying the problem, and continues through the suggestion, development and evaluation steps to the conclusion step, with feedback loops at every step [61]. The study approach is based on the assumption that in the design of mobile applications the problem space is 'fuzzy' (i.e., the problem to be solved is not well defined) and therefore the application evaluation and conclusion steps cannot be completed without understanding how the new mobile artifact or mobile application may fit in a number of possible service use scenarios [23][62].

Methodologically, the research can be viewed as a sequence of two distinct but related investigative phases. The first phase is concerned with the investigation of selected mobile services and the adoption process associated with these. The predominant research thinking underpinning this phase of the work is objectivist although research methods more aligned to an interpretivist approach are deployed. For example the hypotheses formulated and tested statistically in the quantitative studies on mobile banking and mobile gaming adoption highlighted the role of perceived usefulness and of compatibility with customer lifestyle requirements and expectations as adoption antecedents [38][45]. The exploratory analysis of the quantitative survey data gathered for the study of mPayment adoption indicated that the more customers were aware of the service the more likely they were to become regular users and create demand for it, and also leading to improving the design of the service (e.g., user interface), and the design application supporting it used (e.g., security concerns) [63]. The qualitative study of mLearning adoption identified support for mobile lifestyle and providing rich but relevant information as the key contributors to the perceived service value and therefore likely demand drivers [41]. Finally, the research review of LBS mapping LBS development stages to customer expectations and requirements showed that while customer expectations about the quality of the service were high customers were are likely

 ISBN: 978-1-61208-100-7

to adopt even a 'low tech' service if their requirements were met [64].

While empirical results from prior work have informed the development of the model proposed earlier an interpretivist approach will be adopted for the second phase of the investigation where rich subjective qualitative data will be gathered and analyzed with respect to the new model variables [65, pp. 121-134; p. 172][66, pp. 87-116]. Thus with respect to the overall methodology the study can be classified as exploratory in design [67] and following a mixed methods approach [68, p. 642]. It is believed that using a research strategy combining qualitative and quantitative research methods will facilitate a better understanding and interpretation of the relationships between the model variables [68, p. 653].

The contribution of this work to the body of knowledge is a comprehensive explanatory theory of mobile services and applications adoption; in addition, the model may be used to complement the design science evaluation of a mobile application by developing and validating frameworks for evaluating the service value potential of the application from a service perspective [69].

REFERENCES

[1] R. C. Nickerson, U. Varshney, J. Muntermann, and H. Isaac, "Taxonomy development in infromation systems; Developing a taxonomy of mobile applications," Proc. 17th European Conf. Information Systems, June 2009.

[2] S.-J. Hong, J. Y. L., Thong, J.-Y. Moon, and K.-Y. Tam, "Understanding the behaviour of mobile data services consumers," Information Systems Frontiers, vol. 10, no. 4, 2008, pp. 431-445.

[3] G. M. Giaglis, P. Kourouthanassis, and A. Tsamakos, "Towards a classification framework for mobile location services," in Mobile Commerce: Technology, Theory, and Applications, B. E. Mennecke, and T. J. Strader, Eds. Hershey, PA: Idea Group Publishing, 2003, pp. 67-85.

[4] N. Kreyer, K. Pousttchi, and K.Turowski, "Characteristics of mobile payment procedures," Proc. 13th Int. Symp. Methodologies for Intelligent Systems (ISMIS 2002), June 2002, pp.10-22.

[5] O. Källström, "Business solutions for mobile e-commerce," Ericsson Review, vol. 2, 2002, pp. 80-92.

[6] E. Turban, J, Lee, and D. Viehland, Electronic Commerce: A Managerial Perspective, NY: Prentice Hall, 2004.

[7] B. Anckar and D. D'Incau, "Value-added services in mobile commerce: An analytical framework and empirical findings from a national consumer survey," Proc. 35th Hawaii International Conference System Sciences (HICSS'02), vol. 7, IEEE Comp. Soc. Press, Jan. 2002, pp. 1087-1096.

[8] K., Siau, E.-P. Lim, and Z. Shen, "Mobile commerce: Promises, challenges and research agenda, " J. of Database Management, vol. 12, no. 3, 2001, pp. 4-13.

[9] A. Tsalgatidou and J. Veijalainen, "Mobile electronic commerce: Emerging issues", Proc. 1st Int. Conf. Electronic Commerce and Web Technologies (EC-Web 2000), Sept. 2000, pp. 477 – 486.

[10] E. Swilley and R. E. Goldsmith, "The role of involvement and experience with electronic commerce in shaping attitudes and intentions toward mobile commerce," Int. J. of Electronic Marketing and Retailing, vol. 1, no. 4, 2007, pp. 370- 384.

[11] J..Y. L. Thong , S. J. Hong , K.Y. Tam, "The effects of post-adoption beliefs on the expectation-confirmation model for information technology continuance," Int. J. of Human-Computer Studies, vol. 64, no. 9, 2006, pp. 799-810.

[12] S. E. Chang, Y.-J. Hsieh, T.-R. Lee, C. K. Liao, and A.-T. Wang, "A user study on the adoption of location based services," LNCS 4537, Berlin: Springer, 2007, pp. 276-286, doi 10.1007/978-3-540-72909-9_32.

[13] H. de Vos, T. Haaker, and M. Teerling, "Consumer value of context aware and location based services," Proc. 21st Bled Conference. eCollaboration: Overcoming Boundaries through Multi-Channel Interaction, June, 2008, pp. 50-62.

[14] K. Pousttchi and D. G. Wiedermann, "What influences consumers' intention to use mobile payments?," Proc. 6th Annual Global Mobility Round Table (GMR 2007), June 2007, www.marshall.usc.edu/assets/025/7534.pdf (May 1, 2010).

[15] N. Mallat, "Exploring consumer adoption of mobile payments: A qualitative study:", The J. of Strategic Information Systems, vol. 16, no. 4 2007, pp. 413-432.

[16] J. Peppard and A. Rylander, "From value chain to value network: Insights for mobile operators," European Management J., vol. 24, no. 2-3, 2006, pp. 128-141.

[17] M. Akesson, "Value proposition in m-commerce: Exploring service provider and user perceptions," Proc.6th Annual Global Mobility Roundtable (GMR 2007), June 2007, www.marshall.usc.edu/assets/006/5574.pdf (May 1, 2010).

[18] S. Mallenius, M. Rossi, and V. K. Tuunainen, "Factors affecting the adoption and use of mobile devices and services," Proc. 6th Annual Global Mobility Roundtable (GMR 2007), 2007, http://www.marshall.usc.edu/assets/025/7535.pdf (May 1, 2010).

[19] W. Li and R. J. McQueen, "Barriers to mobile commerce adoption: An analysis framework for a country-level perspective," Int. J. of Mobile Communications, vol. 6, no. 2, 2008, pp. 231 – 257.

[20] A. Samtani, T. T. Leow, H. M.Lim, and P. G. J. Goh, "Overcoming barriers to the successful adoption of mobile commerce in Singapore," Int. J. of Mobile Communications, vol. 1, No 1-2, 2003, pp. 194-231.

[21] E. Seeman, M. O'Hara, J. Holloway, and A. Forst, "The impact of government intervention on technology adoption and diffusion: The example of wireless location technology," Electronic Government, vol. 4, no. 1, 2007, pp. 1-19.

[22] T. Laukkanen, "Comparing consumer value creation in Internet and mobile banking," Proc. 4th Int. Conf. Mobile Business (ICMB'05), IEEE Comp. Soc. Press, July 2005, pp. 655-658, doi 10.1109/ICMB.2005.28.

[23] P. Pedersen, "An adoption framework for mobile commerce," in Towards an E-Society: E-Commerce, E-Bsuiness, and E-Government, B. Schmid, K. Stanoevska-Slabeva, and V. Tschammer, Eds. NY: Kluwer/IFIP, 2001, pp. 643-655.

[24] K. Heinonen and M. Pura, "Classifying mobile services," Sprouts: Working Papers on Information Systems, vol. 6, article 42, 2006.

[25] S. J. Barnes, "The mobile commerce value chain in consumer markets, " in m-Business: The strategic Implications of Wireless Technologies, Burlington MA: Elsevier, 2006, pp. 13-37.

[26] K. Petrova, "Mobile payment: Towards a customer-centric model," LNCS 5176, Berlin: Springer, 2008, pp. 12-23.

[27] K. Petrova, "A study of the adoption of mobile commerce applications and of emerging viable business models," Managing Modern Organizations with Information Technology: IRMA 2005 Proc., M. Khosrow-Pour, Ed. Hershey, PA: IGI Global, May 2005, pp. 1133-1136.

[28] T. J. Gerpott and K. Kornmeier, "Determinants of customer acceptance of mobile payment systems," Int. J. of Electronic Finance, vol. 3, no. 1, 2009, pp. 1-30.

[29] T. Dahlberg and N. Mallat, "Mobile payment service development: Managerial implications of consumer value perceptions," Proc. 9th European Conf. Information Systems (ECIS 2002), June, 2002, pp. 649-657.

[30] N. Kreyer, N., K. Pousttchi, K., and K. Turowski, "Standardized payment procedures as key enabling factor for mobile commerce," LNCS 2455, London: Springer, 2002, pp. 400-409.

 ISBN: 978-1-61208-100-7

[31] L. Perusco, and K. Michael, "Control, trust, privacy, and security: Evaluating location-based services," IEEE Technology and Society Magazine, vol. 26, no. 1, 2007, pp. 4-16.

[32] V. Koutsiouris, C. Polychronopoulos, and A. Vrechopoulos, "Developing 3G location based services: The case of an innovative entertainment guide application," Proc. 6th Int. Conf. Management of Mobile Business (ICMB'07), IEEE Comp. Soc. Press, July 2007, pp. 1, doi: 10.1109/ICMB.2007.38.

[33] S. Laforet, and X. Y. Li, "Consumers' attitudes towards online and mobile banking in China" . Int. J. of Bank Marketing, vol. 23, no. 5, 2005, pp. 62-380.

[34] T. Laukkanen, "Bank customers' channel preferences for requesting account balances," Proc. 40th Annual. Hawaii Int. Conf. System Sciences (HICSS'07), IEEE Comp. Soc. Press, Jan. 2007, p.148a, doi: 10.1109/HICSS.2007.101.

[35] T. Laukkanen, S. Sinkkonen, P. Laukkanen, and M. Kivijarvi, "Segmenting bank customers by resistance to mobile banking,", Int. J. of Mobile Communications, vol. 6, no. 3, 2008, pp. 309 – 320.

[36] P. Luarn and H. H. Lin, "Toward an understanding of the behavioural intention to use mobile banking," Computers in Human Behaviour, vol. 21, no. 6, 2004, pp. 340-348.

[37] K. Pousttchi and M. Schurig, "Assessment of today's mobile banking applications from the view of customer requirements," Procs. 37th Annual Hawaii Int. Conf. System Sciences (HICSS'04), IEEE Comp. Soc. Press, Jan. 2004, pp. 70184.1, doi: 10.1109/HICSS.2004.1265440.

[38] K. Petrova and S. Yu, "SMS banking: An investigation of the factors influencing future use," Int. J. of e-Services and Mobile Applications, in press.

[39] J. Taylor, M. Sharples, C. O'Malley, G. Vavoula, and J. Waycott, "Towards a task model for mobile learning: A dialectical approach," Int. J. of Learning Technology, vol. 2, no. 3, 2006, pp.138-158.

[40] A. Kukulska-Hulme, M. Sharples, M. Milrad, I. Arnedillo-Sánchez, and G. Vavoula, "Innovation in mobile learning: A European perspective,". Int. J. of Mobile and Blended Learning, vol. 1, no. 1, 2009, pp. 13–35.

[41] K. Petrova, "An implementation of an m-learning scenario with short text messaging: Analysis and evaluation," Int. J. of Mobile Learning and Organization, vol. 4, no. 1, 2010, pp. 83-97.

[42] K. Petrova, "Mobile gaming: Perspectives and issues," in Encyclopaedia of E-Business Development and Management in the Digital Economy, I. Lee, Ed. Hersey, PA: IGI Global, 2010, pp. 789-800.

[43] M. D. Kleijnen, K. de Ruyter, and M. Wetzels, "Factors influencing the adoption of mobile gaming services," in Mobile commerce: Technology, Theory, and Applications, B. E. Mennecke and T. J. Strader, Eds. Hershey, PA: Idea Group Publishing, 2003, pp. 202-217.

[44] H. A. Hashim, S. H. Ab Hamid, and W. A. Wan Rozali, "A Survey on mobile games usage among the institute of higher learning (IHL) students in Malaysia," Proc. 1st IEEE Int. Symp. Information Technologies and Applications in Education (ISITAE'07), IEEE Press, Dec. 2006, pp. 40-44, doi: 10.1109/ISITAE.2007.4409233.

[45] K. Petrova and H. Qu, "Playing mobile games: Consumer perceptions," Proc. 2nd Int. Conf. e-Business (ICE-B 2007), INSTICC Press, July 2007, pp. 209-214.

[46] S. A. Paul, M. Jensen, C. Y. Wong, and C. W. Khong, "Socializing in mobile games," Proc. 3rd Int. Conf. Digital Interactive Media in Entertainment and Arts (DIMEA 2008), ACM Press, Sept. 2008, pp. 2-9, doi: 10.1145/1413634.1413641.

[47] H. B-L. Duh, V. H. H. Chen, and C. B. Y. Tan, "Playing different games on different phones: An empirical study of mobile gaming," Proc. 10th ACM Int. Conf. Mobile HCI (MobileHCI08), ACM Press, Sept. 2008, pp. 391-394, doi: 10.1145/1409240.1409296.

[48] I. Ha, Y. Yoon, and M. Choi, "Determinants of adoption of mobile games under mobile broadband wireless access environment," Information and Management, vol. 44, no. 3, 2007, pp. 276-286.

[49] M. D. Kleijnen, K. de Ruyter, K., and M. Wetzels, "Consumer adoption of wireless services: Discovering the rules, while playing the game," J. of Interactive Marketing, vol. 18, no. 2, 2004, pp. 51-61.

[50] M. Turner, B. Kitchenham, P. Brereton, S. Charters, and D. Budgen, "Does the technology acceptance model predict actual use? A systematic literature review," Information and Software Technology, vol. 52, no. 5, 2010, pp. 463-479.

[51] T. Dyba, N. B. Moe, and E. Arisholm, "Measuring software technilogy usage: Challenges of conceptualization and operatianlization," Proc. Int. Symp. Empirical Software Enginnering (ISESE 2005), IEEE Comp. Soc. Press, Nov. 2005, pp. 447-458, doi: 10.1109/ISESE.2005.1541852.

[52] K. Petrova, "Understanding the success factors of mobile gaming," in Encyclopaedia of mobile computing & commerce, vol. 1, D. Taniar, Ed. Hershey, PA: IGI Global, 2007, pp. 497-503.

[53] M. Mikkonen, S., Väyrynen, V. Ikonen, and M. O. Heikkilä, "User and concept studies as tools in developing mobile communication services for the elderly," Personal and Ubiquitous Computing, vol. 6, no. 2, 2002, pp. 113 – 124.

[54] A. Osterwalder and Y. Pigneur, "Modeling value propositions in e-business," Proc. 5th Int. Conf. Electronic Commerce (ICEC'03), ACM Press, Sept. - Oct. 2003, pp. 429-436.

[55] J. Spohrer and P. P. Maglio, "The emergence of service science: Toward systematic service innovations to accelerate co-creation of value,",Production and Operations Management, vol.17, no. 3, 2008, pp. 238-246.

[56] H. Demirkan, R. J. Kauffman, J. A. Vayghan, H.-G. Fill, D. Karagiannis, and P. P. Maglio, "Service-oriented technology and management: Perspectives on research and practice for the coming decade," Electronic Commerce Research and Applications, vol. 7, no. 4, 2009, pp. 356-376.

[57] EurekaAlert (November 2, 2009). Taking a touching approach to transport ticketing and home care for elderly. Eurekalert. http://www.eurekalert.org/pub_releases/2009-11/e-tat110209.php (November 30, 2009).

[58] P. P. Maglio and J. Spohrer, "Fundamentals of service science", J. of the Academy of Marketing Science, vol. 36, no. 1, 2008, pp. 18-20.

[59] J. Spohrer, "Services sciences, management, and engineering (SSME) and its relation to academic disciplines," in Services science: Fundamentals, challenges and future developments, B. Stauss, K. Engelmann, A. Kremer, and A. Luhn, Eds. Berlin: Springer, 2008, pp. 11-40.

[60] T. Abe, What is service science? Research report No 246. The Fujitsu Research Institute. Economic Research Center, 2005, Tokyo. http://jp.fujitsu.com/group/fri/downloads/en/economic/publications/report/2005/246.pdf (24 December 2009).

[61] V. Vaishnavi and W. Kuechler, W., Design research in information systems, 2004-2205. desrist.org/design-research-in-information-systems/ (December 24, 2009)

[62] K. Petrova, "Mobile learning as a mobile business application," Int. J. of Innovation and Learning, vol. 4, no. 1, 2007, pp. 1-21.

[63] K. Petrova and R. Mehra, "Mobile payment: An exploratory study of customer attitudes," Proc. 6th Inte. Conf. Wireless and Mobile Communications (ICWMC 2010), 2010, in press.

[64] K. Petrova and B. Wang, "Location-based services deployment and demand: A roadmap model," unpublished.

[65] M. Myers, Qualitative Research in Business and Management. LA: Sage , 2009.

[66] S. J. Taylor and R. Bogdan, Introduction to Qualitative Research Methods, 3rd ed. NY: John Wiley, 1998.

[67] U. Sekaran, Research Methods for Business. A Skill Building Approach, 4th ed. NY: John Wiley, 2003.

[68] A. Bryman ands E. Bell, Business Research Methods. Oxford, UK: Oxford University Press, 2003.

[69] S. L. Vargo, P. P. Maglio, and M. A. Akakaa, "On value and value co-creation: A service systems and service logic perspective," European Management J., vol. 26, no. 3, 2008, pp. 145-152.

Improved Spatial and Temporal Mobility Metrics for Mobile Ad Hoc Networks

Elmano Ramalho Cavalcanti and Marco Aurélio Spohn
Systems and Computing Department
Federal University of Campina Grande, Brazil
Email: {elmano,maspohn}@dsc.ufcg.edu.br

***Abstract*—This work shows that two well-known spatial and temporal mobility metrics for mobile ad hoc networks (MANETs) have drawbacks, possibly leading to invalid results. Based on the concept of spatial dependence in the absence of movement among mobile nodes, we propose mobility metrics able to promptly capture spatial and temporal dependence among mobile nodes. Through simulation, we compared the proposed metrics over a diversified set of synthetic mobility models. The results revealed that our spatial metrics can capture spatial dependence in scenarios having different levels of node pause time. Our temporal metric also demonstrated to be better suited for capturing different levels of temporal dependence, without being biased by node speed. Thus, the proposed mobility metrics can accurately capture spatial and temporal node behavior in MANETs.**

***Index Terms*—ad hoc network; mobility metric; spatial dependence; temporal dependence;**

I. Introduction

To support the growth and development of mobile ad hoc networks (MANETs), researchers from industry and academia have designed a variety of protocols, spanning the physical to the application layer. Analytic modeling and simulation are amongst the most used methods for evaluating MANET protocols. The former has limitations due to the lack of generalization, and the intrinsic high level of complexity [5]. The latter is by far the most used method for designing and evaluating MANET protocols.

A mobility model is one of the most important components in the simulation of MANETs. This component describes the movement pattern of mobile nodes (e.g., people, vehicles), impacting on protocol performance [2], [4], [11], [15], topology and network connectivity [3], [8], [16], data replication [10], and security [7]. Bai et al. [2] demonstrated that the performance of a protocol can vary dramatically depending on the adopted mobility model.

Mobility models can be classified into four categories: random, temporal-based, spatial-based (or group-based), and with geographic restriction [1] (Figure 1). Aiming at measuring quantitatively and qualitatively mobility models, one can use mobility metrics.

Bai et al. [2] proposed a framework to analyze the impact of mobility on performance of routing protocols for MANET. They proposed two metrics to quantify the spatial and temporal dependence of mobile nodes. Since then, several works has been based on these metrics for many purposes [13], [14], [17], [19], [20]. However, we show that those metrics have important drawbacks (Section III). After that, we introduce spatial and temporal metrics that overcome the described limitations, and also propose another spatial metric, based on the average distance among nodes (Section IV).

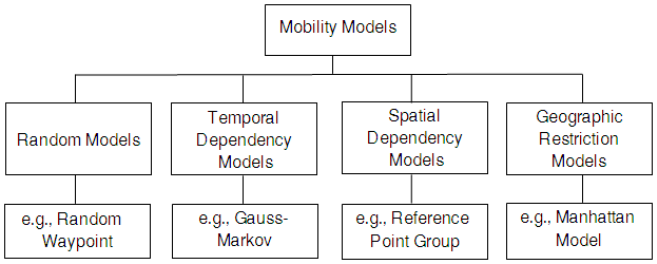


Fig. 1. Categories of mobility models in MANETs [1].

In order to evaluate the proposed metrics, we conducted an extensive simulation using four well know synthetic mobility models (Section V). Afterwards, we perform a comprehensive analysis of metrics behavior (Section VI).

II. Terminology

The following terminology is needed to define the mobility metrics, and it will be used throughout this paper:

- T - Simulation time;
- N - Number of mobile nodes;
- X, Y - Length and width of the scenario;
- R - Radio communication range;
- $x(i,t)$ is the x-coordinate of node i at time t (idem for $y(j,t)$).
- $\theta(i,t)$ is the velocity angle of node i at time t.
- $v(i,t)$ is the velocity of node i in the time t and $v(i,t_0..t_k)$ means that the velocity of node i remains constant from t_0 to t_k.
- $Cos(i,j,t)$ is the cosine of angle between the velocities of nodes i,j:

$$Cos(i,j,t) = \frac{\vec{v}(i,t) \bullet \vec{v}(j,t)}{|\vec{v}(i,t)| \cdot |\vec{v}(j,t)|} \quad (1)$$

- $SR(i,j,t)$ is the speed ratio between nodes i,j at time t:

$$SR(i,j,t) = \frac{min(\vec{v}(i,t), v(\vec{j,t}))}{max(\vec{v}(i,t), v(\vec{j,t}))} \quad (2)$$

 ISBN: 978-1-61208-100-7

- $D(i,j,t)$ is the Euclidean distance between nodes i,j at time t:

$$D(i,j,t) = \sqrt{(x(j,t) - x(i,t))^2 + (y(j,t) - y(i,t))^2} \quad (3)$$

- $\rho(M_p, m)$: indicates the Pearson correlation between the parameter p of the mobility model M and the metric m.

III. Related Work

Bai et al. [2] proposed, in their IMPORTANT framework, two mobility metrics that should be able to quantify spatial and temporal movement dependence among mobile nodes. Both metrics are based on the cosine similarity between the velocities of nodes (Equation 1).

The first one is the Degree of Spatial Dependence between nodes i,j at time t ($DSD(i,j,t)$), defined in Equation 4.

$$DSD(i,j,t) = Cos(i,j,t) \bullet SR(i,j,t) \quad (4)$$

Therefore, the average degree of spatial dependence (DSD) is given as the average between all nodes during the simulation. Group-based mobility models (e.g., RPGM [9]) should present high values for DSD.

The second mobility metric proposed by Bai et al. is the Degree of Temporal Dependence (DTD) (Equation 5), which is calculated similarly to DSD but it considers the difference of velocities between two time slots. Thus, the current velocity of a mobile node is dependent on its past moving pattern. This metric reflects the smoothness of node movement.

$$DTD(i,t,t') = Cos(\vec{v}(i,t), \vec{v}(i,t')) \bullet SR(\vec{v}_i(t), \vec{v}_i(t')) \quad (5)$$

Temporal mobility models (e.g., Gauss-Markov [12]) should present high values for DTD, while strongly random models should have null DTD (i.e., zero). For the former models, node velocity changes incrementally, unlike the abrupt changes occurring in random models (e.g., Random Waypoint).

Based on the work by Bai et al. [2], Zhang et al. [20] extended and developed the concept of a very similar spatial mobility metric, called Spatial Dependence (SD). The authors used this metric in the design of a distributed group mobility adaptive clustering algorithm (i.e., DMGA) [20]. However, both DSD and SD present the same limitation, which is described in next section.

A. Limitations on Previous Metrics

The main limitation on the DSD metric (Equation 4) is that it does not consider spatial dependence (correlation) in the absence of node movement. While two nodes i,j are pausing, their correlation is always zero (i.e., $Cor(i,j,t) = 0$), what is not necessarily true because nodes i and j might have paused (i.e., switched to velocity zero) just because there is some dependence between them.

To demonstrate that the assumption may be wrong, consider two mobile nodes, B and C, which are moving in accordance to the movement pattern of their leader, node A (Figure 2). At time t_0, nodes B and C are inside node's A transmission range and they all start moving to points A', B', and C' where $v(A, t_0..t_9) = 15m/s$, $\theta(A, t_0..t9) = 0$, $v(B, t_0..t_9) = 18m/s$, $\theta(B, t_0..t9) = \beta$, $v(C, t_0..t_9) = 16m/s$ and $\theta(C, t_0..t9) = \alpha$. By that time, they stop from t_{10} to t_{19}. For this group mobility scenario, the DSD metric just captures the correlation during the movement period (i.e., from t_0 to t_9), while the correlation is considered null during pause times (Table I). There is a clear spatial dependence among nodes during pause times, but it is not captured by the DSD metric. From t_0 to t_9 the total degree of spatial dependence is .7034. By time $t = 19$, DSD has decayed to .3517. In case the nodes continue paused for an additional 10 s, the DSD decreases to .2345. Thus, the higher the node pause time, the lower the metric value. Therefore, it is paramount considering the correlation during pause periods.

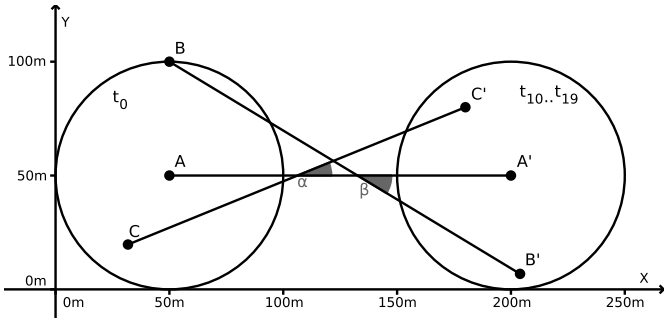


Fig. 2. Example of a group mobility scenario.

TABLE I
PAUSE CORRELATION PROBLEM IN A GROUP MOBILITY SCENARIO.

Nodes	Movement (t:0..9)	Pause (t:10..19)
	v(i,t)	
A	15 m/s	0 m/s
B	18 m/s	0 m/s
C	16 m/s	0 m/s
	DSD(i,j,t)	
A,B	0.71	0
A,C	0.87	0
B,C	0.53	0

IV. Contributions

We propose the *Improved Degree of Spatial Dependence* (IDSD), a spatial mobility metric which is able to capture both movement and pause correlation among mobile nodes.

The second contribution is the proposal of the *Improved Degree of Temporal Dependence* (IDTD), a temporal mobility metric based on DTD [2]. Besides the pause correlation problem, DTD was not able to distinguish temporal from atemporal mobility models [2]. We verified that this is due to improperly computing that metric: instead of computing $DTD(i,j,t)$ for each time slot t, it should only be computed when the velocity $v(i,t)$ changes in magnitude or direction. With this simple modification, IDTD can, in addition to other benefits, distinguish temporal and atemporal mobility models.

IDTD is also substantially less impacted by node speed. In addition to that, it has higher correlation to parameter

α, a memory level parameter commonly defined in temporal models such as Gauss-Markov [12] and Semi-Markov Smooth [21].

Our third contribution is a novel spatial mobility metric, named Degree of Node Proximity (DNP), which is capable of distinguishing group-based mobility models from others. Besides that, simulation results show that DNP is less impacted by node pause time than DSD.

A. Pause State Movement Dependence

It is reasonable to consider that spatial dependence between two nodes i, j at a pause time step t, $DSD(i, j, t)$, will be equal to the average of the last K values. The higher the average pause time, greater is the value of K. Thus, the Improved Degree of Spatial Dependence metric equation is given by:

$$IDSD(i,j,t) = \begin{cases} PC(i,j,t) & \text{if } \vec{v}(i,t) = \vec{v}(j,t) = 0, \\ DSD(i,j,t) & \text{otherwise.} \end{cases} \tag{6}$$

where $PC(i, j, t)$ is the pause correlation between nodes i, j at time t. It is computed as follows:

$$PC(i,j,t) = \frac{1}{K} \sum_{k=t-K}^{t-1} DSD(i,j,k) \tag{7}$$

where K is a function of the average pause time, a typical mobility model input parameter[1].

B. Improved Degree of Temporal Dependence

As explained previously, the metric DTD (Equation 5) should be computed only when the node velocity changes, otherwise, it will not be able to promptly catch the temporal node behavior of a mobility model. Thus, we have that IDTD$(i, t) = 0$ if $v(i, t) = v(i, t-1)$ and $\theta(i, t) = \theta(i, t-1)$, and IDTD$(i, t)$ = DTD(i, t) otherwise. Therefore, the Improved Degree of Temporal Dependence ($IDTD$) metric is defined as follows:

$$IDTD = \frac{1}{P} \sum_{i=1}^{N} \sum_{t=1}^{T} IDTD(i,t) \tag{8}$$

where P is the number of tuples (i, t) such that $IDTD(i, t) \neq 0$.

C. Degree of Node Proximity

We propose a spatial mobility metric based on the distance between pairs of nodes, called Degree of Node Proximity (DNP). Let AD be the average distance between all nodes during the simulation, and MAD be the maximum average distance expressed in units of transmission range, R. Formally speaking:

[1]In fact, some mobility models have the maximum pause time (MPT) parameter instead of average pause time (APT). For those cases, pause time generally has an uniform probability distribution function, and then $APT = MPT/2$.

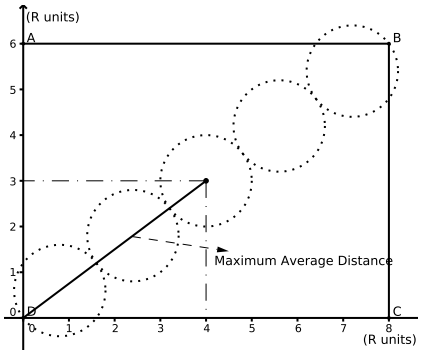


Fig. 3. Example of Maximum Average Distance (MAD).

$$AD = \frac{1}{N(N-1)/2} \sum_{i=1}^{N} \sum_{j=i+1}^{N} \frac{\sum_{t=1}^{T} D(i,j,t)/R}{T} \tag{9}$$

$$MAD = \frac{\sqrt{X^2 + Y^2}}{2\dot{R}} \tag{10}$$

Suppose a scenario where its width is 600 m, its length is 800 m, and the node transmission range is 100 m. Then, the maximum average distance, MAD, is given by $5R$ (or 500 m). Figure 3 illustrates exactly this situation.

The proportion about AD and MAD gives a notion about the degree of mobility dependence. When the average distance among the nodes is constantly low, then this probably means that nodes follow some sort of group-mobility movement. For this reason, we define our spatial mobility metric DNP as expressed in Equation 11.

$$DNP = 1 - \frac{AD}{MAD} \tag{11}$$

DNP values normally range from 0 to 1. Spatial mobility models (e.g., RPGM [9]) should present high DNP values, while other models should present lower values. Next, we present an extensive simulation using these metrics and a heterogeneous set of mobility models.

V. Simulation

To verify the ability that our proposed mobility metrics have to capture spatial and temporal dependence among mobile nodes, we selected the following mobility models (the same displayed in Figure 1):

- Random Waypoint (RWP) [6]: is probably the simplest and most used mobility model in MANET simulation studies. It has just three parameters: minimum and maximum speeds, and maximum pause time.
- Reference Point Group Mobility (RPGM) [9]: is a group-based model where the movement of the leader of a group influences the movement of all its members. The distance between the leader and his members should not be greater than a threshold, called maximum distance from center (MDC). RPGM is more applicable for battle field or rescue operations scenarios.

 ISBN: 978-1-61208-100-7

TABLE II
MOBILITY MODELS SELECTED FOR SIMULATION.

Feature	RWP [6]	RPGM [9]	GM [12]	MAN [2]
Randomness	high	moderate	variable	moderate
Group-based		X		
Temporal			X	
Grid-based				X

- Gauss-Markov (GM) [12]: in this model the velocity of mobile node is assumed to be correlated over time and modeled as a Gauss-Markov stochastic process. GM is a temporally dependent mobility model whereas the degree of dependency is determined by the memory level parameter α ($0 < \alpha < 1$).
- Manhattan (MAN) [2]: is a grid-based model where nodes follow specific paths (e.g., streets) distributed in a rectangular grid. It is suitable for modeling the movement of vehicular wireless networks.

The results presented in this paper depend on some assumptions which are required for computing mobility metrics:

i. Communication between nodes is always bidirectional during the simulation.
ii. R is constant and equal for all nodes.
iii. N is constant during the simulation.
iv. The scenario has a two-dimensional square geometry.

Table II summarizes the main characteristic of the selected mobility models. RPGM and MAN models are classified as having moderate randomness. The former, because the movements of regular nodes are limited to their leader's, and in the latter node movements are limited due to obstacles spread over the scenario (e.g., city blocks). Gauss-Markov (GM) presents variable randomness, since it depends on the value of the memory parameter α.

BonnMotion [18] was employed for mobility scenario generation producing the synthetic traces for the mobility models. For all scenarios, 100 nodes moved over an area of 1000m x 1000m for a period of 900 seconds. Transmission range was set to 100, 150, and 200 meters. For the RPGM model, the number of nodes per group (NG) was set to 10, 25, and 50, which represents scenarios with 2, 4, and 10 groups of mobile nodes. Maximum pause time and memory parameter were set to a large range of values (see Table III).

All graphs present results with a confidence level of 99%, based on 10 repetitions for each one of more than 1,400 generated mobility scenarios. In some situations, the interval length is smaller than the symbol used in the legend, making it barely visible.

VI. ANALYSIS

Firstly, we compare the performance of the temporal metrics DTD and $IDTD$. Then we show how the node pause time affects the spatial metrics DSD and $IDSD$. Lastly, we also show that our proposed metric, degree of node proximity (DPN), is able to differentiate the mobility models used in our simulation, and that it is not impacted by node pause time.

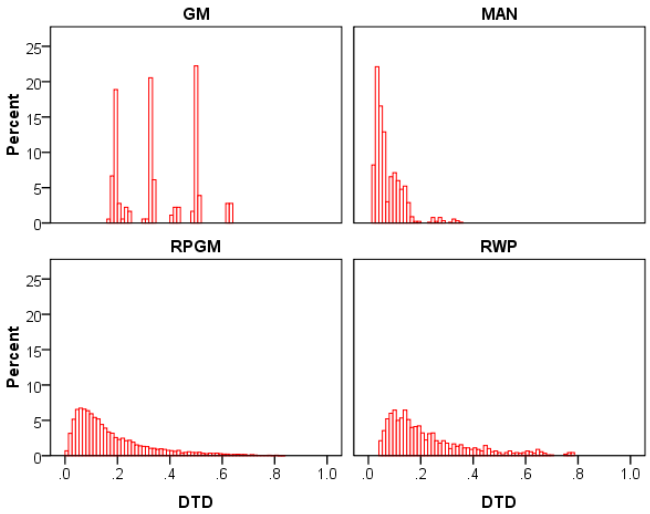


Fig. 4. Degree of Temporal Dependence (DTD) percentage histograms.

A. Temporal Metrics

Table IV shows the basic descriptive statistics for the mobility metrics. In general, random models showed moderate DTD values, what was not expected. For some scenarios, the DTD value for Random Waypoint and RPGM even surpassed GM's. The percentage histogram of DTD clearly reveals this shortcoming (see Figure 4). On the other hand, $IDTD$ properly identified the Gauss-Markov model as the unique temporal model among all under consideration, and the metric correctly considered that the other models should have values close to zero (Table IV).

The second problem with DTD metric is that it is very little impacted when changing the memory parameter α in the Gauss-Markov model (Figures 5, 6, and 7). The unique visible change happened when $\alpha = .99$. However, $IDTD$ demonstrated a higher correlation with α (.91 versus .35, Table V), and consequently is better in capturing different levels of temporal dependence than DTD.

The third problem is that DTD decreases with the increment of node speed. When maximum node speed (S) is 10 m/s, DTD is, on average, nearly 0.5. When S increases from 20 to 30, DTD decreases from 0.3 to 0.2 (Figures 5, 6, and 7). Nevertheless, this relationship is almost imperceptible with the $IDTD$ metric.

B. Spatial Metrics

As stated in Section III-A, the degree of spatial dependence DSD does not capture pause state spatial dependence (presented in Section IV-A).

Figure 8 shows the different effect that the variation of maximum node pause time MPT causes on DSD and $IDSD$ in the RPGM model with 10 groups of 10 nodes each. At point $MPT = 0$, both DSD and $IDSD$ have the same value, because nodes never stop moving. As the node pause time increases, DSD quickly decreases. However, $IDSD$ increases a little bit and keeps at about the same level

TABLE III
CONFIGURATION OF MOBILITY MODELS' INPUT PARAMETERS FOR SIMULATION.

PARAMETER - unit	Gauss-Markov [12]	Random Waypoint [6]	RPGM [9]	Manhattan [2]
Simulation Time (T) - s	900			
Number of nodes (N)	100			
Transmission range (R) - m	100, 150, 200			
Scenario's length (X) - m	1000			
Scenario's width (Y) - m	1000			
Minimum speed (s) - m/s		1, 3, 5		
Maximum speed (S) - m/s		10, 20, 30		
Average speed (AS) - m/s	f(S)[a]			6, 11, 16
Speed Standard Deviation (SSD)	f(S,AS) [b]			f(s,AS) [c]
Maximum pause time (MPT) - s		0, 100, 200, 300, 400, 500, 600, 700, 800, 900		
Number of nodes per group (NG)			10, 25, 50	
Memory Parameter (α)	.0, .2, .4, .6, .8, .99			
Number of rows (NR)				10
Number of columns (NC)				10
Max. deviation from leader (MDC) - R			1	
Speed change probability (SCP)				10%
Total number of experiments	2,700	540	8,100	2,700

[a] It is a function of maximum speed (S).
[b] It is a function between average (AS) and maximum speed (S).
[c] It is a function between average (AS) and minimum speed (s) for MAN model.

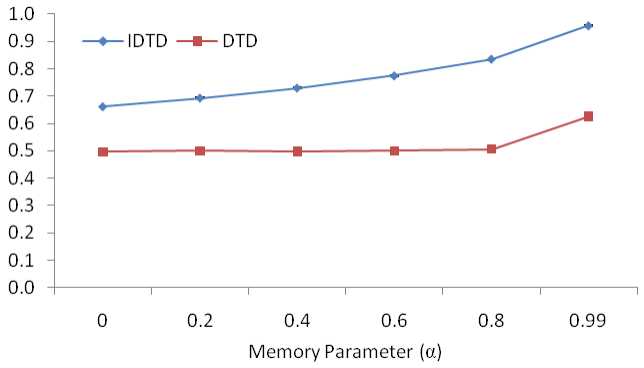


Fig. 5. Effect of memory parameter on the temporal mobility metrics ($S = 10m/s$, $R = 150m$).

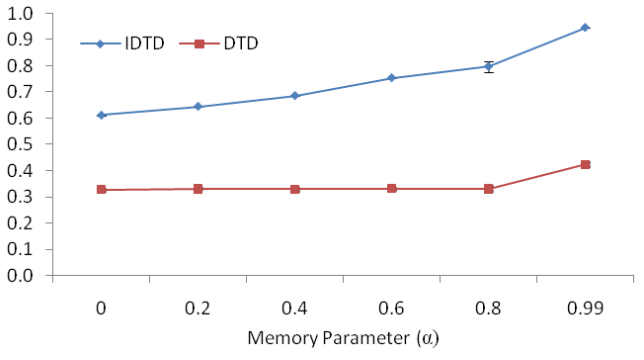


Fig. 6. Effect of memory parameter on the temporal mobility metrics ($S = 20m/s$, $R = 150m$).

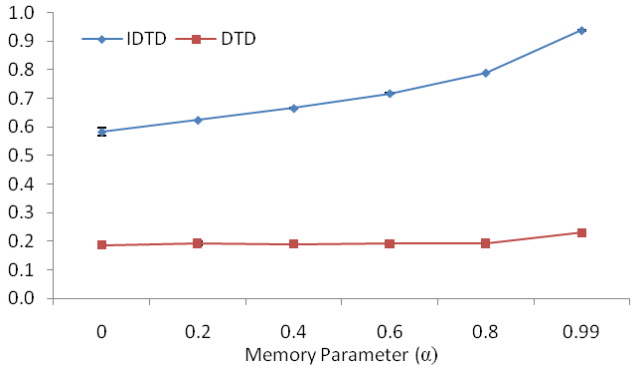


Fig. 7. Effect of memory parameter on the temporal mobility metrics ($S = 30m/s$, $R = 150m$).

until $MPT = 500s$, when then it starts decreasing. Similar behavior also happens in Figures 9 and 10. Although $IDSD$ also decreases, this occurred in a much more slower fashion than occurred with DSD. This is due to the smaller correlation between MPT and $IDSD$ ($\rho(RPGM_{MPT}, DSD) = -.58$ and $\rho(RPGM_{MPT}, IDSD) = -.32$, Table V).

Therefore, $IDSD$ presents more accurate values for spatial dependence among nodes than DSD. For most real scenarios, where MPT is low or moderate, $IDSD$ keeps nearly the same value as for $MPT = 0$. Even in unusual scenarios, where nodes stay longer paused than moving, $IDSD$ still presents higher spatial dependence values.

Concerning our second spatial mobility metric, Degree of Node Proximity (DNP), its histogram clearly distinguished the spatial dependency model (RPGM) from others, and showed similar patters for RWP and MAN models (Figure 11). GM presented the lowest DNP standard deviation.

Figure 12 shows the effect that MPT causes on DNP for all the mobility models that have that input parameter.

 ISBN: 978-1-61208-100-7

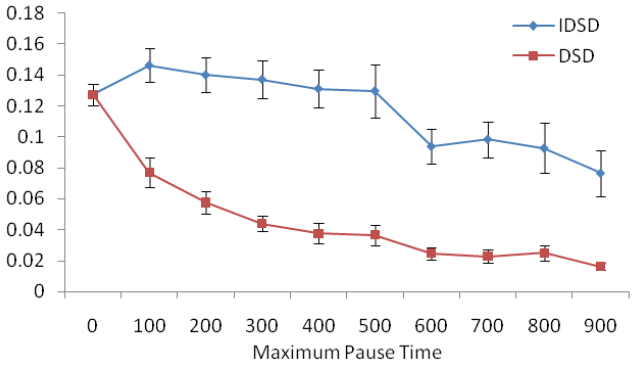


Fig. 8. Effect of pause time on the spatial mobility metrics in RPGM with 10 groups ($s = 3m/s$, $S = 20m/s$, $R = 150m$).

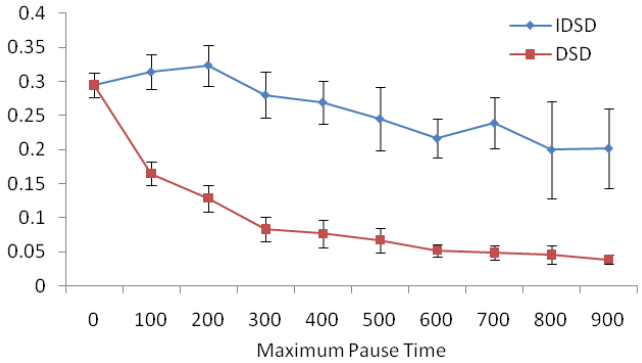


Fig. 9. Effect of pause time on the spatial mobility metrics in RPGM with 4 groups ($s = 3m/s$, $S = 20m/s$, $R = 150m$).

In the RPGM model, MPT caused a constant small drop in DNP. In the RWP model, DNP has a considerable drop for $MTP = 100s$, but then it remains approximately constant. On the other hand, the DNP in the MAN model was little affected by MPT.

Comparing the relationship between $IDSD$ and MPT, and between DNP and MPT, the difference is that in the last one, there is a constant small decay of DNP, instead of in the $IDSD$, when it starts to decay for higher MPT values. Anyway, both $IDSD$ and DNP are better than DSD, as they are extensively less affected by MPT.

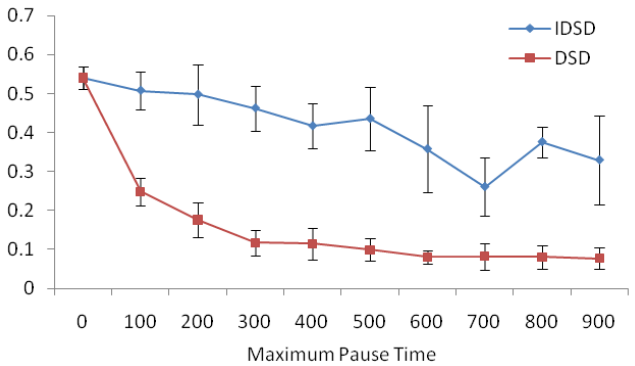


Fig. 10. Effect of pause time on the spatial mobility metrics in RPGM with 2 groups ($s = 3m/s$, $S = 20m/s$, $R = 150m$).

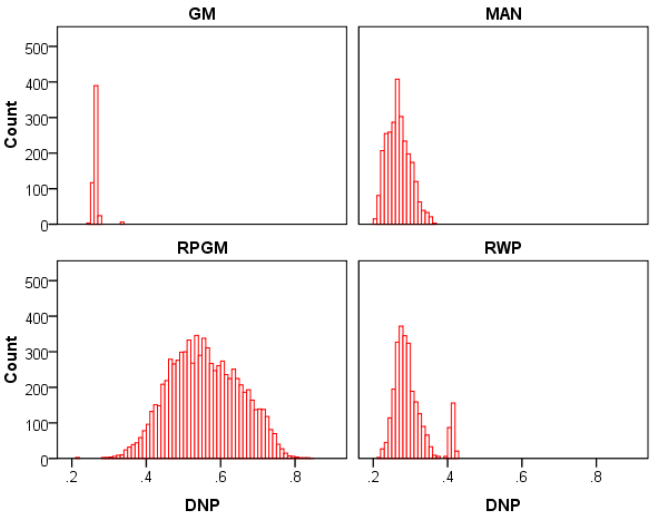


Fig. 11. Histogram-DNP.

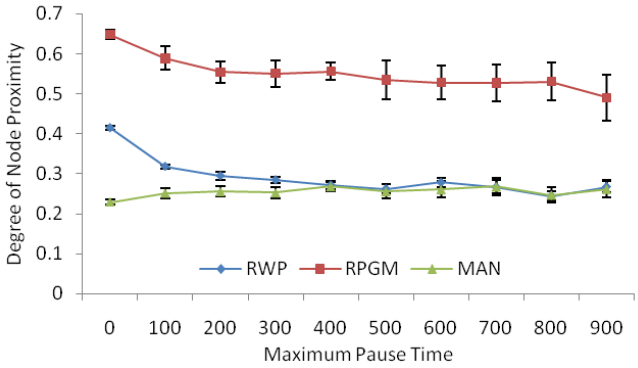


Fig. 12. Effect of pause time on the degree of node proximity metric.

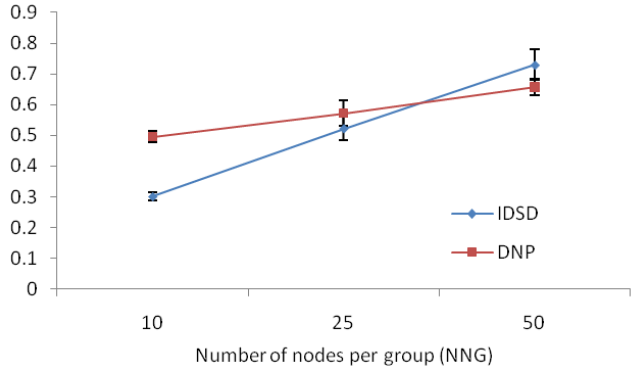


Fig. 13. Effect of number of nodes per group on metric IDSD and DNP (RPGM).

TABLE IV
DESCRIPTIVE STATISTICS FOR THE MOBILITY METRICS.

Metric	Model	Mean	STD	Min	Max
Degree of Spatial Dependence (DSD)	RWP	.007	.005	-.008	.029
	RPGM	.117	.126	.005	.805
	GM	.001	.005	-.012	.017
	MAN	.013	.012	-.004	.058
Degree of Temporal Dependence (DTD)	RWP	.226	.157	.042	.783
	RPGM	.175	.143	.003	.833
	GM	.355	.137	.17	.632
	MAN	.076	.056	.02	.353
Improved Degree of Spatial Dependence	RWP	.010	.006	-.008	.032
	RPGM	.279	.176	.024	.881
	GM	.001	.005	-.012	.017
	MAN	.017	.011	-.004	.059
Improved Degree of Temporal Dependence	RWP	.000	.000	.000	.001
	RPGM	.000	.000	.000	.003
	GM	.744	.112	.53	.958
	MAN	.012	.025	.001	.105
Degree of Node Proximity (DNP)	RWP	.299	.044	.245	.415
	RPGM	.558	.080	.374	.744
	GM	.271	.023	.261	.337
	MAN	.268	.023	.223	.321

TABLE V
CORRELATION MATRIX BETWEEN INPUT PARAMETERS AND MOBILITY METRICS.

Metric	Model	R	s	S	AS	MPT	NG	α
DSD	RWP	.18	.00	-.64		.25		
	RPGM	-.33	-.09	-.16		-.58	.38	
	GM	-.07		.08	.08			-.17
	MAN	.16	.02		-.66	-.06		
DTD	RWP	-.66	-.12	-.14		-.28		
	RPGM	-.11	-.37	-.39		-.66	.00	
	GM	.72		-.28	-.28			.35
	MAN	-.54	-.04		-.25	-.32		
$IDSD$	RWP	-.06	.04	-.64		.31		
	RPGM	-.59	.00	-.05		-.32	.67	
	GM	-.07		.08	.08			-.17
	MAN	.05	.02		-.68	.00		
$IDTD$	RWP	.00	.10	.05		-.51		
	RPGM	.07	.10	.04		-.51	.00	
	GM	.00		-.23	-.23			.91
	MAN	.00	.03		.06	-.58		
DNP	RWP	.00	-.06	-.12		-.82		
	RPGM	.24	.01	-.03		-.49	.77	
	GM	.00		-.22	-.22			.31
	MAN	.00	-.00		.36	.69		

VII. CONCLUSION AND FUTURE WORK

In this paper we introduced the concept of pause state movement dependence, which considers the possible existence of spatial dependence among nodes in a mobile ad hoc network (Section IV-A). From this concept, we proposed the Improved Degree of Spatial Dependence ($IDSD$) mobility metric. $IDSD$ revealed to be better than DSD [2] at capturing the spatial dependence in scenarios having different patterns of node pause times (Section VI-B).

We also proposed another spatial mobility metric, Degree of Node Proximity DNP, which also presented better results than DSD. Besides this, we also proposed a new temporal mobility metric, called Improved Degree of Temporal Dependence ($IDTD$) that demonstrated to be better than DTD [2] in three aspects: capturing different levels of temporal dependence, properly identified the Gauss-Markov model as the unique temporal model among all under consideration in this work, correctly setting other models to produce values near zero, and it is not influenced by node speed.

For future work we plan to investigate the use of the proposed mobility metrics in the design of mobility-aware adaptive routing protocols for mobile ad hoc networks.

REFERENCES

[1] F. Bai and A. Helmy, *Wireless Ad Hoc and Sensor Networks*. Kluwer Academic Publishers, June 2004, ch. A Survey of Mobility Modeling and Analysis in Wireless Adhoc Networks.

[2] F. Bai, N. Sadagopan, and A. Helmy, "IMPORTANT: A framework to systematically analyze the impact of mobility on performance of routing protocols for adhoc networks," in *Proc. of IEEE INFOCOM*, 2003.

[3] C. Bettstetter, "On the connectivity of ad hoc networks," *The Computer Journal*, vol. 47, no. 4, pp. 432–447, July 2004.

[4] J. Boleng, W. Navidi, and T. Camp, "Metrics to enable adaptive protocols for mobile ad hoc networks," in *ICDCS Workshops*, 2002, pp. 293–298.

[5] A. Boukerche and L. Bononi, "Simulation and modeling of wireless, mobile and ad hoc networks," in *Mobile Ad Hoc Networking*. Wiley-IEEE Press, 2004, pp. 373–410.

[6] J. Bronch, D. Maltz, D. Johnson, Y.-C. Hu, and J. Jetcheva, "A performance comparison of multi-hop wireless ad hoc network routing protocols," in *MobiCom*, 1998, pp. 85–97.

[7] S. Capkun, J.-P. Hubaux, and L. Buttyan, "Mobility helps peer-to-peer security," *IEEE Transactions on Mobile Computing*, vol. 5, no. 1, January 2006.

[8] M. Grossglause and D. Tse, "Mobility increases the capacity of ad-hoc wireless networks," *IEEE/ACM Transactions on Networking*, vol. 10, pp. 477–486, 2001.

[9] X. Hong, M. Gerla, G. Pei, and C.-C. Chiang, "A group mobility model for ad hoc wireless networks," in *ACM MSWiM*, August 1999, pp. 53–60.

[10] J. L. Huang and M. S. Chen, "On the effect of group mobility to data replication in ad hoc networks," *IEEE Transactions on Mobile Computing*, vol. 5, no. 5, pp. 492–507, 2006.

[11] V. Lenders, J. Wagner, and M. May, "Analyzing the impact of mobility in ad hoc networks," in *REALMAN*. New York, NY, USA: ACM, 2006, pp. 39–46.

[12] B. Liang and Z. Haas, "Predictive distance-based mobility management for pcs networks," in *Proc. of IEEE INFOCOM*, April 1999, pp. 1377–1384.

[13] Y. Lu, H. Lin, Y. Gu, and A. Helmy, "Towards mobility-rich performance analysis of routing protocols in ad hoc networks: Using contraction, expansion and hybrid models," in *IEEE International Conference on Communications (ICC)*, June 2004.

[14] K. Maeda, K. Sato, K. Konishi, A. Yamasaki, A. Uchiyama, H. Yamaguchi, K. Yasumoto, and T. Higashino, "Getting urban pedestrian flow from simple observation: realistic mobility generation in wireless network simulation," in *ACM MSWiM*. ACM, 2005, pp. 151–158.

[15] G. Ravirikan and S. Singh, "Influence of mobility models on the performance of routing protocols in ad-hoc wireless networks," in *IEEE VTC*, 2004, pp. 2185–2189.

[16] P. Santi, *Topology Control in Wireless Ad Hoc and Sensor Networks*, 1st ed. Wiley, 2005.

[17] C. Shete, S. Sawhney, S. Hervvadkar, V. Mehandru, and A. Helmy, "Analysis of the effects of mobility on the grid location service in ad hoc networks," in *IEEE International Conference on Communications (ICC)*, June 2004, pp. 4341–4345.

[18] C. Waal and M. Gerharz, *BonnMotion - a mobility scenario generation and analysis tool*, University of Bonn, http://bonnmotion.iv.cs.uni-bonn.de/, 2009, last access date: July 23, 2010.

[19] S. Williams and D. Hua, "A group force mobility model," *Simulation Series*, vol. 38, no. 2, pp. 333–340, 2006.

[20] Y. Zhang, J. Ng, and C. Low, "A distributed group mobility adaptive clustering algorithm for mobile ad hoc networks," *Computer Communication*, vol. 32, no. 1, pp. 189–202, 2009.

[21] M. Zhao and W. Wang, "A unified mobility model for analysis and simulation of mobile wireless networks," *ACM-Springer Wireless Networks (WINET)*, vol. 15, no. 3, pp. 365–389, April 2009.

 ISBN: 978-1-61208-100-7

A Heap-based P2P Topology and Dynamic Resource Location Policy for Process Migration in Mobile Clusters

Y. Mohamadi Begum, M.A. Maluk Mohamed
Software Systems Group
M.A.M. College of Engineering
Anna University
Tiruchirappalli, India
ssg_mohamadi, ssg_maluk@mamce.org

***Abstract*— A mobile cluster experiences disruption in execution of long-running applications due to its highly dynamic nature. Process migration handles such dynamism to have seamless computing with minimal disruption. The challenge in process migration is that it should take considerably less time and techniques adopted for static networks are not suitable for mobile networks. This work is a novel effort that organizes the cluster as a heap-based super P2P structure and process state is transferred in terms of object migration between the peers. Also, while migrating processes, load balancing is dynamically done. As the mobile cluster has heterogeneous nodes with varying processing capabilities, we devise a mechanism for computing the capabilities of these nodes. Considering the capability and current load of the nodes the right destination for process migration is chosen and thus we attempt at a better location policy for the migrated process.**

Keywords- DHT; load balancing; mobile cluster; P2P networks; process migration.

I. INTRODUCTION

A mobile cluster (MC) is a Network of Workstations (NOW) or nodes that may be both stationary as well as mobile. The mobile nodes (MN) communicating over a cellular network, may leave or enter a cell any moment of time, making it difficult to run long-running applications on the MC. There are basically three issues that need to be addressed because of the mobility and the resource-poor nature of the nodes. Firstly, the mobile nodes may enter into doze mode or voluntarily disconnect from the entire network affecting the overall cluster availability and performance. Secondly, disconnections can be abrupt, where the device may enter into a region, out of coverage. However, during such disconnections, the communication link would be maintained by the Communication Subsystem (CS) of the Mobile-OS. Since the seamless communication is maintained by the CS, the out of coverage issue has no effect on the on-going computation. That is, we can simply move the computation along with the mobile node. Finally, there may be disruption in service due to the sudden failure of the node, which requires periodic checkpointing of the processes running on that node and applying migration strategy discussed in this paper. Hence, in this work, we consider only the voluntary disconnections as an appropriate issue that needs to be resolved.

This work focuses on pre-determined sign-off occurring due to such voluntary disconnections. By anticipating such disconnections, the MN has two options, namely, one that is followed in Coda file system [2], where it pro-actively downloads any data that is required so as to function independently of the network in carrying out the task assigned. Coda attempts at distributed file sharing applications and not compute-intensive applications. It assumes higher bandwidth communication with its servers. Further, it requires user's prediction on future needs for its cache management policy. The second option for the MN is to checkpoint its process state so that it transfers the same to another node for resuming execution. In our work, we choose the second option of process migration (PM) [3]. PM is associated with moving a process state from one node to another for resuming execution on the latter. In systems with only static nodes, there are a number of implementations for process migration. However, in MC those techniques adopted for static nodes are not applicable or even irrelevant owing to the mobility constraints. It adds to the complexity when process migration is to be coupled with load balancing.

The impact of mobility [1] on distributed computations is severe that it requires a totally different approach for both PM and load balancing. The migration cost is typically a function of address space size of the process and nevertheless includes the cost of locating an apt destination node. The cost incurred in locating such a node in a static network is obviously less compared to a dynamic one. In a MC, what magnifies the cost is the way the devices communicate with each other. Here the task is assigned to a mobile device only through its Base Station (BS). Hence, a copy of the program code as well as the static data is already available with the BS. Whenever a migration request comes from a MN, the BS pro actively sends this to the most eligible device in its cell based on its computational power and its current load. After receiving the process state from the MN, the BS transfers the same to the destination. The goal of load balancing is to assign to each node tasks

 ISBN: 978-1-61208-100-7

proportional to its performance. A MC whose nodes are highly heterogeneous, comprising of a combination of various resources, requires an efficient location policy to determine a suitable node as a destination for the migrating process.

Distributed scheduling related to process migration [3] decides on when to migrate, which process, and where to migrate. Some popular distributed scheduling policies like sender-initiated, receiver-initiated, random policies are not apt for resource-constrained mobile cluster. We propose a novel approach for minimizing such overheads using Peer-to-Peer (P2P) systems. Such systems share computer resources by direct exchange, rather than requiring the intermediation or support of a centralized server or authority. In our work, this property of P2P systems is exploited to implement a new location policy for dynamic process migration.

We begin by building a hierarchical structured P2P cluster of static workstations and mobile devices. The overlay network thus arrived by employing Distributed Hash Table (DHT) concept helps in linking resources (nodes) and enables easy sharing of processing among them. Traditionally Distributed Hash Table (DHT) is used only to find and share content / file / data only. In our work, we use the DHT to assist a mobile node find a destination node (to which it can off-load its current processes) when it voluntarily gets disconnected. A binary heap (max-heap) is built and a selection algorithm is written to obtain the node with maximum spare computational capacity. The contributions of this paper can be summarized as follows.

1. We propose a process migration policy for a dynamic MC built on a cellular network when it is processing a long-running, compute-intensive application.

2. While migrating processes we devise a more accurate mechanism for dynamic load-balancing considering the variations in computing power of the different mobile nodes.

3. A DHT-based approach for location policy on where to migrate is described.

4. Considering the resource-poor nature of the mobile device, we have formulated an algorithm for a node to accept a migration request or not.

5. We show how large prime numbers can be generated on a mobile cluster while striving to harness the idle time of the nodes. Also we demonstrate an efficient programming way of storing such large numbers in memory-poor mobile nodes.

This work does not spell any mechanism for PM, but a facility that reduces the time taken for PM in a very dynamic network and also increases the throughput of the system. We also restrict our work to homogeneous PM in which we migrate between nodes of same architecture. This restriction is reasonable as our concern is not with the mechanism but with the policy of when and where to migrate the process. The rest of the paper is organised as follows. Section 2 presents the related work in this area of research. In section 3 we provide the background and also highlight the rationale behind building a mobile P2P cluster. Section 4 provides system analysis and design of the mobile cluster. Section 5 accounts for the test application of large prime generation and a comparison of performance of the system with existing ones. Section 6 concludes and gives an insight into the future enhancements.

II. RELATED WORK

PM is extensively surveyed in [3, 4], which present a number of approaches. Some recent efforts on PM include [5] in which the authors describe the use of system-wide pointers and global dynamic data structures for migration. Gobelins DSM [6] moves processes or threads among cluster nodes using the distributed shared memory concept. But it again does not deal with mobile devices. Checkpointing for mobile computing systems has been discussed in [7, 8, 9]. An evaluation of different checkpointing protocols is done in [10]. One of the factors that amount to the cost of PM in a dynamic network is the time taken for deciding on the destination. All previous works on PM only focus on time taken for state transfer, but not for locating the destination. In [11], the issue of timeliness for rerouting and multicast when handoff occurs in a MC is discussed. A model for overcoming such issues in MC is presented in [12]. P2P networks [13] are self-organising structures apt for realising such clusters. Some important DHT-based P2P systems are found in [14 – 16] and are focused on fixed networks only. In [17], the authors propose a load balancing scheme for heterogeneous cluster using mobile agents. An algorithm for load balancing in heterogeneous dynamic P2P systems using the concept of virtual servers is presented in [18]. P2P networking in mobile environment is explored in JXME [19]. In [20], the mobile devices are assumed as low-performance nodes and hence only used to redirect their requests to their associated static nodes. However, in reality we cannot afford to keep mobile nodes without any useful processing. Hierarchical P2P systems are said to improve scalability. Such systems are discussed in [21, 22, 23]. Heaps based on the concept of a complete binary tree, are a good choice for implementing selection and priority based algorithms. A distributed heap-based data structure called CONE [24] has load balancing properties and is layered on Chord DHT. To the best of our knowledge, none of these works focus on a process migration facility for mobile clusters. Thus, this work is a new attempt at such a design.

III. BACKGROUND

A. *Problem Definition*

The fact that the mobile devices are becoming powerful in terms of computing cannot be overlooked. This calls for mechanisms to run high-end applications on such devices while taking care of their inherent mobility. The mobility issue can be seen in two perspectives. Firstly, the devices voluntarily disconnect themselves or move around; thus they leave or join the network / cell anytime. Secondly, disconnections can occur suddenly or the device itself may fail suddenly. Voluntary disconnection forces the termination of long-running applications. The problem is severe when such applications get terminated especially when they are nearing completion. PM eliminates this problem whereby the mobile node transfers the non-static part of the computation

 ISBN: 978-1-61208-100-7

state to the BS and the BS finds another compute node in its cell to resume the computation.

In systems with only static nodes, there are a number of implementations for PM. Nevertheless, there should be a justification between the choice of running a process from the start after disconnection and PM because the time taken for PM is not always negligible. It includes the time taken for choosing a destination node with sufficient capability while maintaining load balancing and transferring the process state to such a node. In static networks the time for negotiations between the nodes to choose a destination is generally a constant depending on the type of the network and therefore very negligible. To choose a destination node in mobile clusters, techniques like polling, random or nearest neighbor selection policies are not applicable or even irrelevant because of inherent mobility and other constraints posed by mobile nodes. This work attempts to define a dynamic resource location policy so as to choose an eligible and efficient destination node in a highly dynamic and heterogeneous network.

B. *Mobile P2P Cluster*

In designing a mobile cluster, P2P system is an attractive architectural alternative to the traditional client-server computing. It solves the problem of server being down or becoming overloaded. Therefore a number of critical, real-time, computationally high-end applications can be successfully implemented on a mobile P2P cluster. Further, P2P network is self-organizing, which is a key advantage for a dynamic network. They offer efficient search/location of nodes and load balancing facilities. These characteristics make P2P network a good choice for performing process migration.

IV. SYSTEM DESIGN

A. *Basic Model*

The mobile cluster (MC) contains a set of mobile nodes (MN) and static nodes or mobile support stations called as Base Stations (BS). Static wired network connects BSs to each other whereas a cellular network connects the MNs. In such a network, there are multiple cells and each cell has multiple channels to communicate with many mobile nodes. Also each cell is equipped with a BS that governs multiple MNs in that cell. Any MN stays in connection with at most one BS at any given time and communicates with other MNs and BSs only through the BS to which it is currently connected.

B. *Building a P2P Mobile Cluster*

We consider the mobile cluster to adopt a Super-Peer network model. In this model, the BS is categorized as super-peer. A typical distributed application is submitted to the cluster through a designated coordinator that is nothing but the BS. The BS distributes the application to the peers. The peers do not choose which processes to host. The processes are allotted to peers depending on their capability and current computational load. The peers periodically inform and update their load information to their super peer, the BS. Fig. 1 depicts BS as super-peer and all other nodes organized as a heap.

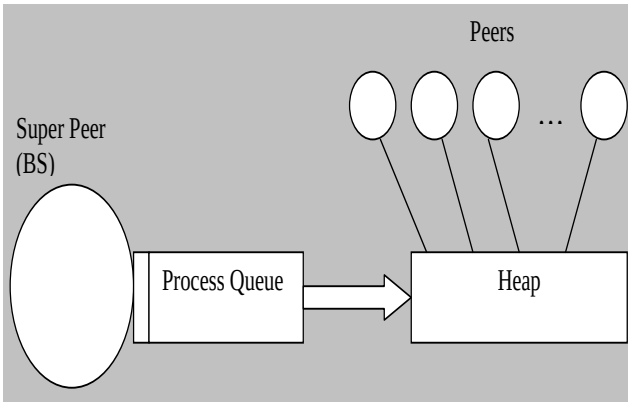


Figure 1. Super Peer and Peers

1) *DHT Design:*

When a mobile node sends a migration request to BS, the BS as a super-peer determines the destination on behalf its client (the source). In designing such structured networks, Distributed Hash tables (DHT) are employed. They are distributed data structures for building robust P2P applications. Conventional DHT maps keys to values, store key/value pairs and retrieve values using the given keys and are used in general for file storage and sharing. In contrast, in our work, every node in the mobile cluster stores a hash table and the resultant DHT performs two functions: (i) organizes cluster nodes as a max-heap and (ii) distributes processes to various nodes considering their current load and capability.

Each node in the cluster is known by a 128-bit identifier that is unique in the cluster. Each process is assigned a unique identifier that remains unchanged even when migrated. On applying consistent hash function, we allow nodes to join and leave the cluster with minimal disruption. The hash function determines the node identifier by hashing the IP address of the node and a key for each process by hashing the process identifier. The node identifiers are arranged in the form of a max-heap. A max-heap is a complete binary tree in which at every node the data stored at the node is no less than the data at either child. The heap is constructed based on the spare capacity S_i available at each node i. We show in Section 4.2.2 on how to calculate this spare capacity. Thus the root node always has maximum spare capacity. The DHT supports basically three operations, namely join, leave and migrate. The heap is reconstructed whenever a node joins/leaves or a process migrates and is done by heapify(). heapify() maintains the max-heap property that the spare capacity of parent node is always more than that of its child nodes.

The node that joins the cluster is added to the bottom of the heap, keeping up the shape of a complete binary tree. If it has more spare capacity than its parent, it is swapped with it. Then the node continues to move up until it finds a place so as to maintain the heap property. The node before leaving the cluster initiates migrate operation forwarded to the BS. Also

 ISBN: 978-1-61208-100-7

when a process migrates to a new node, that node now has less spare capacity and therefore moves down in the heap. The nodes maintain information that enables communication to their neighbor nodes. That is, except the leaf nodes, all other nodes keep an account of their child nodes.

The super peer maintains a heap manager (HM) which is responsible for keeping up-to-date information on the current load and spare capacity of every node in the cell. Every node that enters or leaves a cell causes an imbalance in the load and the spare capacity of all other nodes is bound to change and this is updated by the HM through appropriate function calls. At any point in time in a heap it is always the root node that has maximum spare capacity. This facilitates in searching a destination node among n peers in O (1) time; joining and leaving of nodes consumes O (log N) time where N refers to total number of peers.

2) Capability of Heterogeneous Nodes:

We illustrate an effective way of computing the capability of nodes in the cluster and thereby determine the potential candidate to take up the migrated process. A benchmark program is run on all the N nodes of the cluster. The execution time is recorded for every machine, say e1, e2, eN. Since the power of the node is inversely proportional to the execution time, we take the largest ei. Compute its inverse to make it our unit. We divide the inverse of every other ei by this unit. This gives the relative power of the various nodes of the cluster. For example if the execution times are say: 100 ns, 500 ns, 1 micro sec, 5 micro sec, and 10 micro sec. The inverse of 10 micro sec is 100,000. The computer that gave the result in 100 ns has an inverse of 10,000,000. On dividing 10,000,000 by 100,000 we get 100. Thus we say that the first machine is 100 times more powerful than the last machine.

We find the current load on the processor and find out the spare capacity Si of the node i by subtracting current load from 1. If the CPU utilization is k%, it means (100 - k) % is available for the task to be added. The fact that the utilization is less than 100% indicates that we could add a task to the mobile node. The added task would then use the spare CPU time, without degrading the performance of the system. For example, let the CPU utilization be 0.999 for the node which is 100 times more powerful and that of the slowest machine be 0.1 The fastest machine now has a spare capacity of (1 - 0.999) x 100 = 0.001 x 100 = 0.1. The slowest machine has a spare capacity of (1 - 0.1) x 1 = 0.9. Thus given the current load situation, the slowest machine is 9 times more powerful than the fastest machine. Accordingly the loads are assigned so that both machines complete their assigned tasks at about the same time. In [25], the authors propose Horse power utilization (HPU). Our approach differs from HPU in two ways. Firstly, in HPU the test is done once and the results are used again and again. But here we check the CPU activity register to test for CPU occupancy every time we migrate a process. Testing CPU occupancy of course delays the process migration. However, the decision would be more accurate as we do not model the performance of a very complex system like HPU, but measure it. Secondly, the HPU approach is based on the assumption that the relative power of heterogeneous systems reacts the same way under all load conditions, which is not really true. In our approach, for every migration request we query the CPU utilization and use it to calculate the relative power.

3) Status of the Mobile node:

The queue of migrating processes is held at the processor controlling the base station. We are desirous of removing the waiting process to a free or less loaded mobile node. Since the mobile nodes are resource-poor, we check for their readiness to accept a migrated process or not. As regards the checking of the status of a mobile node whether it could accept a migrating process or not, the decision could be as follows.

1. A time interval is decided for the mobile nodes to inform their current base station of their status on current load, say once every second. As long as the queue of migrating processes is small compared to a predetermined size, the reporting could be once every second, which means once every billion instructions or so.

2. When the migrating process queue becomes too big reduce this time from 1 second to 0.5 second and broadcast this change to all mobile nodes connected to the base station.

3. Continue decreasing the delay between successive reporting until the queue becomes smaller. In case the reporting activity occupies more than a predefined percentage of the processor time, say 5%, we suspend the migrating process and reactivate when the load decreases.

4. When we find that the migration queue is smaller than a predetermined size, we increase the delay again to 1 second between successive reports from the connected mobile nodes.

The above process is dynamic balancing the need for higher efficiency at the mobile nodes against the queue size of the migrating processes.

C. Process Migration

Before the MN signs off, it initiates a daemon that has two responsibilities: one, to inform BS the intention of the node to leave the cluster and second, to save the process image as an object with the process information available in task_struct (in Linux kernel). Using object serialization, the process state is transferred to the destination via the BS. The destination node i will be chosen based on its relative power (Pi) and current CPU load (Li) obtained periodically. The BS receives periodically the Li information from its peers as discussed above.

The algorithm is summarized with the various actions that take place in the following three entities:

(i) At Super-peer:

- Receives migration request from the node that is going to sign-off.
- Determines root node as destination in heap.
- Pro actively sends code and static data to the destination; Receives from MN and transfers process state and dynamic data to destination.

(ii) At mobile node (source):

- Before signing-off: Issues migration request to its BS.

- Checkpoints and saves it current state; Transfers process state to BS.
- On return to its previous cell / another: Gets assigned new process; Starts execution.

(iii) At static / mobile node (destination/root node in heap):

- Receives migration request from its BS
- Receives process state, code and data
- Resumes execution of the migrated process
- Process gets executed on this node until completion; else if the node leaves the cluster the heap gets reorganized and its child node now becomes the root node. The request is passed onto the new root node and process repeats until completion.

V. PERFORMANCE COMPARISON

A. Prime Number Generation

Generating very large primes it is a compute-intensive application. Let us assume that we assign 1 million numbers to each partition. The partition upper bounds then are 1M, 2M, 3M, ... At the beginning of this sequence the time taken by the different partitions would be appreciably different from each other. However, the time difference would decrease as the upper bound of the partition becomes much larger. An important issue is to resolve the memory constraint of nodes while dealing with such huge numbers. If we consider the memory requirement in terms of digits or bytes, it might increase linearly as the range moves away from 1. This is because the number of digits keeps increasing as the range moves away. Even here, since the digits change only over the first lowest weighted 6 digits, the higher value digits being common, a clever programming trick would store only the lowest 6 digits for every prime and store the higher value common digits in a separate place once. Then the memory would remain constant. For example, all the prime numbers between 123000000 to 123999999 are of the form 123xxxxxx, where the xxxxxx alone need to be stored. The required prime is generated by appending 123 to xxxxxx at run time. A similar approach is used in clusters to save the memory access time and can be found in [26].

B. Comparison

The two tasks that make PM time-consuming are the time to negotiate and choose a destination node and the state transfer from source to destination. Location policy in distributed scheduling determines the destination node for the migrated process, and some of these policies include polling, random and nearest neighbor algorithms. These techniques when employed for systems with static nodes have proven performance based on the system workload. However, when applied to mobile clusters these strategies incur more overheads and also sometimes not applicable at all. For example, polling involves checking for the status of a node to accept the migrated process and continues the same until a suitable node is found. Here the best case can be the first node that is approached and the worst case is the node that is finally approached. This strategy causes increased communication cost in the worst case. Further, a node which indicated its willingness to be the destination may choose to sign off later to conserve energy or because it simply chooses to move away. In this case the mobile node needs to repeat the process of polling. Also any communication from a mobile node to another is via the BS and this amounts to a huge overhead. The same issues are true in the case of random algorithms.

The third strategy of contacting the nearest neighbor is again not applicable for a mobile cluster. Here it is difficult to determine the nearest neighbor and also if we do so it will only be an estimate and not an accurate one. Moreover as mobile cluster is very dynamic, it is difficult and time-consuming to determine the right destination. In comparison with these existing approaches, our proposed model of a heap-based DHT approach takes very negligible time to fetch the destination as well as maintains load balance. This is made possible because as the nodes join and leave we organize them in the form of a heap based P2P cluster. Further, using appropriate benchmarking we estimate the power and also with the current CPU utilization rates, the heap gets reorganized. Now as in existing approaches the mobile node does not correspond with any other nodes for their availability. The root node in the heap is simply chosen as the destination node. Even if this node moves, the heap gets reorganized and the next root node is tried. Thus this model chooses the destination node with ease. In the following table, we provide a comparison of some existing systems with ours in terms of various features.

TABLE I. COMPARISON WITH EXISTING SYSTEMS

S.No.	System	Scheduling (Centralized / Distributed)	Load-balancing	Location Policy
1.	Heap	Distributed (P2P)	Continuously and Dynamic	Root node of the Heap
2.	Sprite	Centralized	Only during creation or eviction of a process	History of Idle Time Length
3.	Condor	Centralized & Priority-based	Only during creation or eviction of a process	Polling
4.	Mosix	Centralized	Continuously and Dynamic	Decentralized Load Vector with Load information on a set of random nodes
5.	MPVM	Centralized	Only during creation / eviction of a process and very high load on a node	Idle node availability

The task of transferring the process state is equally challenging and those adopted for static nodes again are not suitable for mobile clusters. The eager (all) approach is used by checkpoint/restart implementations in which the entire process state is transferred at a time to the destination. If we apply this approach in mobile cluster, the transfer is via the BS and there is a possibility that the power-constrained mobile device may go off while the transfer is going on. The eager (dirty), copy-on-reference, pre-copy and flushing strategies again are not applicable because of the dynamic nature, resource-constraints and heterogeneity prevailing in mobile clusters. Our model takes care of these issues by doing two tasks in parallel: as soon as the migration request is received by the BS, it chooses the destination and pro actively sends the static data and code; at the same time the remaining process state alone is extracted from the source and sent to the destination by means of object serialization, reducing the time taken for transferring the process state. Conclusion

The motivation to migrate a process in a mobile cluster is manifold. There are various components that contribute for the longer time taken to migrate a process. By organizing the cluster as a P2P network with nodes arranged as a heap topology, the time taken can be considerably reduced. Normally a migration request is issued to a remote node and only after negotiation, we decide to move the process. This has been avoided since the BS quickens the process migration by choosing the root node from the max-heap. Also, once the migration request is received, the BS sends the constant data and code to the destination even before the process state is dispatched from the source to the BS. This work can be extended for process migration among a finite collection of clusters and thereby a computational grid. Further we can consider migration in a heterogeneous environment because in a mobile cluster the likelihood of heterogeneity among nodes is more.

REFERENCES

[1] B.R. Badrinath, A. Acharya, and T. Imielinski, “Impact of Mobility on Distributed Computations,” ACM SIGOPS Operating Systems Review, vol. 27, no. 2, April 1993, pp. 15-20.

[2] James J. Kistler and M. Satyanarayanan, “Disconnected operation in the Coda file system,” ACM Transactions on Computer Systems, vol. 10, no. 1, Feb. 1992, pp. 3-25.

[3] D.S. Milogicic, F. Douglis, Y. Paindaveine , R. Wheeler , and S. Zhou , “Process Migration,” ACM Computing Surveys, vol. 32, no. 3, Sep. 2000, pp. 241-299.

[4] M. Singhal and N.G. Shivaratri, Advanced Concepts in Operating Systems, 2001, McGraw Hill.

[5] K. Noguchi, M. Dillencourt, and L. Bic, “Efficient Global Pointers with Spontaneous Process Migration,” Proc. 16th Euromicro Conference on Parallel, Distributed and Network-Based Processing (PDP 2008), 2008, pp. 87-94.

[6] G. Vall´ee1, C. Morin, J. Berthou, Ivan D. Malen, and R. Lottiaux, “Process Migration based on Gobelins Distributed Shared Memory,” Proc. Workshop on Distributed Shared Memory (DSM'02), held in conjunction with CCGRID 2002, Germany, May 2002, pp. 325-330.

[7] R. Prakash and M. Singhal, “Low-Cost Checkpointing and Failure Recovery in Mobile Computing Systems,” IEEE Transactions on Parallel and Distributed Systems, vol. 7, no. 10, October 1996, pp. 1035-1048.

[8] B. Gupta, S. Rahimi, and Z. Liu, “A New High Performance Checkpointing Approach for Mobile Computing Systems,” IJCSNS International Journal of Computer Science and Network Security, vol. 6 no. 5B, May 2006.

[9] L. Chen, Q. Zhu, and G.Agrawal, “Supporting dynamic migration in tightly coupled grid applications”, Proc. ACM/IEEE Conference on supercomputing (SC’06), 2006, pp. 28.

[10] A. Agbaria and R. Friedman, “ Model-based performance evaluation of distributed checkpointing protocols,” Performance Evaluation, vol. 65, no. 5, 2008, pp. 345-365.

[11] H. Zheng, R. Buyya, and S. Bhattacharya, “Mobile cluster computing and Timeliness issues,” Informatica, vol. 23, 1999, pp. 5-17.

[12] M.A. Maluk Mohamed, A. Vijay Srinivas, and D. Janakiram, “Moset: An anonymous remote mobile cluster computing paradigm,” Journal of Parallel and Distributed Computing, vol. 65, 2005, pp. 1212 – 1222.

[13] E.K. Lua, J. Crowcort, M. Pias, R. Sharma, and S. Lim, “A survey and comparison of Peer-to-Peer overlay network schemes”, IEEE Communications Surveys and Tutorials, March 2004.

[14] S. Ratnasamy, P. Francis, M. Handley, R. Karp, and S. Shenker, “A scalable content addressable network,” Proc. ACM SIGCOMM, 2001, pp. 161–172.

[15] I. Stoica, R. Morris, D. Karger, M.F. Kaashoek, and H. Balakrishnan, “Chord: A scalable peer-to-peer lookup protocol for internet applications,” IEEE/ACM Transactions on Networking, vol. 11, no. 1, 2003, pp. 17–32.

[16] B.Y. Zhao, L. Huang, J. Stribling, S.C. Rhea, A.D. Joseph, and J.D. Kubiatowicz, “Tapestry: A resilient global-scale overlay for service deployment,” IEEE Journal on Selected Areas in Communications, vol. 22, no. 1, January 2004, pp. 41–53.

[17] M. Abdallah and E. Buyukkaya, “Fair Load Balancing under Skewed Popularity Patterns in Heterogeneous DHT-Based P2P Systems,” Proc. IASTED International Conference on Parallel and Distributed Computing and Systems (PDCS), Cambridge, Massachusetts, USA, November 2007, pp. 484-490.

[18] B. Godfrey, K. Lakshminarayanan, S. Surana, R. Karp, and I. Stoica, “Load Balancing in Dynamic Structured P2P Systems,” Performance Evaluation, vol. 63, no. 3, March 2006, pp. 217-240.

[19] A. Arora, C. Haywood, and K.S. Pabla, “JXTA for J2ME – Extending the Reach of Wireless with JXTA Technology,” Sun Microsystems, March 2002.

[20] S. Zoels, S. Schubert, W. Kellerer, and Z. Despotovic, “Hybrid DHT Design for mobile environments,” Proc. AP2PC Workshop at AAMAS 2006, Hakodate, Japan, May 2006.

[21] G. Erice, E.W. Biersack, K.W. Ross, P.A. Felber, and G.U. Keller, “Hierarchical Peer-to-Peer Systems,” Proc. ACM.IFIP International Conference on Parallel and Distributed Computing (Euro-Par), 2003.

[22] B. Yang and H.G. Molina, “Designing a Super-Peer Network,” Proc. International Conference on Data Engineering (ICDE), 2003.

[23] I. Rimac, S. Borst, and A. Walid, “Peer-Assisted Content Distribution Networks: Performance Gains and Server Capacity Savings,” Bell Labs Technical Journal, vol. 13, no. 3, Fall 2008, pp. 59-69.

[24] Bhagwan, R., Mahadevan, P., Varghese, and G., Geoffrey M Voelker, “CONE: A Distributed Heap-Based Approach to Resource Selection,” Technical Report CS2004-0784, UCSD, 2004.

[25] R.K. Joshi and D. Janaki Ram, “Anonymous Remote Computing: A Paradigm for Parallel Programming on Interconnected Workstations,” IEEE Trans. Software Eng., vol. 25 no. 1, 1999, pp. 75-90.

[26] S. Hwang, K. Chung, and D. Kim, “Load Balanced Parallel Prime Number Generator with Sieve of Eratosthenes on Cluster Computers,” Proc. 7th IEEE International Conference on Computer and Information Technology (CIT 2007), 2007, pp. 295-299.

 ISBN: 978-1-61208-100-7

EaST: Earth Seismic Tomographer

Building a network of volunteer smart mobile devices for seismic travel times Earth tomography

Ida Bifulco, Rita Francese, Ignazio Passero and Genoveffa Tortora
Dipartimento di Matematica e Informatica, DMI
University of Salerno
Fisciano (SA), Italy
ibifulco@unisa.it, francese@unisa.it, ipassero@unisa.it, tortora@unisa.it

***Abstract* — The recent diffusion of smart mobile devices deeply influences current technological landscapes also supported by a blooming market economy. New forms of users, interaction styles and ubiquitous paradigms are growing with this technological revolution. Connected mobile devices equipped with accelerometers represent, for geological Research, the opportunity to increase the information on several phenomena that are difficult to study because of the limited availability of observation data. Information provided from a cloud of mobile sensors, randomly localized on the territory, may contribute to extend the fixed nature and the very limited number of traditional Earth observation points. This paper describes the Earth Seismic Tomographer system, a mobile application that analyzes triaxial accelerometer data, aiming at collecting travel time information on earthquake events. The system relies on the voluntary participation of users that devote personal mobile resources to detect and provide seismic data to the central server. The Earth Seismic Tomographer adopts neural network classifiers to separate user movement from the seismic signal. The proposed system will increase the amount of information on seismic events enabling Earth scientists to study problems still undetermined with the currently available data.**

Keywords - Mobile Smart Devices, Earthquakes, Seismic Tomography, Accelerometer Sensing, Neural Networks.

I. INTRODUCTION

Technological market and progress are depicting novel scenarios in which users and applications exploit increased degrees of connectivity and ubiquity. In several field of parameter estimation [2], for example, it would be indispensible for researcher to exploit the increased amount of data provided by new mobile smart mobile devices.

Indeed, new generation mobile devices are, almost always, equipped with accelerometers, orienteer and camera and often are GPS localized or, at least, estimate their position by triangulating GSM cells or WiFi repeaters.

This paper presents the Earth Seismic Tomographer system (EaST), a mobile application that aims at providing geologists with a redundant amount of data on seismic events by recording ground acceleration data.

The problem of estimating physical parameters values from experimental data is a crucial matter in many geophysical investigations. In geophysical literature, this problem is denoted as model inversion and the goal is to combine information arising from physical theories and from experimental results, in order to infer some characteristics of given Earth properties. In general, theories are represented by a set of equations relating values of unknown properties to the physical parameters observed in the experiment.

In particular, *seismic tomography* [13] is a technique aiming at reconstructing the velocity structure ($s(r)$ in equation 1) of a body, given the measurement of travel times (T in equation 1) of waves that have propagated through that body.

$$T = \int_{r[s]} s(r)ds \qquad \textbf{(1)}$$

In this expression, s is the slowness and is defined as the reciprocal of the velocity: s = 1/v. The slowness, used instead of the velocity, keeps the integrand linear respect to the quantity we aim at retrieving.

It would be tempting to conclude from (1) that the relation between the travel time and the slowness is linear. However, this is wrong because the integration in (1) is along the path on which the waves travel. The rays are curves of stationary travel time, and hence the ray location depends on the slowness as well.

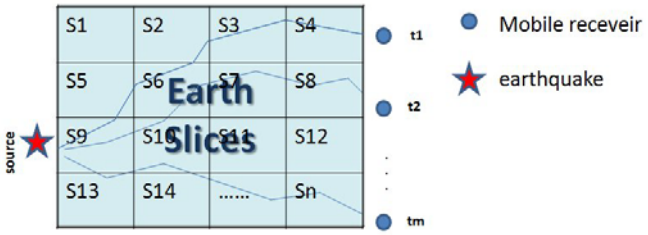


Figure 1. The discrete slowness model for seismic tomography.

Seismic tomography is thus a nonlinear inverse problem: the unknown slowness is present in (1) both in the travel time integral and in the integrand, where it determines the ray position $r[s]$.

When one divides the earth-model in cells where the slowness perturbation is assumed constant, the discrete form of (1) can be written as:

$$\delta T_i = \sum_j l_{ij} s_j \qquad \textbf{(2)}$$

 ISBN: 978-1-61208-100-7

In this expression, the subscript i labels the different travel times used in the inversion, while j is the cell index and l_{ij} is the length of ray i through cell j. Figure 1 depicts this discrete version with ray paths, homogeneous slowness cells, seismic source and receivers.

In the seismic tomography problem, the aim is to reconstruct, using (2), the Earth velocity model ($V_i=1/s_i$ for $i=1, \ldots, n$) from a set of travel time measurements. In the ideal case, an exact theory exists that prescribes how the data should be transformed in order to effectively reproduce the model. For some selected examples such a theory exists assuming that the required infinite and noise free data sets would be available.

Nevertheless, the model that we aim at determining is a continuous function of the space variables. This means that the model has infinitely many degrees of freedom while, in a realistic experiment, the amount of data that can be used for the determination of the model is usually finite. A simple count of variables shows that the available data do not carry sufficient information to determine uniquely the model. The fact that in realistic experiments a finite amount of data is available to reconstruct a model with infinitely many degrees of freedom necessarily means that the inverse problem is not unique in the sense that many models explain the data equally well. The model obtained from the inversion of the data is therefore not necessarily equal to the true model to seek.

EaST project aims at improving the availability of seismic data, providing more information about seismic events, about their effects on the territory and contributing to improve Earth phenomena models. In the case of a seismic event, a network of volunteer sensing devices sends toward the server the temporized acceleration tracks. The server offline elaborates these tracks to detect the seismic rays travel times providing a new amount of information for the Seismic tomography and other geophysical problems.

The remainder of the paper is organized as follows: Section II presents and discusses similar research approaches; Section III details the system, while Section IV concludes the work.

II. RELATED WORK

This section resumes the state of the art related to seismic signals detections and the approaches aiming at automatically understanding user movements from accelerometer data.

The earthquake detection is the goal of the *Quake-Catcher Network* project [16]. It is a collaborative initiative for developing the world's largest, low-cost strong-motion seismic network. The QCN is collecting a great amount of data on seismology by combining new Micro-Electro-Mechanical Systems technology with volunteer seismic stations distributed computing. The system has a twofold nature since it utilizes two kinds of sources:

- laptop built in sensors (the authors refer that only Apple and ThinkPad laptops offer this feature),
- specific sensing components attached to internet-connected computers (that need to be bought and configured).

QCN originated from an idea of Elizabeth Cochran [3], a seismologist at UC Riverside. The project started in 2006, and is operative since April 2008, now being one of the largest and densest earthquakes monitoring system.

Actually, QCN is the only system similar to EaST in terms of expected diffusion and limited cost. Obviously, there are other specific Earth monitoring networks, but they are based on a small number of accurate but geographically concentrated observation points. Differently from Cochran et al., we completely rely on ordinary mobile devices and on already available hardware sensors to increase the amount of information on seismic events enabling Earth scientists to study problems still undetermined with the currently available data. Additionally, in our case, the adoption of mobile devices causes the perturbation of seismic data with user movements and it is necessary to adopt some detection mechanism.

The neural networks (NNs) are software classifiers inspired to biological nervous system [10]. NNs simulate the human brain activities with a self-adapting system composed of simple elements, the neurons, connected with each other and operating in parallel. The theory of NNs has a broad application sphere in several fields including pattern recognition, identification, classification, speech, vision, and control systems. The main interesting feature of NNs is in their self-adaptation mechanism that can solve problems that are difficult for conventional computing or human beings. As in nervous systems, the connections between elements determine the overall network function. The training phases of a network adapt it to the desired function by adjusting the values of the connections between neurons.

Earthquake detection and classification belong to a class of problems where artificial NNs may be useful [4].

Indeed, the most important characteristic of a NN is its capability to learn from examples, so that NNs are powerful tools in approaching problems, like the earthquake related class, that are difficultly described using a classic algorithmic strategy.

In [17], Romero proposed two NN applications at earthquake detection problem: a simple NN (Perceptron [18]) classifying earthquakes recorded at the Bardonecchia (North Italy), and an auto associative NN that has proved useful to build an adaptive neural trigger for earthquake detection.

Sharma et al. present and evaluate the precision of several approaches for detecting seismic event signals in presence of background noise [20]. In their work, the authors examine several trigger algorithms, ranging from a very simple amplitude threshold type, to sophisticated pattern recognition, adaptive methods and NN based approaches. All the detection triggers have been tested on natural events and on artificial ones (e.g., underground nuclear explosions).

In our case, we do not have previous knowledge about the sensitivity of devices and their response to seismic events, but we can exclude perturbations due known user movements. As a dual approach of previous ones, the

 ISBN: 978-1-61208-100-7

adoption of NNs let us filter out, on the client device, a good number of false positive detections by training neural classifiers on typical user movements.

In [11], user activity classification is performed analyzing acceleration magnitude and rate of change, as well as piecewise correlating the three components of acceleration. A simple Multi-Layer Perceptron is trained for classification.

Fabian et al., train a set of NNs to recognize six typical human activities: resting, typing, gesticulating, walking, running and cycling [5]. They collect body-part acceleration values reading three MotionBand devices attached to the test subjects. In our case, we have a single accelerometer moving with the user, as in [22] where Yang et al. adopt a multilayer feed-forward NN as activity classifiers. They propose an effective activity recognition method using acceleration data collected from a single triaxial accelerometer.

III. THE PROPOSED SYSTEM

EaST is a client server system aiming at collecting travel time information on earthquake events. Each client is a mobile receiving station and sends seismic acceleration data to the central server. To prevent false positive seismic triggers due to user movement, several NNs, trained on typical user activities, detect perturbations that can interfere with the seismic detection. The configuration of clients is fully dynamic: the server stores all configuration data and provides updates to all clients. It is always possible to change the client behaviour simply acting on the server, an easy and user transparent operation compared to the distribution of a new release of the client. This runtime dynamic configuration is not limited to simple processing parameters but also to the architecture and the configuration of adopted NNs. As NN engine, EaST system embeds Joone [9]. The core engine of Joone is suitable for small devices, having a small footprint and is executable on Personal Java environments. The framework enables to serialize a NN object (structure and weights) to a file. On update, the server propagates new NNs to each client.

Actually, the client application is only available for Google's Android, an open source Linux based platform for mobile devices [1], but we foresee to develop EaST client versions for other mobile platforms. The Android architecture has been developed by the Open Handset Alliance [15], a federation of device manufacturers, mobile operators and other companies (i.e., semiconductors manufacturers, software developers, etc.). The Alliance ensures that the same implementation of the software is suitable for all devices that support the Android software stack. This feature greatly improves diffusion expectations of EaST system.

An interesting feature of Android is the availability of free API libraries for controlling the device hardware. The Android SDK includes APIs for location-based hardware (such as GPS), camera, network connections, Wi-Fi, Bluetooth, accelerometers, touch-screen, and power management [1].

Accurate location and time for each client are crucial for the effectiveness of measurements. At this aim, EaST clients interrogate the Android Location service searching for GPS localization that is affected by an error lower than 20 m. As an alternative, if GPS is not available (often in indoor locations) clients use the network localization, whose precision is of some hundreds of meters. In any case, the error esteemed for the localization is communicated to the server. In the case GPS service were available, the time provided by satellites is used also for periodically synchronizing the client clock. Alternately, the time is periodically synchronized with the Network Time Protocol time as in [3], with a verified a precision of 20 ms [6].

A. *The instrument and the phenomenon sensitivity*

To tune and test the system, we use mobile devices equipped with a Qualcomm processor working at 528 MHz, 512 MB ROM and 288 MB RAM, an internal GPS receiver, a G-sensor accelerometer, a digital compass, a 320x480 pixel touch-screen and an opposite 3.2 megapixel integrated camera. As additional memory, the devices support SD cards up to 16 GB. In particular, the devices have an AK8976A 3-axis accelerometer, even if the SDK ensures the application to be suitable for all Android devices that are equipped with a similar low cost hardware component.

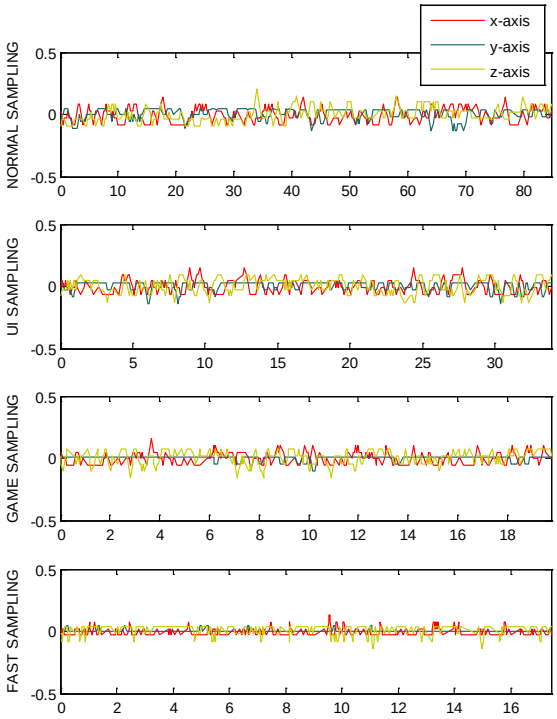


Figure 2. The noise affecting accelerometer for 300 measurements at different sampling rates, on a time scale in seconds.

The Android SDK lets programmers to choose between four sampling frequencies with a progressively decreasing sampling period: NORMAL, UI (i.e., User Interface suitable sampling rate), GAME and FASTEST. It is important to point out that Android system is a multitasking one and, therefore, the sampling frequency is not constant.

The coordinate space adopted for acceleration sensing is the OpenGL ES [14] coordinate system. When the phone lies

on a table with upward screen, the origin is in the lower-left corner with respect to the screen, with the X axis horizontal and pointing right, the Y axis vertical and pointing up and the Z axis pointing outside the front face of the screen. In this system, coordinates behind the screen are characterized by negative Z values. The axes are not swapped when the device's screen orientation changes.

All values are expressed in *IS* units (m/s^2) and measure the acceleration applied to the phone minus the force of gravity.

If the device lies flat on a table, the force of gravity acts only on the z component of the sensed acceleration. When the device is in a jacket pocket of a user, like in reported samples, the acceleration component A_y is decreased with the force of gravity (- 9.81 m/s^2). The A_x and A_z components measure, respectively, a lateral acceleration and a front rear acceleration respect to the user.

Before starting the development of the system, we sampled the sensor noise and sensitivity at different rates. The objective of the feasibility study was twofold: verifying if earthquakes propagation time enables to obtain meaningful measures from devices spread on the territory and that the phenomena are perceptible in a geographic area wide enough respect to the temporal resolution of measures.

Figure 2 plots the accelerometer values, when the phone is in state of rest, sampled at all the available frequencies. In particular, 300 measurements are represented respect to time in seconds, and increasing frequencies are compared. The g_z component of the force of gravity has been subtracted from the A_z measure for better graphically comparing all the components on a smaller scale. By observing Figure 2, it is possible to note that the FAST SAMPLING subplot appears as the less affected by measurement noise. For this reason, and to obtain the best signal resolution possible, we adopted the fastest frequency available for sampling the acceleration. Even in the fastest sampling case, produced track files are thin respect to modern microSD cards and network transmission rates: 1.70 Mbytes are sufficient to record more than 15 minutes of signal and can be compressed in less than 300 Kbytes. As depicted, the error results lower than 0.2 m/s^2 = *2%g*. These values enable to detect a seismic signal from a wide geographic area surrounding the source.

TABLE I. A SAMPLE OF PEAK GROUND ACCELERATIONS FOR SEISMIC EVENTS

ACCELERATION (%g)					
SOURCE	**M**	**Depth Km**	**10 Km**	**50 Km**	**100 Km**
1	5.0	10	6	1	//
2	6.0	57	20	5	//
3	6.2	60	2.1	1.0	0.5
4	6.8	40	9	4-7	1
5	6.2	10	20	7	3
6	6.8	35	30	10	4
7	6.9	11	40	10	//

Table I resumes the peak ground acceleration measured during some strong seismic events and obtained analyzing the USGS (U.S. Geological Survey) maps [19]. The reported measures of peak ground acceleration are expressed as a percentage of g (gravity acceleration).

Comparing a sample of seismic ground acceleration values (Table I) with current instrument sensitivity, it is possible to esteem that the devices will be capable to detect arrival times for seismic events in a range of 10-50 Km from the source. However, it is important to point out that this estimation is pessimistic, since the ground acceleration is the minimal measured since does not consider the amplifying effect due to building structures and heights.

Table II reports the earthquake transmission velocities of some soils. The values ensure a time delay of several seconds, compatible with the time resolution of the instrument, in a distance of 50 Km, esteemed according to Table I, to be the sensitivity of EaST clients.

TABLE II. EARTHQUAKE TRANSMISSION VELOCITIES, V_P, AGGREGATED BY SOIL CLASSIFICATION

SOILS	
Classes	**V_p (Km/s)**
Rocks	5.7 - 0.85
Dense Clays	1.4 - 0.5
Sands and Gravels	0.7 - 0.1
Loose Clays	0.16 - 0.14

extracted from [21]

B. The detection

The EaST system is based on a mobile and not inertial sensor device set. The localization of sensor is performed via GPS, that provides also time synchronization, or by network triangulations, even if the detection algorithm has still to cope with all the problems due to the user movements.

The detection of a seismic event is performed both on the clients, to avoid overloading the server with false positives, and on the server, where a predetermined fraction of seismic triggers is always sent.

On the client, the detection is based on three subsequent phases characterized by increasing computational complexity and based on the dynamically updated execution parameters α, β, γ:

- *STA/LTA detection*: evaluation of *Short Time Average* (STA) and *Long Time Average* (LTA) ratio,
- *probabilistic positivity*,
- *detection of user activity*.

In particular, α ($\alpha>1$) represents the threshold for STA/LTA detection, β ($0\leq\beta\leq1$) the probability of excluding the NN detection of user activity and γ ($0\leq\gamma\leq1$) the NN sensitivity.

The clients continuously perform the *STA/LTA detection*: they observe a time window of accelerometer signals and compare the relative average values LTA with a shortest sample average value, the STA. When the STA/LTA ratio exceeds α, the client is measuring a rapid change in acceleration signals and an earthquake event is possible.

Probabilistic positivity is the second phase of the algorithm and guarantees that, after *STA/LTA detection*, a known percentage of the receivers will however send the

 ISBN: 978-1-61208-100-7

seismic trigger to the server. This phase enables to avoid starting the next one, and to demand a part of the detection to the server. The server maintains a list of working devices positions, and in presence of a good coverage of suspected area, is able to understand when a seismic trigger is a false positive to discard, according to the diffusion of the alert. Exploiting this server capability, during the *probabilistic positivity* phase, we explicitly exclude with p=γ, the NNs detection; basing on the other near clients, the server understands if the signalled variation in the ratio STA/LTA is due to a seismic phenomenon.

As stated in [20], STA/LTA based detection usually does not perform well at sites with high, irregular and in particular, man caused acceleration noise.

To overcome this limitation, the proposed algorithm combines this approach with the NN analysis of user activities, the *detection of user activity* phase. This enables to obtain good client robustness to user movements, still requiring only a light computation for on-line controlling the accelerometer signals.

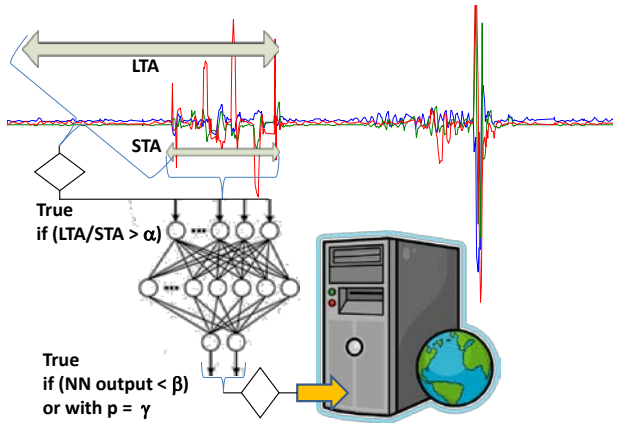


Figure 3. A schematic view of the seismic detection algorithm

As previously stated, the adopted sampling frequency for the sensor is the highest available, and oscillates around 15 measures per second (since Android is a multitasking system, the frequency adjusts according to the system scheduler). This big amount of on-line accelerometer data and the voluntary nature of user participation in the project, impose a strong optimization of required device resources.

Indeed, the evaluation of user activity is the only expensive phase, in terms of required processing time, as four NNs classify the acceleration signal aiming at excluding false positives seismic triggers. This phase is executed only when the STA/LTA ratio exceeds the threshold α and with p= 1-γ.

Figure 3 depicts the detection algorithm, graphically representing the acceleration and the samples adopted for LTA e STA. The execution parameters α, β, γ and the structure and configuration of the NNs are updated on the client during the time synchronization, when necessary.

Further details on the adopted NNs and the training phases are reported in next subsection.

Resuming, when the ratio STA/LTA exceeds α, with p=γ or when the NNs do not recognize any known user movement pattern (NNs output < β), a text file is prepared and sent to the server, communicating the client position, the device orientation, the current time and the acceleration data.

C. NNs and training

The NNs are software classifiers inspired to human brain [10]. NNs simulate the human nervous system activities with a self-adapting system composed of simple elements, the neurons. Each neuron computes a simple threshold function on its inputs. NNs are usually constructed organizing neurons in three layers: input layer, hidden layer and output layer.

According to [17] and [11], the EaST clients embed four simple Back Propagation NNs aiming at recognizing user movements as walking, using a lift, climbing the stairs or moving in a vehicle.

The Back-Propagation BP NNs are trained in a two steps procedure:

- in the first step, the forward propagation by positive model, the NN analyzes a sample and calculates, layer by layer, the input and output,
- in the second step, the error respect to the expected classification is propagated back and proportionally adjusts the weight and neuron threshold.

The operation is repeated until the error reaches the allowable range. The recognizing of user movements is not a simple task, and the variability in user habits does not help to give specific analytic functions or descriptions.

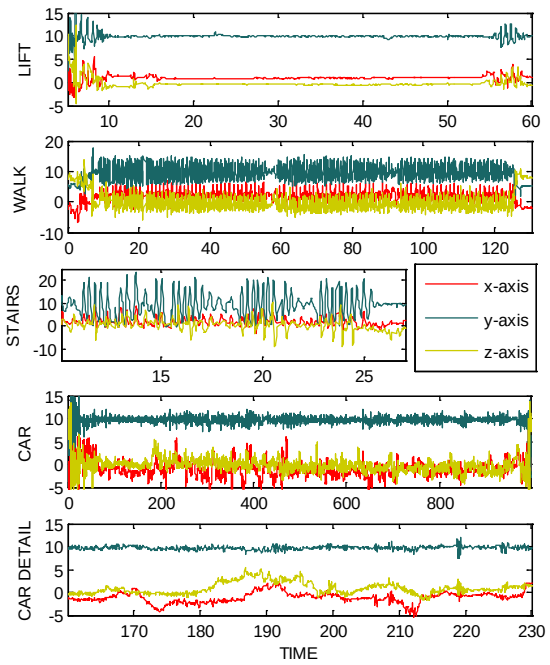


Figure 4. Acceleration sampled during four typical user actions.

In this way, the NNs can learn from historical data, stimulate the movement through neurons network, and approach the practical function by hidden layers to establish a relation model between acceleration and user movements, without considering specific details. Figure 4 reports one

 ISBN: 978-1-61208-100-7

graph for each class of samples on which we trained the first version of EaST clients. The training phase is still on going to obtain a good generalization of NNs to different users. More NNs, trained on other typical user movement patterns, are also going to be prepared form embedding on the EaST client.

During the detection algorithm, when the neural detection is required, the acceleration data are analyzed by the NNs. If the NNs do not classify any known user movement pattern, the client position and the device orientation, the current time and the acceleration data are sent to the server.

IV. CONCLUSION

The EaST system is still in tuning: we are simulating a device population to establish the best values of execution parameters respect to client number and geographical density. Once distributed, the EaST system will be based on the voluntary contribute of users that are asked to offer some resources on their mobile devices to detect seismic waves and provide seismic data to the central server. It is the analogous of other volunteer computing projects such as SETI@home (radio telescope data), Einstein@home (gravitational wave data), and climateprediction.net (testing the accuracy of climate models) that exploit user desktop PC resources to perform heavy scientific computations. In our case, the aim is slightly different since we aim at collecting data and do not require, usually, computational resources to our users.

After the current tuning of the system, the adoption of Android will let us freely distribute EaST application and arrive to a big community of potential users just using the official distribution channel of this platform: the Market. Android has surged to fourth place overall, growing from 1.6% to 9.6% market share from 2009 to 2010 [8].

The future development of clients for other common mobile platforms will ensure a bigger diffusion of the EaST system and an increased amount of data available on seismic events.

Indeed, actual estimations of mobile technology future strongly encourage our and general interest toward these technologies. According to Gartner Inc., an information technology research and advisory company, location based services user number will grow from 96 million in 2009 to more than 526 million in 2012 [7] and the 3G connection will be available on the 61% of mobile devices [12]. In the foreseen landscape, it will be crucial experimenting and proposing new ubiquitous and distributed, often pervasive, applications exploiting available resources.

Besides its interest for Earth scientists, we also expect that the research work connected with the realization and the improvement of the EaST system, will provide also interesting results in human computer interaction field. A better understanding of user habits and movement patterns will improve and propose new forms of life logging, context based services and content providing.

ACKNOWLEDGMENT

The authors are grateful to Donato Cirillo for the support provided during the development of the system.

REFERENCES

[1] Android SDK, retrieved on May 2010 from http://developer.android,com/index.html .

[2] R. C. Aster, B. Borchers, and C. H. Thurber, Parameter Estimation and Inverse Problems, Elsevier Academic Press, 2005.

[3] E. Cochran, J. Lawrence, C. Christensen, and A. Chung, A novel strong-motion seismic network for community participation in earthquake monitoring, Instrumentation & Measurement Magazine, IEEE, vol.12, no.6, 2009, pp.8-15.

[4] F. D. Dowla, S.R Taywr, and RW. Anderson: Seismic discrimination with artificial neural networks: preliminary results with regional spectral data, Bull. Seismol.Soc. Am., 80, 5, 1990, pp. 1346-1373.

[5] Á. Fábián, N. Győrbíró, and G. Hományi, Activity recognition system for mobile phones using the MotionBand device, Conference on Mobile Wireless Middleware, Operating Systems, and Applications, 2008, pp 1-5.

[6] A. Frassetto, T.J. Owens, and P. Crotwell, Evaluating the Network Time Protocol (NTP) for Timing in the South Carolina Earth Physics Project (SCEPP), Seismological Research Letters, vol.74, 2003, pp. 649-652.

[7] Gartner Identifies the Top 10 Consumer Mobile Applications for 2012, retrieved on May 2010 from http://www.gartner.com/it/page.jsp?id=1230413 .

[8] Gartner Says Worldwide Mobile Phone Sales Grew 17 Per Cent in First Quarter 2010, retrieved on May 2010 from http://www.gartner.com/it/page.jsp?id=1372013 .

[9] Joone, Java Object Oriented Neural Engine, retrieved on May 2010 from http://sourceforge.net/projects/joone/ .

[10] N. Kasabov, Foundations of Neural Networks, Fuzzy Systems, and Knowledge Engineering, 1996, The MIT Press.

[11] H. Ketabdar and M. Lyra, ActivityMonitor: Assisted Life Using Mobile Phones, IUI'10, Hong Kong, China, 2010, pp. 417-418.

[12] Mobile Life 2012, retrieved on May 2010 from http://www.bitkom.org/en/Default.aspx .

[13] G. Nolet, Seismic wave propagation and seismic tomography. Seismic Tomography, edited by G.Nolet, Reidel, Dordrecht, 1987, pp. 1-23.

[14] OpenGL ES - The Standard for Embedded Accelerated 3D Graphics, retrieved on May 2010 from http://www.khronos.org/opengles/spec/ .

[15] Open Handset Alliance, retrieved on May 2010 from http://www.openhandsetalliance.com .

[16] The Quake-Catcher Network, retrieved on May 2010 from http://qcn.stanford.edu/ .

[17] G. Romero, Seismic signals detection and classification using artificial neural networks, ANNALI DI GEOFISICA, VOL. XXXVII, N. 3, 1994, pp. 343-353.

[18] F. Rosenblatt, Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms, Spartan Books, New York, 1962.

[19] ShakeMaps Technical Manual, User's Guide, and Software Guide retrieved on May 2010 from http://pubs.usgs.gov/tm/2005/12A01/ .

[20] B. K. Sharma, A. Kumar, and V. M. Murthy, Evaluation of Seismic Events Detection Algorithms, JOURNAL GEOLOGICAL SOCIETY OF INDIA, Vol.75, 2010, pp.533-538.

[21] K. Terzaghi and R. B. Peck, Geotecnica, 1974, UTET.

[22] J. Y. Yang, J. S. Wang, and Y. P. Chen, Using acceleration measurements for activity recognition: An effective learning algorithm for constructing neural classifiers, Pattern Recognition Letters, Elsevier, 29, 2008, pp. 2213–2220.

 ISBN: 978-1-61208-100-7

MMSP: Designing a Novel Micro Mobility Sensor Protocol for Ubiquitous Communication

Dhananjay Singh and Daeyeoul Kim
Division of Fusion and Convergence of Mathematical Sciences
National Institute for Mathematical Sciences,
Daejeon, Korea
singh@nims.re.kr ; daeyeoul@nims.re.kr

***Abstract*—Designing low power and less delay for mobile nodes is one of the most important issues for the ubiquitous sensor networks (USN). The paper presents a novel Micro Mobility Sensor Protocol (MMSP), as an enhanced form of the AODV (Ad hoc on-demand Distance Vector) protocol, in order to improve the quality of mobile IP (IPv6)-USN nodes. An IP-USN node can include various sensors for application monitoring. In this MMSP technique, IP-USN nodes can easily move to monitor their applications within the range of a PAN coordinator, connected to internet based networks. By using this technique, the user can globally receive sensor data on any internet based equipment such as a PDA, notebook, cell phone, etc. The IETF working group has defined, in RFC 4919 and RFC 4944, standards for IPv6 over Lowpan (IP-USN). We carry out a performance analysis in an NS-2 simulator for, e.g., a small number of IP-USN nodes randomly deployed to monitor a number of targets. Each target (IP-USN) may be redundantly covered by multiple sensors. To conserve routing performance of IP-USN networks, we have organized sensors in sets activated successively. This novel protocol can use monitoring and detection with respect of various applications.**

Keywords—Mobility; IP-USN; AODV; Routing; Global communication.

I. INTRODUCTION

Nowadays, technological development has changed our lives in many ways by providing increased comfort and safety, and has affected almost every field of work and education. One of these fields is that of communication; here it has enriched our lives by enabling us to communicate more easily and, to some extent, more cheaply, with people living in different places on the earth. Information technology is still developing easier, faster, and more accurate technologies to improve the quality of our life. The technologies that enable this connection are called ubiquitous computing technologies. The pervasive nature of IP-based networks allows the use of the existing infrastructure. IP based mobile (cell phone) technologies already exist, and can be connected readily to internet based networks, without the need of intermediate entities like translation base station or proxies. Such a method is specified in open and freely available specifications, which is a situation favorable to, or at least better able to be understood by, a wired audience than would be proprietary solutions. Tools for the diagnostics, management, and commissioning of IP-networks, already exist. Due to the rapid development of new paradigm applications, wireless networks are morphing into IEEE 802.15.4–the standard for Lowpan (Low power wireless personal area networks), which are playing an essential role in the realization of the envisioned ubiquitous world. IEEE 802.15.4 (Lowpan) needs to be connected with other Lowpans as well as with other wired networks in order to maximize the utilization of information and other resources. However, the maximum frame size of IEEE 802.15.4 is 127 octets while UDP and IPv6 have big packet size and no space for applications data. The PANs consist of various IP-USN (IPv6 over ubiquitous sensor networks) nodes. As well, one has to overcome problems such as network overhead, node discovery, and security. When that technology is integrated with IPv6, we have a vast amount of possibilities for implementing applications because IP has been used for a long time and technologies related to it already exist, as IP-connectivity is spreading its influence to all kinds of applications [1] [2].

IP-USN has currently become a hot subject for researcher with the advancement in WSN (wireless sensor networks). This is evolving together with global connectivity between IP-sensor devices and IP-network services. The IETF (Internet Engineering Task Force) working group has been standardizing a new development called 6lowpan (IPv6 over low power wireless personal area networks), which refers to an IPv6 integrated with a Lowpan device [2].

This paper proposes a novel mobility approach and analyses the simulation results of IP-USN networks. We have created an NS-2 simulation-based 6lowpan stack. It, in this stack, presents compression techniques, protocol designs, a mobility approach, data binding techniques, and communication between neighboring nodes in the same environment by diffusing throughout a specific field with inter-PAN networks. The aim of this paper is to develop global communications between IP-USN nodes and service

 ISBN: 978-1-61208-100-7

providers. The service provider (user) connects directly and checks the current status of the IP-USN based sensor node (for application- data) with the help of an existing wireless internet-based technologies such as cellular, GPS [16], Wifi [17] services used by PDA, notebook, and cell phone [16].

II. RELATED WORKS

So far, many mobility protocols have been proposed based on IPv6 for tunneling mobile nodes, such as HMIPv6 and MIPv6. These have managed to reduce packet losses while the mobile nodes are moving. HAWAII and Cellular IP networks require mobile nodes to manage mobility through path setup messages. The mobility related packets are used in the IP layer at IP traffic. Researchers have followed different approaches to give connectivity to a mobile user when the user is moving. Integrating mobility with an IP-USN node is very useful for applications. IP-USN node has considered IEEE 802.15.4 networks with internet for global monitoring applications. There are a lot of packet losses due to mobility. Few of the following mobility protocols have been proposed for IP-USN networks [3–5].

NEMO (Network Mobility) is a routing based mobility protocol, which requires a mobile router to support the mobility of a WPAN [5]. NEMO provides connectivity to all mobile nodes. Basic Support ensures session continuity for all the nodes. Mobile Node does this by adding routing capability between its point of attachment (Care-of Address) and a subnet that moves with the Mobile Router. It is non-supportive of multi-homing for Mobile Routers [6].

LoWMob & DLoWMob (Intra-PAN Mobility Support Schemes for 6LoWPAN) use static nodes for multi-hop communication between PAN coordinators and mobile nodes in Intra-PAN networks. LoWMob is a network based mobility approach for mobile 6lowpan nodes in which the mobility of the 6lowpan nodes is handled at networks-side. It ensures multiple-hop communication between PAN coordinators and mobile nodes with the help of static node within a 6lowpan. The distributed version of DLoWMob, which employs mobility, supports the distribution of the traffic connection at the PAN coordinators and optimizes a multi-hop path between sources and destinations [7].

PMIPv6 (Proxy Mobile IPv6) is a network based localized mobility scheme. In this system, MNs' movements and setup require routing states. PMIPv6 uses host based mobility protocols, thus it is good for 6LoWPAN mobility management. PMIPv6 is also compatible with any global mobility management protocols such as Host Identity Protocol (HIP), IKEv2 Mobility, and Multi-homing (MOBIKE) [8].

MUNNA (Mobile Ubiquitous Nodes, Negotiation Agent) are techniques where the devices move within a mobile network. It shares the responsibility by revolving the load of smaller devices to bigger network elements such as a PAN coordinator and a mobile router. It hosts a mobile router and a 6lowpan PAN coordinator. Its main function is to maintain a delegation table which is specially designed to support the mobility of sensor nodes or low capacity devices. MUNNA techniques have objectively analyzed the scalability of the system using the throughput and delay measures for benchmarking their performance under the influence of mobility. The IP-USN nodes must be addressable by any corresponding node, independent of its current whereabouts [9].

III. SYSTEM DESIGN

During the earlier development phase, wireless sensor applications focused mainly on environment and industrial monitoring applications, but now different applications are emerging from all fields. The ubiquitous communication has also witnessed a few new applications for wireless sensor networks, but they are a bit different as far as the issues that need to be addressed. Earlier applications focused mainly on the ways to optimize the power consumption in the network, and gave less priority to the reliability of packet transmission. However, in the global scenario the main purpose shifts from power to reliability. So the design of wireless applications should focus more on the reliability of packet transmission, although this does not mean that power consumption should be ignored.

Fig. 1 describes the ubiquitous communication sensor networks for global connectivity between IP-USN node and service provider. In this system IP-USN nodes are able to move easily within the range of PAN coordinator which is integrated with IPv6-based wired networks. Thus, the service provider can easily get to know the current position and its application data on internet provider equipments. This integration will help realize ubiquity by allowing global to access application data across IP-USN system and wired IP-based networks [10] [11].

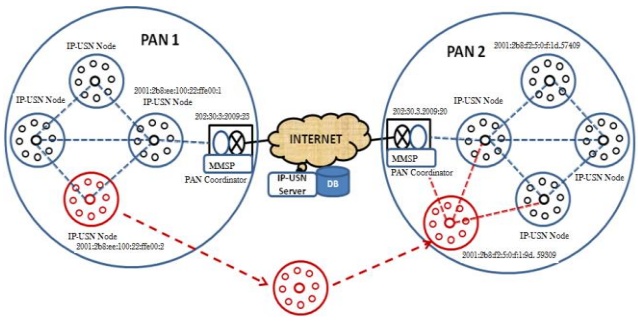


Figure 1. Mobility in ubiquitous communication sensor networks.

In relation to the maximum utilization of resources, they are mainly associated with internet-based networks. These networks are characterized by short range, low bit rate, low power, and low cost. Many devices used are also limited in computational power, memory and energy. The novel MMSP protocol is highly reliant on the availability of other information, such as physical location, global ID, data

 ISBN: 978-1-61208-100-7

gathering, transmitting to the PAN coordinator, etc. The global ID is desirable in sensor networks. Thus, the sensors can be distinguished from other networks. The sensor node has address space for the global ID, which will cause it to establish communication with IPv6 networks. For the operation of some routing protocols, we do need to distinguish sensor nodes to some extent, but a locally unique ID will suffice [12] [13].

IV. MMSP: MICRO MOBILITY SENSOR PROTOCOL

MMSP works as the back bone of the mobile IP. The idea behind the design is to modify the cellular IP in such a way as to get location information at a particular instant in time and to find the estimated velocity during handoff. To find the location of the IP-USN node at a particular instant in time, directional antennae located on the PAN coordinator are used, directed towards the highest roaming probability inside the PAN or smart networks. In this technique, the PAN coordinator (GW) stores the location information of all IP-USN nodes shown in Table 1. The MMSP knows its radius and maintains a routing table for the each of the IP-USN nodes. The intermediate IP-USN node also maintains a route cache as a PAN coordinator. The PAN coordinators broadcast periodic route query messages to detect available IP-USN nodes in its wireless coverage or PAN. Responding to query messages, all IP-USN nodes in the coverage field send route update messages. After the time elapsed during the exchange of both control packets, the MMSP calculates the distance of the IP-USN from the PAN coordinator or of the PAN coordinator from other nodes. The PAN coordinator keeps the location information of all IP-USN nodes. Table 1 shows the present IP-USN radial component in R1 [14] [15].

TABLE 1. PRESENT IP-USN RADIAL COMPONENT IN R1

IP-USN seq. no.	Location in terms of radius in cm	Location in terms of azimuthal angle in Radian (α &β)	Shortest path
IP-USN (A)	R1	α	MMSP 1
IP-USN (A)	R2	β	MMSP1

The angle of the antenna lobe at which it receives maximum strength from a particular IP-USN, is taken as approximately equal to the azimuthal angle between the two, α. The values of the angles are tabulated as the current positions shown in Table 2. After the completion of consecutive control message exchanges, the MMSP again records the R2 and β for the IP-USN node.

The PAN coordinator maintains a routing table for the IP-USN as shown in Table 2 with its position information. All position entries are taken in circular coordinates. Table 2 is updated with R2, β, using these two position values as well as the time delay between the two entities, the approximate velocity of the IP-USN node is calculated and further updated in Table 3.

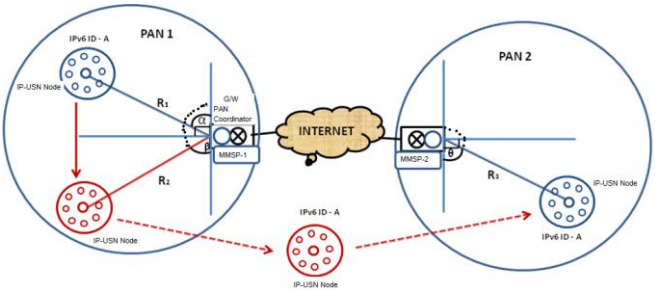


Figure 2. IP-USN mobility in a specific field.

Whenever the MMSP receives a route update packet from the IP-USN, the MMSP updates its route cache. If it receives a route update packet for the first time when a new IP-USN enters its area of coverage, a new entry is made for the IP-USN and the route validation time is set. If the PAN coordinator receives a route update message from an old IP-USN, it refreshes the old route; besides the route update packets, the IP-USN sends a periodic page update packet to the nearest IP-USN.

TABLE 2. MMSP MAINTAINS A ROUTING TABLE IN PAN

IP-USN Seq. no.	Root Validation Time	Current Position (R2, β)	Last Position (R1, α)	Velocity [(R2, β) -(R1,α)]/T

The MMSP makes more for complications in structure due to the greater number of directive antennas instead of one omni-directional antenna. Extra computation will be needed on the part of the PAN coordinator.

TABLE 3. MMSP MAINTAINS A ROUTE TABLE FOR THE OTHER PAN

IP-USN seq. no.	Location in terms of radius in cm	Location in terms of azimuthal angle in radian	Shortest path
IP-USN (A)	R3	θ	MMSP2

In Fig. 3 is presented the communication scheme where the PAN coordinator broadcast a query packet to the IP-USN networks (including approximate receiving signal strength for 1st level) at once and then waits for a reply until the timer expires. The timer is set on the IP-USN according to the velocity, signal strength and distance between IP-USN networks and PAN coordinator. Each level has to define the hop distance between the IP-USN node and the PAN coordinator. The PAN coordinator broadcasts a query packet into the mesh topology. The IP-USN node receives a packet within an area that compares to the signal strength, according to an RSS value which the node joins or establishes a connection to the PAN coordinator. Then, the IP-USN node sends a Query_response (IPaddr) packet to the PAN

 ISBN: 978-1-61208-100-7

coordinator that the IP-USN nodes are joining with the coordinator. Then, the IP-USN nodes adjust their transmission power to the PAN coordinator for further communication processes.

```
ALGORITHEM
// MMSP Coordinator Functions
while (receivingPacket) {
 //calculate distance
        lastLocation = getLastLocation (packet.nodeId);
        currentLocation = getLocation (packet);
        distance = getDistance (currentLocation,
        lastLocation);
 //calculate interval
        lastTime = getLastTime (packet.nodeId);
        currentTime = getCurrentTime (packet);
        timeInterval = getInterval (lastTime, currentTime);
 //update route information
        velocity = distance/timeInterval;
        updateRouteCache(packet.nodeId,location,currentTi
        me, velocity);
}
// IP-USN Functions
while (periodicTimer) {
        broadcast (updatePacket);
}
// Get Distance
getDistance (currentLocation, lastLocation) {
        distance = lastLocation.R*cos(lastLocation.alpha)
        - currentTime.R * cos (currentTime.alpha);
        return distance;
}
```

Figure 3. Pseudo-code for Micro Mobily Sensor Protocol .

V. PERFORMANCE ANALYSIS

The random waypoint mobility techniques are used during movement of IP-USN nodes. Each node moves randomly within define topology field at a random speed. The speeds are uniformly defined between 1 to some maximum speed. Each node start movement by stationary pause time in value of seconds and after reaching the destination it should stop in pause time seconds as its defined instruction. The mobility will be repeated during simulation process. The mobility process will be set before simulation started there thus we can set the distance of nodes. We have evaluated the proposed MMSP (Micro Mobility Sensor Protocol) by developing a complete simulation in NS 2.33 and through numerical analysis. The terrain area is 500 × 500 m2. A total of 25 number of IP-USN nodes are deployed in a 4 ×4 logical grid. The main reason of dividing the whole area into a grid is to examine the IP-USN node behavior at each step. We have used the random way point mobility model and the fluid flow mobility model. The minimum speed of an IP-USN node is 1 m/s, and the maximum speed varies between 20 m/s, 25 m/s, 30 m/s, and 35 m/s. The IP-USN node pause time is 30 sec. The MMSP is used as mobility enabled routing protocol. The simulation is run for 500 seconds and there are 20 simulations run. The performance metrics of interest are the end-to-end delay, packet delivery ratio, and handoff. The packet delivery ratio is the ratio of the number of packets successfully received by the PAN coordinator, out of the ones that are transmitted by an IP-USN node; and for the communication between multiple IP-USN node packets, the success rate is the number of packets that are successfully received by a IP-USN node out of the ones that are transmitted by another IP-USN node and hand off overhead. Figures 4 and 5 have described the end-to-end delay and the packet delivery ratio of the packets between an IP-USN node and the PAN coordinator. The speed of the IP-USN node and the number of hops between them varies. After a certain number of hops, the end-to-end delay increases linearly with the increasing number of hops between the IP-USN node and the PAN coordinator. Also, the end-to-end delay increases when the speed of the IP-USN node increases. This is because as the speed of the IP-USN node increases, the association of the IP-USN node and sensor node breaks, triggering the handoff process. Thus when the IP-USN node moves with high speed, most of the time is spent to complete the handoff process by the newer and the older IP-USN node. MMSP broadcasts packets by bringing traffic into the network that not only causes collisions but also introduces the hidden node problem. The packet delivery ratio, when the IP-USN node is far away from the PAN coordinator, i.e., 5 hops, is just about 0.4 for an IP-USN node moving with the speed of 20 m/s. As the number of hops between the PAN coordinator and the IP-USN node decreases, the packet delivery ratio increases. And when, the IP-USN node comes closer to the PAN coordinator, the success ratio approaches 1, and the end-to-end delay approaches 0.01 seconds. Moreover, it can be seen from Fig. 4 that, when the speed of the IP-USN node is 20 m/s, the packet delivery ratio is better than when the speed is 25 m/s. This is because as the speed increases, the number of handoffs increases, which can lead to a significant packet loss. Also, when the speed increases exponentially, there is a possibility that the IP-USN node will be lost in the PAN. This is because, as the new IP-USN node wakes up for the handoff process, the IP-USN node may have already crossed the new IP-USN node. As shown in Fig. 5, when the IP-USN node is 5 hops away from the PAN coordinator, the packet

 ISBN: 978-1-61208-100-7

delivery ratio at the speed of 30 m/s is almost double than that of speed of 25 m/s.

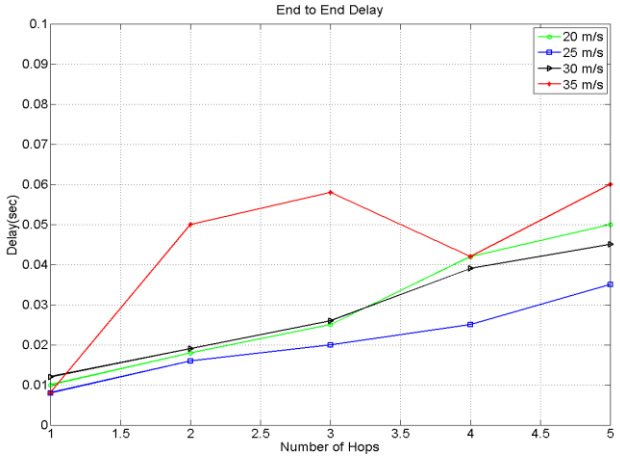


Figure 4 End-to-End delay during IP-USN mobility.

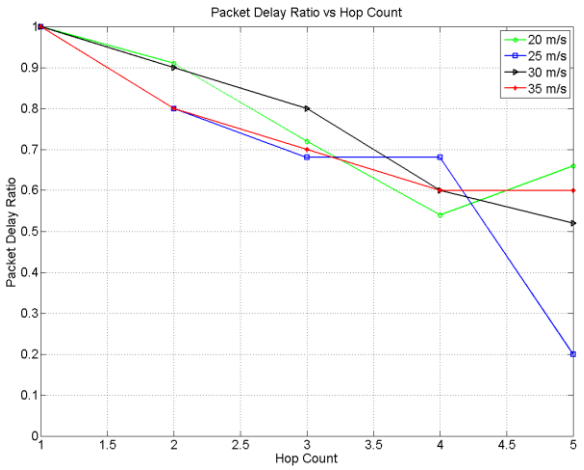


Figure 5. End-to End PDR during IP-USN mobility.

The performance of this novel MMSP in terms of packet delivery ratio and end-to-end delay is good when the IP-USN node is closer to the PAN coordinator. But usually this is not the case, because the IP-USN node can move anywhere within the network. Moreover, as the network size increases, the performance of the novel MMSP algorithm decreases dramatically. Also, as the speed increases, the number of handoffs increases, thus degrading the network lifetime.

VI. CONCLUSION

In this paper, we have proposed a Micro Mobility Sensor Protocol (MMSP) for IP-USN nodes. The node can easily move into the specified field and globally communicate with service providers by PAN coordinator. MMSP is a modified form of the AODV protocol. It comes with a relatively new idea for tackling the increasing performance of various applications such as healthcare monitoring, structural monitoring, location monitoring etc. The application based data packet try to utilize such increasing speed, because currently the increasing speed permits such a kind of system. By using this technique, the service provider globally receives sensor data on internet based equipments such as PDA, notebook, cell phone, etc. This paper is a step towards bringing wireless networking closer to the global communication techniques.

ACKNOWLEDGMENT

This work was supported by NAP of Korea Research Council of Fundamental Science & Technology.

REFERENCES

[1] N. Kushalnagar, G. Montenegro, and C. Schumacher, "IPv6 over Low-Power Wireless Personal Area Networks (6LoWPANs): Overview, Assumptions, Problem Statement, and Goals", RFC 4919, August 2007.

[2] G. Montenegro, N. Kushalnagar, J. Hui, and D. Culler, "Transmission of IPv6 Packets over IEEE 802.15.4 Networks", RFC 4944, September 2007.

[3] H. Soliman, C. Castelluccia, K. El. Malki, and L. Bellier, "RFC-4140: Hierarchical Mobile IPv6 Mobility Management (HMIPv6)", August 2005.

[4] D. Johnson, C. Perkins, and J. Arkko, "Mobility Support in IPv6", RFC-3775, June 2004.

[5] E. Nurvitadhi, B. Lee, C. Yu, and M. Kim, "Adaptive semi-soft handoff for Cellular IP networks", International Journal of Wireless and Mobile Computing archive, Vol. 2, July 2007, pp.109-119.

[6] V. Devarapalli, R. Wakikawa, A. Petrescu, and P. Thubert, "RFC3963 - Network Mobility (NEMO) Basic Support Protocol", Network Working Group, January 2005.

[7] G. Bag, M. T. Raza, K. H. Kim, and S.W. Yoo, "Inter-PAN Mobility Support for 6LoWPAN", Sensors 2009, vol. 9, 2009. pp.5844-5877.

[8] S. Gundavelli, K. Leung,V. Devarapalli, K. Chowdhury, and B. Patil, " RFC-Proxy Mobile IPv6" IETF draft., August 2008.

[9] M. Hasan, A. H. Akbar, H. Mukhtar, K. H. Kim, and D. W. Kim, "A scheme to support mobility for IP based sensor networks", Proceedings of the 3rd international conference on Scalable information systems, vol. 5, 2008, pp.28-38.

[10] Dhananjay Singh, "IP-Based Wireless Sensor Networks for Global Healthcare Monitoring Applications, (Ph.D. thesis) Dongseo University, Busan, Korea, (*published*) Feb. 2010, pages 217.

[11] E. Kim, D. Kaspar, C. Gomez, and C. Bormann, "Problem Statement and Requirements for 6LoWPAN Routing" draft-ietf-6lowpan-routing-requirements-06 (work in progress) March, 2010.

[12] R. C. Wang, R. S. Chang, and H. C. Chao, "Internetworking Between ZigBee/802.15.4 and IPv6/802.3 Network", Proceedings of ACM SIGCOMM 2007 Workshops, pp. 362-367, Japan, August 27-31, 2007.

[13] T. Winter and P. Thubert, "RPL: IPv6 Routing Protocol for Low power and Lossy Networks", Internet Draft, June 2010, pages 103, (draft-ietf-roll-rpl-09).

[14] D. Singh and H. J. Lee, "Design and Performance Evaluation of a Proactive Micro Mobility Protocol for Mobile Networks", chapter in the Book: Handheld Computing for Mobile Commerce: Applications, Concepts and Technologies, IGI Global publisher, USA, Feb. 2010, p.p.328-342.

[15] D. Gao, C. H. Foh, H. Zhang, and L. Liang, "MSRP: A Light Weight Cross-Layer Routing Protocol for IPv6 Wirelss Sensor Networks", Sensors Journal, 2010, doi : 10.3390/ (*In press*).

[16] http://www.pcmag.com/article2/0,2817,2316534,00.asp (July, 2010).

[17] http://www.webopedia.com/TERM/W/Wi_Fi.html (July, 2010).

 ISBN: 978-1-61208-100-7

Handover Scenario and Procedure in LTE-based Femtocell Networks

Ardian Ulvan, Robert Bestak
Department of Telecommunication Engineering
Czech Technical University in Prague
Prague, Czech Republic
ardian.ulvan@fel.cvut,cz, bestar1@fel.cvut.cz

Melvi Ulvan
Department of Electrical Engineering
The University of Lampung
Bandar Lampung, Indonesia
melvi_ulvan@unila.ac.id

***Abstract*—The deployment of Femtocell as the emerging wireless and mobile access technology becomes a solution for the bandwidth limitation and coverage issues in conventional mobile network system (macrocell). In this paper the handover procedure in femtocell network is investigated. The procedure is based on 3GPP LTE specification. Three handover scenarios: hand-in, hand-out and inter-FAP are considered and analysed. In order to achieve the optimize procedure, the handover decision policy based on mobility prediction is introduced and proposed. The reactive and proactive handover strategy is also proposed to mitigate the frequent and unnecessary handover. The result shows that reactive handover is the potential mechanism to mitigate the unnecessary handover.**

Keywords - handover; femtocell; 3GPP-LTE; reactive handover; proactive handover.

I. INTRODUCTION

Femtocell is the emerging network technology, which is defined as a low-cost, low-power cellular base station that operates in licensed spectrum to connect conventional, unmodified mobile terminals to a mobile operator's network. The coverage ranges of femtocells are in the tens of meters. They utilize broadband Digital Subscriber Line (DSL) or cable/fiber to the home (FTTH/FTTx)) Internet connections for backhaul to the operator's core network [1].

The Femto Access Point (FAP), also known as Home Base Station (HBS) or Home Node B (HNB) in 3GPP LTE terminology, is a main device in femtocell network that provides radio access network (RAN) functionality [1]. The FAPs were initially designed for residential use to get better indoor voice and data coverage, improving at the same time the macrocell reliability and promise to be a cost-effective solution. It also increases the peak-bit rate in low coverage areas.

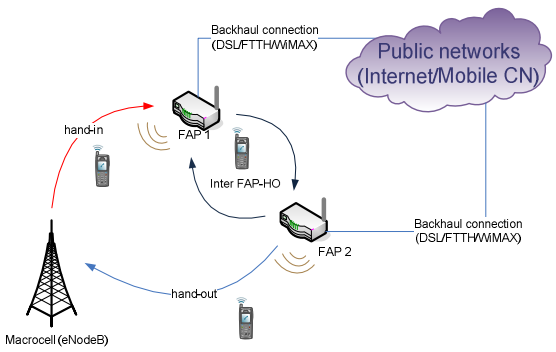


Figure 1. Handover scenario in femtocell networks

Femtocells and the conventional macrocells are seen as isolated networks, but they are not. In this paper, we describe the interaction between femtocells and macrocells in term of handover. The implementation of femtocell may cover the "blank area" and to increase the utilization of wireless capacity which is not covered by macrocell base station. Nonetheless, the availability of hundreds of FAPs in a particular area most likely increases the technological challenges in handover procedure. Another challenge is the mitigation the unnecessary handover since large number of FAPs can trigger the very frequent handovers even before the current initiated handover procedure is completed.

Research and technological development on handovers in macrocell network has been going extensively to provide better Radio Resource Management (RRM). Most of the researches are in the field of cellular networks focused on network-controlled horizontal handover where handover is executed between adjacent cells of the same network.

In term of IP-based wireless network, the research on handover has been done typically in wireless local area network (WLAN) based on WiFi IEEE802.11. Moreover, the client-based handover began to be investigated when the Worldwide Interoperability for Microwave Access (WiMAX) IEEE802.16 networks, 3GPP Long Term Evolution (LTE)/Long Term Evolution-Advanced (LTE-A) as well as Mobile IPv4/IPv6 are introduced. In addition, the inter-system handover or vertical handover is also going investigated intensively. The research in both layer-1 (L1-physical) and layer-2 (L2-Medium Access Control - MAC) is undertaken in order to achieve the most efficient handover and to reduce the handover overhead.

In this paper, the handover between femtocell and macrocell is investigated. Three handover scenarios are considered as shown in Fig. 1. Handover procedure is based on 3GPP LTE specification.

The rest of the paper is organized as follow: Section 2 reviews some related works of handover in femtocel network. Section 3 describes the LTE-based handover in femtocell network. The handover scenarios and optimization proposal are presented as well. In Section 4, the handover signalling flow is analyzed in each scenario. Section 5 provides the proposal of handover optimization algorithm and a performance evaluation of the proposed algorithm, as

 ISBN: 978-1-61208-100-7

well as a result of the performance evaluation. We conclude our work in Section 6.

II. RELATED WORK

In the femtocell network, several research works have been published. The authors in [2] overviewed the 3GPP LTE and the characteristic of Home-eNodeB (HeNB). Their work included the description of mobility support in 3GPP LTE, the handover procedure in LTE and the deployment scenario of HeNB. In addition, some mobility management issues such as handover scenario, mechanism for searching the HeNB in the Closed Subscriber Group (CSG), cell reselection and handover decision parameters have also been described. The work concluded with the recommendation for further work to deal with such issues.

More detail works on handover in femtocell network have also been published. In [3], the work focused on the handover from the macro-tier to the femto-tier in CDMA network. It has been revealed that the User Equipment (UE) may be required to scan the whole femto radio spectrum when switch form macrocell to femtocell, however it is assessed as an expensive operation. To deal with this issue, the cache scheme for femtocell reselection is proposed. By considering the random walk movement, the three user movement models were applied to obtain the UE's movement history. The history included the number of FAP that has been visited. The idea behind this scheme was to obtain the most recently visited order of FAPs stored in the cache. The scheme seems effective in the open subscriber group (OSG) femtocell with plenty of FAPs, however it is relatively inefficient in the femtocell's CSG or in the few number of FAPs.

In order to integrate the femtocell into the system, some modifications on existing network and protocol architecture of Universal Mobile Telecommunication System (UMTS) based macrocell network has been proposed in [4]. The modifications included the change of signal flow for handover procedures and the measurement of signal-to-interference ratio for handover between macrocell and femtocell. The frequent and unnecessary handover is also considered. The analysis is taken on the concentrator-based and without concentrator-based femtocell network architecture. The result shown, the call admission control (CAC) scheme is effective to prevent the unnecessary handover.

In [5], the handover procedure between the HeNB and eNodeB has been proposed to be modified. A new handover algorithm based on the UE's speed and Quality of Service (QoS) is proposed. Three different velocity environments have been considered in the algorithm i.e., low speed (0-15 km/h), medium speed (15-30 km/h) and high speed (>30 km/h). In addition, the real-time and non-real-time traffics have been considered as QoS parameters. The comparison analysis shown that the proposed algorithm has a better performance then traditional handover algorithm in order to reducing the unnecessary handovers and the number of handovers. However, the assigned user velocities seem unrealistic since the HeNB at home deals only with the very low speed (0-5 km/h).

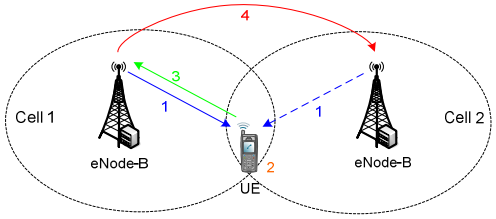


Figure 2. Handover process in 3GPP-LTE

III. HANDOVER IN FEMTOCELL NETWORK

A. *Handover in 3GPP-LTE Macrocell*

The 3GPP LTE for the 4G mobile system specifies the handover procedure and mechanism that support various users' mobility [6] [7]. Handover process is divided into four parts as shown in Fig. 2: UE measures downlink signal strength (blue line 1), processing the measurement results (2) and sends the measurement report to the serving eNodeB (green line 3). The serving eNodeB then makes the handover decisions based on the received measurement reports (red line 4).

The message sequence diagram of the LTE handover procedure is shown in Fig. 3. The handover procedure consists of 3 parts:

- Handover preparation; in this part, UE, serving eNodeB and target eNodeB make preparation before the UE connect to the new cell. The main message and process are described as follows:

1. Measurement control/report (messages 1/2); the serving eNodeB configures and triggers the UE measurement procedure and UE sends measurement report message to serving eNodeB.
2. Handover decision (messages 3/4); the serving eNodeB offers the handover decision based on received measurement report message from UE.
3. Admission control (messages 5/6); the target eNodeB

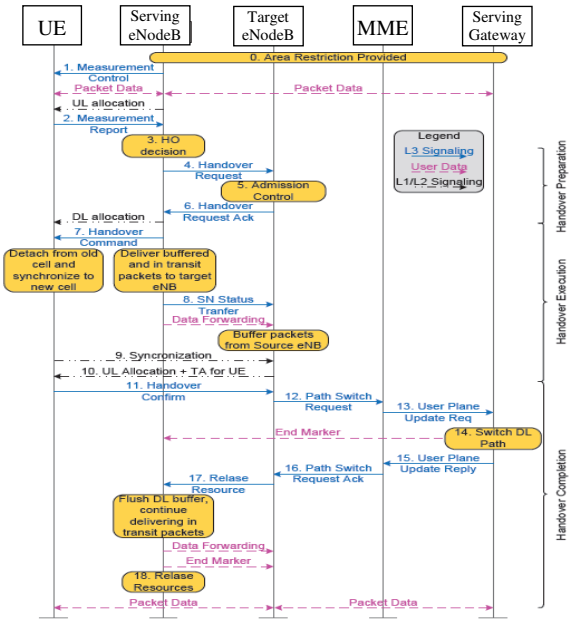


Figure 3. Message sequence diagram of handover Procedure in 3GPP-LTE [6]

performs the admission control dependent on the quality of service (QoS) information and prepares handover with L1/L2.

4. Handover command (message 7); the serving eNodeB sends the handover command to UE.

- Handover execution; on the execution part, the processes are described as follow:

5. Detach from old cell and synchronize to the new cell (messages 8 – 10); UE performs the synchronization to the target cell and accesses the target cell.

- Handover completion; this part includes the following processes:

6. Handover confirm and path switch (messages 11 – 16); the serving-Gateway switches the path of downlink data to the target side. For this, the serving-Gateway exchanges massage with Mobility Management Entity (MME).
7. Release resource (messages 17/18); upon reception of the release message, the serving eNodeB can release radio and control of related resources. Subsequently, target eNodeB can transmit the downlink packet data.

B. Handover Scenario in Femtocell Network

All mobile systems including the femtocell network implement a handover procedure to support the user's mobility. The handover, in one side allows communication during user's movement in the network. On the other side, it significantly increases signalling overhead in the network.

According to [8], it most likely that the soft handover will not be implemented in femtocell due to limited frequency allocation for femtocells. In addition, due to technological challenges and system operator requirements, the initial 3GPP specification for handover in femtocell focused on one direction only that is from FAP to macrocell eNodeB [9].

Despite having some constraints, in this paper we consider all possible handover scenarios between eNodeB and FAP and between FAPs. There are three possible handover scenarios in femtocell, as depicted in Fig. 1:

- Hand-in; this scenario presents the handover where an UE switch out from macrocell eNodeB to FAP.
- Hand-out; represents the handover that is performed from FAP to macrocell eNodeB.
- Inter-FAP handover; it corresponds to the scenario of handover from one FAP to another FAP. In this scenario all FAPs are assumed to be placed at the same location and served by the same service provider.

C. Decision Policy of Handover

One of the challenges in the handover procedure is corresponded to the handover decision mechanism. The common metrics for handover decision mechanism include carrier to Interference-and-Noise Ratio (CINR), Receive Signal Strength Indicator (RSSI) and Quality of Service (QoS). However, those metrics are quite demanding to deal with advanced handover requirement, for instance the fast handover in femtocell network that consist hundreds of possible target FAP. Therefore, the new handover decision mechanism metrics is necessary to be determined.

The handover decision option basically are network-controlled handover in which the decision to implement handover is taken by the eNodeB (in case of hand-in) or FAP (in case of hand-out and inter-FAP) to which the UE is currently attached. However, the support of client-based handover in which initiated by the UE becomes more common. This option gives the handover process more efficient, since any changing of necessary parameters or events (such as CINR, RSSI, coverage, the QoS provided by the network, the probability of next position, etc.) can be monitored by the UE from its wireless interfaces, then use them to decide to trigger the handover.

In network controlled mode, the serving eNodeB decides to perform handover to target FAP by comparing the RSSI that received by UE and the RSSI from the FAP. However, when the CSG is deployed, other parameters e.g., service cost, load balancing, and speed status of UE, which might influenced the handover decision should also be considered. Since the femtocell system offers the different billing models, the user's billing is sum up by whether user is using the FAP. Therefore it is important for UE to handover to the accessible FAP fast.

In the load balancing point of view, when a large number of active UEs are located in a given cell, available resources may be insufficient to meet the QoS for the real time service. But, it may offer the good performance for the best effort service. Particularly, in the FAP case where the available user is limited, if the available resource is too short for UE to handover to CSG cell, then it needs to handover to another accessible FAP or to macrocell eNodeB.

D. Proposed Mobility Prediction

The mobility prediction of UE may also be considered for the handover decision. In this paper we introduced the movement prediction mechanism as an additional parameter for handover decision procedure. This parameter is sent in the system information broadcast of serving cell. This decision mechanism can be applied on all handover scenarios.

Knowing the current position and velocity of an UE can obviously help to estimate where the UE is heading, thus the next position of UE to where the handover might be performed can be predicted.

In this handover decision procedure, it is assumed that the UE is able to periodically (e.g., every 1s) send its position to the serving cell (either eNodeB or FAP) during its moving. In the mean time, the serving cell maintains database of all possible target cell to where the handover might be performed. The probability of transition from one cell to another is modelled as a Markov process as approximated in (1):

$$\boldsymbol{p}_{\mathbf{n}} = [\boldsymbol{p}] \times [\mathbf{P}_{\mathbf{n-1}}] = [\boldsymbol{p}_{\mathbf{n-1}}] \times [\mathbf{P}] \quad (1)$$

where $\boldsymbol{p_n}$ is denoted as the probability of UE's position after $\boldsymbol{n}$ transitions, $\boldsymbol{p}$ is the initial distribution matrix, $\boldsymbol{P_{n-1}}$ is

denoted as current transition probability matrix, p_{n-1} is the initial distribution after n transitions and P is the original transition probability matrix. Detail of mobility prediction method for optimized handover process can be found in [10].

Using this method, the likely path of an UE can be estimated in advance, so both the handover probability and the remaining time before handover can be derived.

Upon receiving the prediction result, serving cell seeks all possible target cells. One of the neighbouring cells is assigned as the predicted target cell, to where the handover is triggered. Serving cell then performs coordination with the predicted target cell via backbone. If the target cell is available for handover, the UE will proceed the handover process.

E. Proposed Proactive and Reactive Handovers

Since the handover procedure may be initiated by either the cell (eNodeB/FAP) or the UE, therefore two handover strategies i.e., proactive and reactive handover [11] [12], are proposed to be applied to trigger the handover

a) Proactive Handover

In the proactive handover strategy, the handover may occur any time before the level RSSI of current eNodeB reaches the handover hysteresis threshold (HHT). The proactive handover strategy attempts to estimate network characteristics of a specific position before the UE reaches that position. Assumed the UE discovered that the new target eNodeB's RSSI (or FAP's RSSI) overpasses the origin one from its serving eNodeB/FAP. The UE calculates the time left before the normal handover is triggered, then triggering the handover earlier before HHT. This strategy is expected to minimize packet loss and high latency during handover.

b) Reactive Handover

Due to small FAP's coverage, its lower power and the density of FAPs, the UE in femtocell system will facing the very frequent and unnecessary handover since the UE will move from one FAP to other FAP repeatedly. To mitigate the overhead of handover, the reactive handover scenario is applied. Reactive handover tends to postpone the handover as long as possible, even though it has discovered the new RSSI signal. The handover is triggered only when the UE (almost) lose its serving eNodeB/FAP signal.

IV. HANDOVER PROCEDURE AND SIGNALLING FLOW

The LTE-based handover procedure within the femtocell network is obviously intended to minimize the handover interruption time. The handovers are also designed to be seamless when occur to/from other technology platforms (2G/3G, WiMAX, etc.).

Several functional elements take part during the handover process. The evolved UMTS Terrestrial Radio Access Network (E-UTRAN) is the key element since it provides all system functionalities included the physical (PHY), medium access control (MAC), radio link control (RLC), and packet data control protocol (PDCP) [13]. It consists a single node i.e., eNodeB or HeNB/FAP. It also provides radio resource control (RRC) functionality that corresponds to handover procedure.

E-UTRAN interacts with the Evolve Packet Core (EPC) system that consist the Mobility Management Entity (MME), Serving Gateway (SGW) and Femto Gateway (Femto-GW). The interaction between all functional elements of EUTRAN and EPC is depicted in Fig. 4.

Mobility Management Entity (MME) is the key control node for the LTE access network [13]. In handover process, MME is responsible for choosing the serving-Gateway for an UE at the initial attach and at time of intra-LTE handover involving Core Network (CN) node relocation.

Another element that takes part in handover process is serving-Gateway that responsible to route and forward user data packets. The serving-Gateway is also acting as the mobility anchor for the user plane during handovers and as the anchor for mobility between LTE and other 3GPP technologies.

The last element is called Femto Gateway that provides the gateway through which the FAP gets access to mobile operator's core network. Femto-GW is responsible for protocol conversion and also creates a virtual radio network control (RNC) interface to the legacy network without requiring any changes to CN elements. It is physically located on mobile operator premises [14].

In addition, 3GPP also specified two standard interfaces i.e., X2 and S1 interfaces, for the Evolved Packet System (EPS). The X2 interface provides capability to support radio interface mobility and shall support the exchange of signaling information between eNodeB macrocells. Therefore, for handover between eNodeB macrocells, the procedure is performed without EPC involvement. Preparation and exchange of signaling flows in the handover procedure are directly between eNodeB using X2 interface. On the other hand, the S1 interface supports many-to-many relations between EPC's elements (MME/SGW) and eNodeB. Moreover S1 is also used for the communication

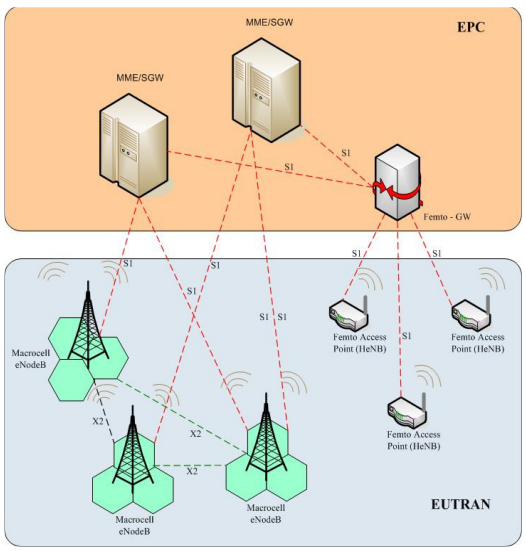


Figure 4. E-UTRAN Architecture including the deployment of Femtocell

between FAP/HeNB with the MME/SGW through the Femto-GW. Specifically, the connection to MME is using S1 control plane (S1-C) interface and the connection to SGW is using S1 user plane (S1-U) interface.

The handover within eNodeB macrocell can happen without restriction. In contrast for FAP, since the CSG is applied not every UE can access the FAP. The handover procedure consist a set of signaling flow that exchanging from one element to others. In case of the proposed handover scenarios, we also proposed the typical signaling flow for each scenario.

A. *Hand-in Procedure*

The handover from macrocell into femtocell is quite demanding and complex since there are hundreds of possible targets FAPs. In hand-in procedure, the UE needs to select the most appropriate target FAP. The interference level should be considered as a decision parameter. Moreover, the proposed mobility prediction is also considered in handover decision to optimize the handover procedure. The signaling flows of the proposed handover procedure for hand-in can be shown in Fig. 5.

B. *Hand-out Procedure*

Handover procedure from FAP to macrocell eNodeB is relatively uncomplicated. The UE has no option to select the target cell since there only the macrocell eNodeB. When the RSSI from eNodeB is stronger than FAP's RSSI, the UE will be connected directly without a complex interference calculation and authorization check as in hand-in scenario. The handover signaling flows is depicted in Fig. 6.

C. *Inter-FAP Procedure*

The procedure for inter-FAP handover is similar to hand-in procedure since the UE will facing hundreds of possible target when out of its serving FAP. For this procedure we also proposed the mobility prediction as the handover decision policy.

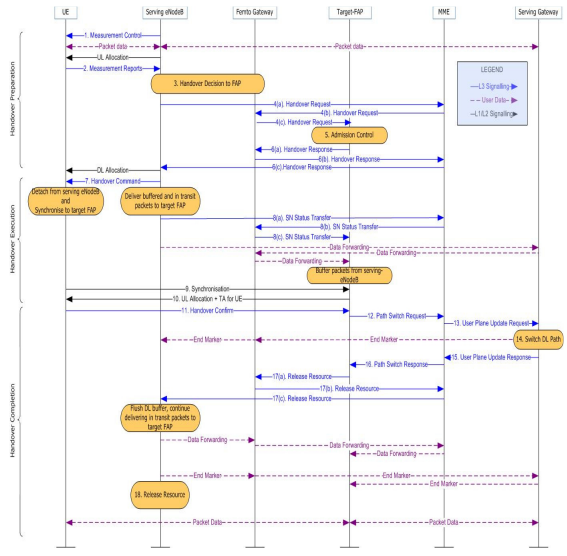


Figure 5. Signaling flow of hand-in (handover from macrocell to femtocell)

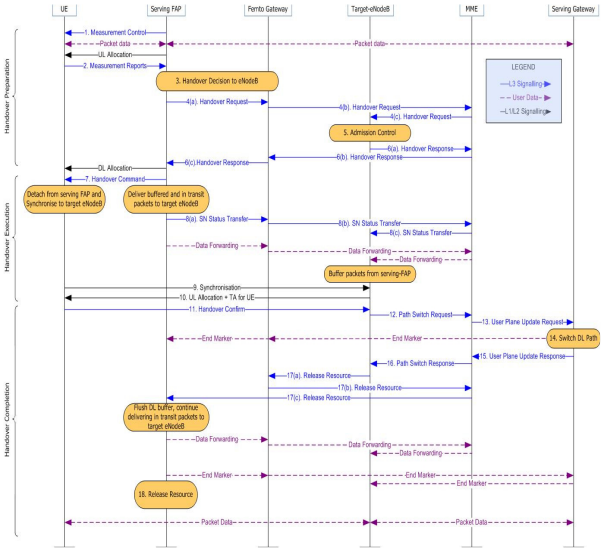


Figure 6. Signaling flow of hand-out (handover from femtocell to macrocell)

V. HANDOVER OPTIMIZATION

A. *Optimization Algorithm*

As already discussed, hand-in and inter-FAP is more complex than the handout. The main constrain on those scenarios is the handover interruption time due to delay on selection of target FAP. Another issue is the possibility of unnecessary handover and the very frequent handover due to the small coverage and low power of FAP.

To cope with these constraints, we proposed the mobility prediction method as mentioned in previous section. Knowing in advance where an UE is heading allows the system to take proactive steps. The mobility prediction mechanism often involves investigation how UEs physically move and it can estimate the final position of the UE. Once the final position of the UE is predicted, then the system will or decide to perform the handover to the nearest available FAP. This method will enhance the conventional handover decision mechanism which is based only on signal quality (RSSI/CINR) and QoS. The unexpected impact of handovers can be mitigated by deploying the reactive handover. In [4]

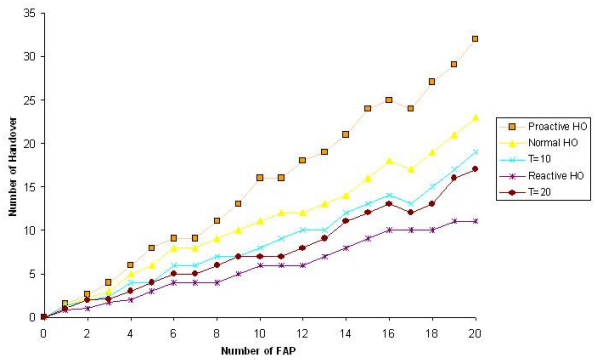


Figure 7. Preliminary performance of reactive/proactive handover

the call admission control has been proposed by forcing the UE to stay for the particular time at the new connected FAP (typically 10 seconds and 20 seconds have been assigned as the threshold time). In reactive handover, the handover will be postponed as long as possible until the UE reach the target FAP as the result of mobility prediction.

The pseudo code of optimization algorithm can be seen below. For the UE speed, we consider the maximum speed of 10 km/h.

```
1.  INITIALISATION # HO algorithm
2.  EXAMINE V       # V is the speed of UE
3.  IF V > 10 Km/h
        NO HAND-IN
4.  ELSE IF V> 5 Km/h
        PERFORM MOBILITY PREDICTION
        IF Traffic = Real-Time
                PERFORM PROACTIVE HO
        ELSE IF Traffic = Non Real-Time
                PERFORM REACTIVE HO
5.  ELSE IF Traffic =Real-Time
        PERFORM PROACTIVE HO
        IF Traffic = Non Real-Time
                PERFORM REACTIVE HO
6.  ELSE
        PERFORM NORMAL HO
    RETURN
```

B. *Preliminary Result and Discussion*

In order to verify the performance of handover procedure, the proposed algorithm has been analyzed. We manage some assumption regarding the UE mobility and the femtocell. The movement prediction of UE is approximated based on Markov-chain as stated in (1). The random way point has been used for the mobility model. Number of FAP is assumed 20 and 3 for eNodeB macrocell. The shape of FAP coverage area is assumed to be circular, with the coverage radius equal to 10 m. All entities are located in the area of 1 Km^2 with non uniform FAP density. In addition, though the random waypoint mobility model is used in the prediction process, it can be assumed that the UE does not walk randomly, but rather several paths could be followed. The result based on Matlab simulation is depicted in Fig. 7.

As can be seen, the number of handover increases almost linearly when the number of FAP is increased. The reactive handover has the lowest number of handovers compared to other schemes. Though it has proven that the performance of reactive handover is better to mitigate the unnecessary handover, the further study is still needed when this algorithm is integrated with the RF and traffic criteria that have been assigned as the handover initiation policy by the 3GPP standard.

VI. CONCLUSION

In this paper the handover procedure on LTE-based femtocell has been investigated and analyzed in three different scenarios, i.e., hand-in, hand-out and inter-FAP. The hand-in and inter-FAP scenarios are quite demanding than hand-out since plenty of target FAPs were involved in the handover process. It is a challenge to make a selection of the target FAP. The mobility prediction mechanism can be used to predict the heading position of the UE and then estimate the target FAP to which the UE may be connected. The reactive handover is the potential mechanism to mitigate the unnecessary handover. The further work is needed to find the most optimize handover procedure by integrating the proposed scheme and algorithm with the handover decision criteria specified by the standard.

ACKNOWLEDGMENT

This paper is part of research work that supported by the FP7 ICT-248891 STP FREEDOM project funded by the European Commission.

REFERENCES

[1] Broadband Forum TR-196, “Femto Access Point Service Data Model”, April 2009.

[2] H. Kwak, P. Lee, Y. Kim, N. Saxena, and J. Shin, ”Mobility Management Survey for Home-e-NB Based 3GPP LTE Systems”, Journal of Information Processing Systems, Vol. 4, No. 4, pp. 145-152, December 2008. ISSN 1973-91 3X

[3] H.Y. Lee and Y.B. Lin,”A Cache Scheme for Femtocell Reselection”, IEEE Communication Letters, Vol. 14, No. 1, pp. 27-29, January 2010.

[4] M.Z. Chowdury, W. Ryu, E. Rhee, and Y.M. Jang., ”Handover between Macrocell and Femtocell for UMTS based Networks”, in proceeding of 11th International Conference on Advanced Communication Technology, Gangwon-Do, South Korea, 2009. ISBN: 978-89-5519-139-4.

[5] H. Zhang, X. Wen, B. Wang, W. Zheng, and Y. Sun, “A Novel Handover Mechanism between Femtocell and Macrocell for LTE based Networks“, in proceeding of 2nd International Conference on Communication Software and Networks, pp 228-231. IEEE Computer Society, 2010. ISBN: 978-0-7695-3961-4/10.

[6] 3GPP-TS36.300 v8.5.0, “E-UTRAN Overall Description”. 2008.

[7] 3GPP- TS 23.401 v9.4.0, “GPRS Enhancement for E-UTRAN Access”. 2010.

[8] D. Chambers, “Which Handover Modes do Femtocells Need First?”, Think Femtocell, 2008. Can be accessed at http://www.thinkfemtocell.com/System/which-handover-modes-do-femtocells-need-first.html. Last accessed on May 2010.

[9] D.N. Knisely, T. Yoshizawa, and F. Favichia, ”Standardization of Femtocells in 3GPP”, Femtocell Wireless Communications, IEEE Communications Magazine, September 2009. ISSN: 0163-6804.

[10] A. Ulvan, M. Ulvan, and R. Bešťák, ”The Enhancement of Handover Strategy by Mobility Prediction in Broadband Wireless Access”, In *Proceedings of the Networking and Electronic Commerce Research Conference (NAEC 2009)*. Dallas, TX: American Telecommunications Systems Management Association Inc., 2009, pp. 266-276. ISBN 978-0-9820958-2-9.

[11] P. Bellavista, A. Corradi, and C. Giannelli, “Adaptive Buffering based on Handoff Prediction for Wireless Internet Continuous Services”, The 2005 International Conference on High Performance Computing and Communications (HPCC-05), Sorrento, Italy, September 2005.

[12] A. Corradi, P. Bellavista, and C. Giannelli, “Mobility Prediction Project”, Universita degli Studi di Bologna. http://lia.deis.unibo.it/Research/SOMA/MobilityPrediction/. Last accessed on May 2010.

[13] Motorola, “Long Term Evolution (LTE): Overview of LTE Air-Interface“. Technical White Paper. Can be accessed at http://business.motorola.com/experiencelte/pdf/LTEAirInterfaceWhitePaper.pdf . Last accessed on June 2010.

[14] 3GPP TS 22.220 v9.4.0, “Service requirements for Home Node B and Home eNodeB”. 2010.

 ISBN: 978-1-61208-100-7

The Collaborative Gaming for Business in Pervasive Networks

Kazuhiko Shibuya
Graduate School of Business Sciences, University of Tsukuba, Japan
CQC01205@nifty.ne.jp, sibuya@gssm.otsuka.tsukuba.ac.jp

***Abstract*— Ubiquitous collaboration should be provided for pervasively children and business people in not relation to their socio-economical conditions. This ubiquity is freely and openly to access various chances and resources in networked communities for global citizens. New business and knowledge based economy can be engendered by these active people. In this regard, I will discuss needs, backgrounds, design, and practices of gaming collaboration for educing business senses. It was designed as one of the network based ubiquitous collaboration (e.g. *Ubiquitous Jigsaw*). I will articulate the Ubiquitous Jigsaw toward organizational development in business. And finally all things considered, I introduce practices of business gaming using this collaboration.**

Keywords; Collaborative Learning, Gaming, Organizational Development in Business, Ubiquitous Computation

I. LEARNING OPPORTUNITIES FOR YOUTH INTERWEAVING WITH GLOBAL NETWORKS

Our global society has been underlying in invisible networked linkages interweaving with enormous others who belonged in the global village. At the end of past millennium, Abowd et al had proposed the three goals of ubiquitous computation such as *natural interfaces*, *context-aware applications*, and *automated capture and access* [1]. Fortunately in technical revel, ubiquitous computation has been widening its research areas and adaptable ideas enriched by technical progress during this decade. These technologies provide citizens to achieve more effective mobile communication, wireless networks and interconnection anytime and anywhere. Indeed, the ubiquitous computation requires managing both human and physical networks.

Network oriented services can heighten its potentials for not only mutual collaboration but acquaintanceships for matching people and positions in business and social life. Collaborations in online groups will be perceived as core of the feasibilities of networked cooperation (e.g. Wikipedia). Personal networks and interconnections with possibly other partners will be exactly the very fundamental core in business as well as daily life. And these available activities will be more pervasive and heighten the affinity of our life.

Otherwise it goes without saying that we can already start new business through internet commerce using some handy devices and we can shake hands with various business partners beyond physical distance and our cultural differences. However, it is quite necessary that we enforce people and children to not only handle with these intelligent devices and services but cultivate business minds. As OECD [8] and the Pearson's studies [9] had said that education of computer skills and world-wide internet based education would be one of the most favorite fundamentals for future children. Similarly, Yunus [14] also said that computational power using mobile phone can be expected as one of key successes factors against poverty in developing countries. As they make sense actually the nature of communication and socially oriented activities, efficient actual experiences on the bases of computational educations will be also beneficial to take further steps such as ventures and other borderless business for young generations in not only developing but developed countries.

However, ubiquitous computation is still higher cost than any other solutions and more difficult for little pupils and elderly people in both developing and developed countries.

So then, as I will answer to these requests, I've designed the Ubiquitous Jigsaw [12][13]. And I paved more reasonable way to experience collaborative learning that can be operated by low-cost and interactively styles on demands. For example, these are *educational gaming* and *business simulation styles* for not only business people but ordinary citizens and pupils without any expensive devices and computational power, when educators can't prepare much educational content and attune learning conditions in advance because of cost, time, human factors and others.

Further I often proposed that educators should append practical education about networking and collaborative activities among heterogeneous members whether they possibly receive computational services or not. That is because there are strong needs to teach informational techniques and fundamentals of business senses as well as harmonized coordination beyond their cultural and social differences among a lot of participants and partners all over the worlds. Namely, they have needs of a collaboration tool for learning opportunities to enhance business competences and to inspire network based business for the youth and other ambitious business leaders.

In these regards, I intend to articulate my collaborative learning (e.g. *Ubiquitous Jigsaw*) and its applied gaming for business on the bases of these motivations after next chapter.

II. THE UBIQUITOUS JIGSAW AS NETWORK BASED COLLABORATION

A. The Goal for this Collaborative Learning

I had implemented the Ubiquitous Jigsaw methodology that actualizes collaboration in networked community [12] [13]. It is one of the computer supported collaborative learning (CSCL) and refined educational activities of the Jigsaw classroom [2] on the bases of technologies of ubiquitous computing [1][16]. As widely known, an original collaborative learning of this style has been providing pedagogical solutions for cultural and racial conflicts and

 ISBN: 978-1-61208-100-7

cultivating cooperative minds beyond differences among students in classroom. Needless to say, Ubiquitous Jigsaw had inherited from this fundamental concept, and it had been conceptualized to heighten the potentials for borderless and omnipresent computational networking era.

As my implementation of this software architecture, I will consider to step it forward business collaboration. Indeed, our activities in actual societies and business scenes are very dynamic rather than static in classroom. And these situations often motivate us to prepare both those minds for cooperation and computational skills for adaptation in global competitive and/or collaborative relationships very much.

Nowadays, informational resources for business already require three factors such as *data resources*, *fundamental information architectures*, and *human resources on the bases of business and computer competences* at least. Data resources can be easily managed by computational systems using cloud computing, online knowledge-database (e.g. Wikipedia, digital libraries, open course-ware, and etc.) and applicable data centers, for example. Besides fundamental information architecture can be also depend on a part of these services, and mobile devices (e.g. mobile phone, laptop computer, iPad and etc.) based communication is conceived to heighten dynamic management and collaborative activities in business. However, only human resources must be trained by appropriately practical ways, and properly interconnection and its rewiring among a lot of experts and partners is the very fundamental core of business in competitive environment. Hence, business competences can be also defined following standpoints. Here, these are enhancement of own core competences, finding job, interconnecting with preferred partners in business, adaptation in organization, and innovative creativity in heterogeneous group.

So, the learning service of the Ubiquitous Jigsaw aims to cultivate the third factor as business competences. And a part of computational services in Ubiquitous Jigsaw can be depended on the first and second factors as ubiquitous oriented framework in particular [12]. Participants will experientially understand those points by this collaborative learning activity through social and computational networking.

B. The Framework of Ubiquitous Jigsaw

In brief, the collaboration of Ubiquitous Jigsaw has a potential to enhance collaborative activities more effectively [13]. Apparently, collaborative learning consists of various objectives to encourage each participant's motivation and understanding to a greater extent than an ordinary learning style and environment. It is noteworthy that collaborative learning in ubiquitous computational environment can engender more *interactive*, *experiential*, *spatiotemporal*, and *distributed* aspects for those who want to learn and solve problems while coordinating with others. Applying methodological and technical concepts, coordination with computational implementations can enhance assistance with more appropriate educational services from each individual level to partners, group, and collective level. Excellence of this methodology can outperform traditional collaborative works to show an obviously pervasive combination with current computational theories.

Even though many past researchers have explored collaborative learning and related computational support in diversified contexts and academic disciplines so far, collaborative learning has yet to be explored widely in terms of its pervasive possibilities. Learning environments can not be restricted physically, spatiotemporally, or in terms of boundaries in this advanced networking era. In addition, ubiquitous computation includes some subcomponents such as location based services, social context awareness oriented significance, spatiotemporal dynamics of human behavior, and integration with other machinery and computer technologies. These constructs of ubiquitous computation are applicable to educational services.

A lot of educational networks such as online learning communities and e-learning constructs have inferred to be based on reciprocal acquaintanceships. The online learning community has been investigated in terms of traditional perspectives. However, I will intensify the learning community to include novel features such as reciprocity and self-organizing patterns, for example. Reciprocity is the basis of mutual trust and well-disposed understanding; at least one educational purpose would nurture the sensibilities of children using these satisfactory interactions. Moreover, self-organizing perspectives in social networks suggest the actual cutting edge of the network. I envision collaborative peer to peer (P2P) like network patterns using mobile devices. Such a network is likely to generate and change its state and structure autonomously because mutual relationships in collaborative learning should form the foundation of students' personal initiatives, rather than compulsive ones. In fact, reciprocity and the self-organizing nexus are two facets of a collaborative-learning network. For those reasons, I believe that collaborative learning must include reciprocal relationships and a self-organizing network structure.

III. FOR GROUNDING ON THE BUSINESS

A. Educing Business Senses and Motivations

Next, I should articulate both academic backgrounds and practices of this gaming collaboration for business. Actually, grounding of fundamental educations for cultivating business competences is quite necessary.

Generally speaking, I consider that recent career educations may lack some considerable factors. Namely, these are the network principle and social selection by preferred partners. The former is fundamental human relationship in social and business activities; otherwise the later is individual preference as well as reflexive mind by standing on the other's view. Both it seems that these factors are essential to motivate development of the youth.

So, as mentioned before, I am willing to apply a framework of the Ubiquitous Jigsaw for organizational development and relative learning activities in business. The concept of collaborative learning will be very harmonious with internet based services and cooperative education. This functionalizes as the platform of networked collaboration for training and brushing up business senses. Especially, I

 ISBN: 978-1-61208-100-7

discuss following four issues after this section in order to understand theoretical backgrounds of my business collaboration.

1.*Social Adaptation*

As cooperative activities, students may reflectively understand the meaning of social networking and group adaptation in terms of others' perspectives.

2.*Interconnections*

Know-How and Know-Who as social interconnections with preferred partners.

3.*Heterogeneity and Creativity*

Management of heterogeneous group members and enhancement of group creativity.

4.*Innovative Mind*

Can collaborative activities lead young students into educing innovative mind for business?

B. Social Adaptation: To Realize it by Networked Collaboration

First, the collaboration is to literally cooperate and co-work with enormous other members within groups and societies. In social and organizational psychology, networked collaboration for this purpose will envisage a core problem on social adaptation to be group members.

Needless to say, as we already known very well, social relationships is the very fundamental for social adaptation. Social network can be almost considered that social members are often dynamically connecting by each person's preference and judgment for the sake of establishing appropriate relationships. Absolutely, social network in business is a kind of implicitly or explicitly selection process among preferred partners. And these social ties can enlarge to social structural conditions such as various communities and group hierarchical relationships. The preference-based selection has been clarified by many empirical findings. For example, Robins,G, et al [10] had discussed and exemplified their findings of social preference and selection process of preferable partners in social networks.

With this in mind, Leary,M.R&Baumeister,F.R [6] recently discussed using their theory of socio-meter. Their theory supposes some hypotheses on both self-esteem and social adaptation in social relationships. It describes that mental awareness of oneself is affordable to amplify and estimate other's mind in society. That is, theory hypnotizes that this function was acquired in evolving process in order to monitor other mental conditions for own adaptation. In other words, self-esteem, self-consciousness and identification are assumed to be symmetry mirror between self and other members.

So then, social selection and exclusion in the community is often quite serious problem for adaptation of group members. Social members in their daily lives must seek to join more preferable and valuable social relationships. Namely a part of their social adaptations is dependently underlying in other member's evaluation and preferences, and vice versa. As social preference is one of critical factors for various behaviors in social relationships, especially we can select favorite partners to satisfy with our self-esteem and motivation. For example, Rudich&Vallacher [11] had found some suggestions that subjects were apt to choose an interaction partner in order to maintain own self-esteem and self-enhance motivation.

In these concerns, we can expect that students will realize to commit affiliation networks and they will maintain and improve their conditions if they will be educed properly. So it can provide possibly experiential learning services not only collaborative working but job seeking activities by computational assistances on demand.

C. Interconnection: Know-How and Know-Who as Preferred Partners

Here, it is worth saying that I can point out a fact on the power of interconnection in business. First, in business and daily life, we sometimes rewire effective pathways to exchange social capitals, information and knowledge in the process of reorganizing human relationships. As suggested before, social networks include fundamental processes of social selection to interconnect with preferred partners. Preferred partners are almost acknowledged as professionals, experts, business partners and favorable key persons in order to achieve specific business goals. So then, it should rewire and interconnect with nodes as those partners if we hope to success in business. And students should learn these factors if they wonder how to interconnect with those partners who have abilities to achieve business tasks by more effectively ways.

Secondly recruit and employment are indeed to append or rewire social nodes with preferred partners in human relational networks. So, job position and career are based on network principle and social selection in order to interconnect with relationships [4]. It is a key factor to obtain new positions and open the gate for another career path. This is also beneficial for business education using computational services as suggested earlier. Being social member literally means to be an interconnected node in relationship nexus among other peoples, organizations and societies. Hence, students are needed to be preferred candidates as nodes in social relationships before job seeking. Without saying, applicants are required to attain properly social skills, abilities, potential, and not limited to these personal properties if they want to be hired. Hence it is necessary that we encourage the youth (and of course older people) to brush-up social skills and enhance their potentials, motivations and social abilities in social relationships.

As an engineering case, Morisue,M, Nakano,Y &Tarumi,H [7] had developed a tool for job hunting. But it was not enough that they designed it standing on this social network principle. There are already information providing sites (e.g. Linked-In and etc.), it may be inspired from these services.

Comparatively, the Ubiquitous Jigsaw should not only just provide much job information for university students. In contrast, this framework aims to inspire effective chances for collaborative business such as LLP (Limited Liability Partnerships), LLC (Limited Liability Company) and NPO activities interconnecting with other members and experts.

Exactly human resources itself can be perceived as a kind of knowledge base (sets of know-how and expertise skill and memorized data), and it is more valuable than job

 ISBN: 978-1-61208-100-7

information itself. These styles can not success unless participants cooperate with each other, namely it is a good example case for collaborative learning by the design of Ubiquitous Jigsaw. Further it has possible advantages for getting openly assistances to interlink with unlimited experts and unknown partners as well as online knowledge bases (e.g. digital libraries) through pervasively services.

D. Boosting up Organizational Uncertainty and Heterogeneity to Group Creativity

Taken together, I also expect organizational potentials and group creativity in networked collaboration. In due process of both educational and business purposes, there are needs to enhance members' creativities and educe potentials of participants. Indeed, networked globalization engenders more heterogeneous and complex conditions in multiple levels from personal to social and business situations [3]. Heterogeneous network per se holds attractive potentials to engender novel idea and breakthrough in business and collaborative activities. Heterogeneities of social network consist of different and independent individuals who embrace their original attitudes, opinions and backgrounds. Heterogeneous and uncertain factors will be possible to facilitate group creativities and potential of members, but it often contains some hardly troubles.

Absolutely, in global era, it is quite necessary to coordinate with ambivalent process of both remaining heterogeneously and even reducing uncertainty in group and organization. The traditional jigsaw method and cooperative activities had to conquer the differences of cultural, socio-economical, racial and other social properties [2], and they reported that their method was very effective to improve despairingly conditions in education. Similarly, as collaborative learning by the Ubiquitous Jigsaw will be also expected to harmonize with those ambivalent conditions, these heterogeneous conditions will inspire participants to conquer any troubles caused by group uncertainty and they regenerate it as dynamism for new business.

E. Comparing with SECI Model: From Collaboration to Innovative Minds

As already known, there is one of the most famous key tools named the SECI model, which proposed by Nonaka,I, for knowledge management in business organization and activities. The concept of Ubiquitous Jigsaw collaboration and the SECI model are not irrelevant to each other. Both concepts devoted to enhance the collaboration in inner and inter groups and communities underlying in actual knowledge and expertise linkages. The former focuses on the educational collaboration and sharing experience among people in ubiquitous and pervasive networks. In contrast, the later articulated the knowledge itself such as implicit and explicit aspects from individual to group, and theorized as an originally framework of knowledge based management in business process.

In similarly, a learning design by communities of practice [17] [18] apparently shares the same root of both cooperative designs of activities in social contexts. But contrary, Ubiquitous Jigsaw can be designed to manage more pervasively computational assistances and online services than those two models in advance. So then, as taking mobile devices and like this, collaboration style by Ubiquitous Jigsaw can be regarded as seamless bridging between actual and networked communities interweaving with pervasively other participants through internet.

IV. COLLABORATIVE GAMING PRACTICES FOR BUSINESS

A. Design And Practices

Here, I introduce detail practices of handy gaming collaboration on the bases of methodological backbone of the Ubiquitous Jigsaw as noted earlier chapter. Namely, it is an applicable services named as *business gaming collaboration*. It was designed to be openly for peoples' participation and avoided affecting their socio-economical backgrounds. And it originated interactional gaming either with or without special devices, it can be adapted as a collaborative gaming in both online and offline conditions.

Gaming itself has been nowadays widely recognized as an experiential learning style assuming a lot of social contexts. This collaborative gaming can be characterized by as following four types of *cards* such as *Task, Data Access, Hint,* and *Ability*. These types of cards are represented specific meaning such as practical allocated assignments, approval for special abilities (represented as own competences), approval for extra database access (represented as knowledgebase and know-what), and providing some hints (represented as know-how) respectively. Educator prepares printed cards or electrical ones in advance, and these are allocated for each small group and participatory student.

- Task Cards
  - These cards are assigned for each student, and they must achieve tasks printed on each card. For example, educators prepare cards on the bases of big themes such as "*The World Heritages in Europe*", and "*Your living City*" in advance. Each assigned card should be considered difficulty level, time-consumption, and other factors, as educator like. And finally, each group shall finish writing a report combining and integrating with achieved tasks by each participatory student on the whole. It is likely jigsaw-puzzle.
- Data Access Cards
  - These cards are represented as "Know-What" and access for knowledgebase. It can permit to access specific knowledgebase such as encyclopedia and internet resources for solving each task. Participants learn the nature of knowledge-intensive economy.
- Hint Cards
  - These cards are represented as "Know-How" and it permits students to teach any other student and he can obtain Points from him. Of course, any students can help other students without this Hint Cards, but in this case the total cost of payment to reward

 ISBN: 978-1-61208-100-7

Points for helping student is higher than use of Hint Cards.

- Ability Cards
  - These cards are represented as "Own Competences" and it allows students to perform extra-powerful specialties illustrated in each card only once. There are many types of cards such as following, for instance.
  - New Interconnection
    - Permitting to interconnect with one partner from other group members. If successfully, they can cooperate with each other during interconnection. And this linkage can also enrich them to adapt more than disconnection. Of course, as you still realized, this card is represented as "Know-Who" and human relationship among business partners. Literally this interlink implies interconnection among experts like LLP and internet based partners.
  - Void of Interconnection
    - Canceling and disconnecting an interconnection between already interlinked groups. When you have this card, you can obstacle for any other group opponents and interlinks. It means cooperation and competitions in business process.
  - Rewiring Interconnection
    - Permitting to rewire with an existing node and new node. It likes Watts's Small World model. When you have this card, you can take a node between any other group opponents and interlinks by force. It means cooperation and competitions in business process.
  - Assigning Extra Tasks
    - As you have this card, you can force an opponent to draw extra task cards from a deck of task cards. So then, this opponent will not be able to finish faster than your group.
  - And a lot of ability cards.

Participatory students cooperate with each other among in-group members in order to achieve own assigned tasks in this process of gaming collaborative learning. Further they strategically decide how to interconnect with other group partners and be competitive against other groups. Through gaming process, participants can discuss their assignments in belonged group and behave competitively among other groups. They also can obtain and interexchange cards and Points with other out-group partners in classroom and business workspaces if possible. And they often access internet resource and database in order to accomplish their assignments. Finally, the most performed group that finished faster than any other groups and obtained by totally Points will be a champion in gaming.

As figure 1 illustrated a part of gaming process in the condition using computational services, contacting partners as appropriate 'Know-what' and 'Know-who' nodes who are reliable experts is a foundation of ubiquitous jigsaw using mobile devices in a socially networked relation if possible. In addition, I would like to explore the way in which these social network structures can emerge. Social relationships have been considered as a kind of 'small world phenomenon' [15]. Collective dynamics of social network and its 'small world' patterns may also appear in collaborative learning networks. Sharing and diffusing experiential knowledge is easier than in a regular network structure if a small world structure can be found everywhere. Furthermore, appropriate social networks can lend the comfort of educational support to solve problems and understand unfamiliar things.

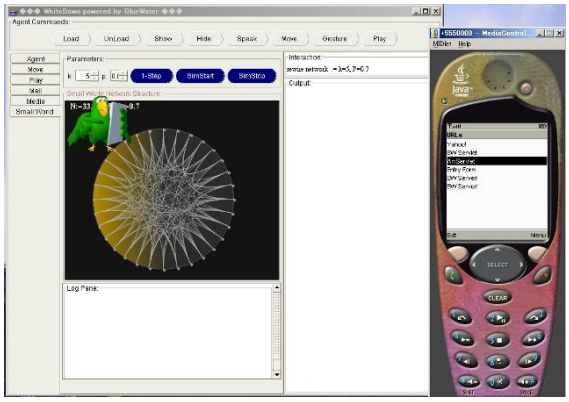


Fig. 1: Testing for Interpersonal Networks

As you can easier imagine a business model like LLP, various linkages as interconnected networks among experts holding high professional skill and competences will be good leading for taking-off own business. Therefore, the current design of experiential collaboration paved the way to facilitate understanding and acquiring fundamental business skills and abilities for students.

B. Clalifying Experiential Data

Methodology has to equip own standard and procedure for application, utilization and evaluation. This collaborative gaming should be also institutionalized the canonical way. Especially, data quantification for evaluation is certainly significant for the quality control of gaming. So, I have been conducting to standardize a first version of this collaborative gaming in various conditions and small-groups (three or five Japanese participants in each group).

Here, the most importance is that survey and observation data in this gaming can be examined in terms of the *Top Ten Open Research Questions* proposed by experimental economist Camerer,C.F [3]. These questions were originally game theoretical bases, but it is able to apply widely toward social experimental studies. And especially his cast question 5, 8, 9 and 10 are relevant to the specification of Ubiquitous Jigsaw. Namely this collaboration consists of participants' cognitive decision making, heterogeneous memberships as team work, solving complicated educational assignments and multicultural differences. Other collaboration and gaming were not able to clarify participants' mental factors properly. As analyzed results in terms of considerable questions, it

 ISBN: 978-1-61208-100-7

could unveil a part of participants' cognitive and behavioral factors more detail.

5.What exactly are people thinking in games?

8.How do teams, groups, and firms play games?

9.How do people behave in very complex games?

10.How do socio-cognitive dimensions influence behavior in games?

After this gaming, I could verify the experiential results such as quality data that observed participants' activities in gaming and quantity data by few answers in survey after gaming. And many of them felt performed well, and I got answers that they could almost understand the nature of collaboration by means of this gaming. Hence, these data of results could shed light on the top ten questions as I listed above.

Further I have a plan to translate the cards of this gaming for multilingual version (current version is almost written in Japanese). And I will prepare both printed cards and electrical online ones to attempt for pupils in developing and developed countries and multicultural contexts in future works.

C. Discussion

The nature of this collaborative gaming will be also adapted interoperating pervasively. This thinking paradigm exactly matches with open innovations in business. That is because business models such as LLP and internet based business are realized as interlinks of many expertise across professionals. And any business scenes can not be imaged without any networking and interactively skills. And computational architecture and massive educational contents has been decreasing its costs, and many of these services will be provided freely by cloud computing or online resources in business.

The ubiquity is freely and openly to access various chances and resources in networked communities for citizens of all over the worlds. The highest merit of this gaming collaboration can be practiced to open easier participation for youth, pupils and any citizens without any expensive devices and difficult constraints for learning opportunities. It is the very good fit for pupils' preliminary literacy education either with or without touching mobile device. And in developing countries, poor pupils can learn by the reasonable alternatives using cards. Of course, when participants can ideally receive supports by available ubiquitous computing, it can nourish more vivid experience for students and business people.

V. CONCLUDING REMARKS

Ubiquitous collaboration should be achieved by pervasively children and business people in not relation to their socio-economical conditions. This ubiquity is freely and openly to access various chances and resources in networked communities for global citizens. New business and knowledge-intensive economy can be engendered by these active people. In this concern, I have discussed needs, backgrounds, design, and practices of gaming collaboration for educing business senses by the Ubiquitous Jigsaw. Business education and organizational development will be possibly constituted in experiential daily process. These actual experiences of collaborative activities can give chances of social adaptation for people in actual contexts of social relationships. And finally I am willing to lead studies on discussed matters in future works.

REFERENCES

[1] Abowd,G.D&Mynatt,E.D 2000 Charting Past, Present, and Future in Ubiquitous Com-puting *ACM Transactions on Computer-Human Interaction 7.1.29-58*

[2] Aronson, E., Blaney, N. Stephin, C., Sikes, J. & Snapp, M. 1978 *The jigsaw classroom* SAGE

[3] Camerer,C,F 2003 *Behavioral Game Theory: Experiments in Strategic Interaction* Princeton University Press

[4] Hofstede,G&Hofstede,J 2005 *Cultures and Organizations: Software of the Mind [Re-vised and Expanded 2nd ed.]* McGraw Hill

[5] Granovetter,M 1995 *Getting a job : A study of contacts and Careers(2nd ed.)*, The University of Chicago Press

[6] Leary,M.R&Baumeister,R.F 2000 The Nature and Function of Self-Esteem: Sociometer Theory, in Zanna,M.P (Eds.) *Advances in Experimental Social Psychology Vol.32* Academic press

[7] Morisue,M Nakano,Y &Tarumi,H 2005 Job-hunting Support with Enhanced Informal Communication within a Department, *IEEE AMT-2005.199-202*

[8] OECD 2006 OECD Work on Education 2005-2006 http://www.oecd.org/dataoecd/35/40/30470766.pdf

[9] The Pearson Foundation 2010 *The Digital World of Young Children: Emergent Literacy (White Paper) http://www.pearsonfoundation.org/*

[10] Robins,G, Elliott,P&Patterson,P 2001 Network models for social selection processes *Social Networks 23.1.1-30*

[11] Rudich,E.A&Vallacher,R 1999 To Belong or to Self-Enhance? Motivational Bases for Choosing Interaction Partners *Personality and Social Psychology Bulletin 25.11.1387-1404*

[12] Shibuya,K 2004 A Framework of Multi-Agent Based Modeling, Simulation and Computational Assists in Ubiquitous Environment *SIMULATION :Transactions of The Society for Modeling and Simulation International 80.7-8.367-380*

[13] Shibuya,K 2006 Collaboration and Pervasiveness: Enhancing Collaborative Learning Based on Ubiquitous Computational Services, including as Chapter 15 (p.369-p.390), in Lytras, M&Naeve,A (Eds.) *Intelligent Learning Infrastructures for Knowledge Intensive Organizations: A semantic web perspective*, IDEA group Publishing

[14] Yunus,M 2007 *Creating a World Without Poverty* Public Affairs

[15] Watts,D.J 1999 *Small Worlds* Princeton University Press

[16] Weiser, M 1993 Some computer science issues in ubiquitous computing *Communications of the ACM 36.75-84*

[17] Wenger,E 1998 *Communities of Practice: Learning, Meaning, and Identity* Cambridge University Press

[18] Wenger,E,McDemott,R&Snyder,W.M 2002 *Cultivating Communities of Practice* Harvard University Press

 ISBN: 978-1-61208-100-7

On-the-Fly Ontology Matching for Smart M3-based Smart Spaces

Alexander Smirnov, Alexey Kashevnik, Nikolay Shilov
SPIIRAS
St. Petersburg, Russia
{smir; alexey; nick}@iias.spb.su

Sergey Balandin, Ian Oliver, Sergey Boldyrev
Nokia Research Center
Helsinki, Finland
{sergey.balandin; ian.oliver; sergey.boldyrev}@nokia.com

***Abstract* — Proper functioning of smart spaces demands semantic interoperability of the knowledge processors connected to it. As a consequence it is required to develop models that would enable knowledge processors to perform on-the-fly translation and interpretation between the internal and smart spaces ontologies. The paper presents our solution to the above stated problem, which has been implemented for Smart-M3 platform.**

Smart spaces; ontology matching; semantic similarity; semantic interoperability; Smart-M3

I. INTRODUCTION

Smart-M3 is an open source software platform [1] that aims to provide "Semantic Web" [2] information sharing infrastructure between software entities and various types of devices. The platform combines ideas of distributed, networked systems and Semantic Web [3]. The major application area for Smart-M3 is the development of smart spaces solutions, where a number of devices can use a shared view of resources and services [4], [5]. Smart spaces can provide better user experience by allowing users to easily bring-in and take-out various electronic devices and seamlessly access all user information in the multi-device system from any of the devices.

The simplified version of the Smart-M3 smart spaces reference model is shown in Figure 1. The Knowledge Processors (KPs) represent different applications that use the smart space. The smart space core is implemented by one or several Service Information Brokers (SIBs) interconnected into the common space. The information exchange is organized through transfer of information units (represented by RDF triples) from KPs to the smart space and back. The information submitted to the smart space becomes available to all KPs participating in the smart space. The KPs can also transfer references to the appropriate files/services into the smart space, since not all information can be presented by RDF triples (e.g., a photo or a PowerPoint presentation). As a result the information is not really transferred but shared between KPs by using smart space as a common ground.

However, real implementation of any smart spaces solution faces a number of problems. Let's consider a simple case study when a user is having his/her mobile device with KP running on it. Assume that the user KP (UKP) is configured to make some presentation. The LCD projector represented in the smart space by Projector KP (PKP) is a key enabler of this functionality. But in order for the presentation to be shown, the UKP has to share the information about the presentation location (URI) with the PKP. As a result the following conditions have to be fulfilled:

1) The UKP has to know that the PKP is a part of the smart space;
2) The UKP has to share the presentation's URI in such a way that the PKP can understand it.

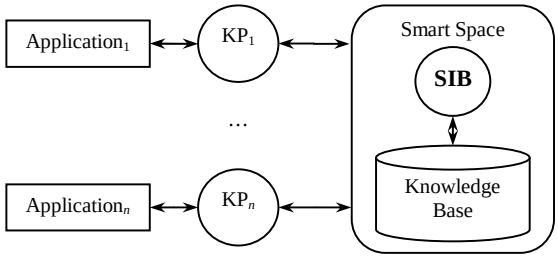


Figure 1. Smart space based on Smart-M3: simplified reference model.

Even from this simple example we can see that proper functioning of the smart space requires semantic interoperability between knowledge processors connected to it. As a consequence it is required to develop a model that would enable knowledge processors to translate on-the-fly between their internal and smart space's ontology to a certain extent. In this paper, we present our solution for the above stated problem.

II. SIMILARITY IN SMART SPACES

All the similarity metrics in the performed state-of-the-art review are based on the two information retrieval metrics of precision and recall. As it was mentioned earlier, in case of smart spaces it is necessary to avoid false alignments, so the precision should be preferred above the recall. This is achieved via choosing the right threshold value. The possibility of choosing the right threshold value has to be taken into account in the development of the matching models.

Since in smart spaces most of knowledge processors are problem-oriented it should be proposed to utilize reusable ontology patterns for ontology creation. This would enable unification and standardization of the ontologies and significantly simplify ontology matching.

Table 1 summarizes the results of the state-of-the-art.

 ISBN: 978-1-61208-100-7

TABLE 1. STATE-OF-THE-ART SUMMARY

Criteria	Possibilities / Requirements
Agent anonymity	Matching approach for different ontologies Matching of different versions of the ontology
Information/ knowledge representation format	Format independent RDF KIF Graph-based formats
Ontologies aligned	Any Lightweight Large-scale
Automation	Semi-automatically Automatically
Algorithm complexity	High complexity Low complexity $O(N^2)$, N – the number of elements in the contexts of the concept to be matched
Precision	Supported N/A
Matching method class(es)	Contextual: distribution-based graph-based structural similarity structural propagation Linguistic: similarity-based Statistical: data type compatibility Combined: pattern-based heuristics rule-based
Usage of synonyms	No synonyms Synonyms supported: Thesaurus-based WordNet-based
Ontology element matching	One-to-one Any to any entities matching One to any entities matching supported
Internet usage	Internet is used Internet is not used Internet can be used

Based on it, the following concluding remarks can be made.

The goal of ontology matching is basically solving the two major problems, namely:

1) Ontology entities which have the same name can have different meaning.

2) Ontology entities which have different names can have the same meaning.

For this purpose a number of techniques are applied in different combinations. These techniques include:

1) Identification of synonyms

2) Similarity Metrics (name similarity, linguistic similarity)

3) Heuristics (for example two nodes are likely to match if nodes in their neighborhood also match)

4) Compare sets of instances of classes instead compare classes

5) Rules: for example, if class A1 related to class B1 (relation R1), A2 related to class B2 (relation R2) and B1 similar to B2, R1 similar to R2 therefore A1 similar to A2.

As a result of matching the following types of elements mapping proximity can be identified:

1) One-to-one mapping between the elements (Associate-Professor to Senior-Lecturer)

2) Between different types of elements (the relation AdvisedBy(Student, Professor) maps to the attribute advisor of the concept Student)

3) Complex type (Name maps to the concatenation of First Name and Last Name)

All methods can be separated into the following four groups.

A. *Linguistic methods*

These methods are focused on determining similarity between entities based on linguistic comparison of these entities (count of the same symbols estimation, estimation of the longest similar parts of words, etc.).

B. *Statistical methods (instance based)*

These methods compare instances of the ontology entities and based on this estimation entities can be compared.

C. *Contextual methods*

The aim of the contextual similarity is to calculate a measure of similarity between entities based on their contexts. For example if parents and children of the ontology classes are the same consequently the classes also the same.

D. *Combined methods*

These methods combine specifics of two or three of the above methods.

In the M3 approach, there is no strict definition of instances and differentiation of them is not an easy tasks. Because of this reason the techniques and methods relying on instances were not considered for further development. Hence, the developed models presented below integrate all of the above techniques (except those dealing with instances) and propose a set of combined methods having features of the linguistic and contextual methods.

III. PRINCIPLES FOR CREATING ONTOLOGICAL DESCRIPTION FOR SMART SPACE KNOWLEDGE PROCESSORS

In this section, seven principles for creating ontological description for smart space knowledge processors are proposed. The correspondences between the principles and the criteria are indicated in Table 2.

A. *Synonyms*

Synonyms of the used in the ontological description terms have to be provided. Synonyms can be provided as additional RDF-triples. For example:

RDF Triple:
(“URI”, “is”, “http://myexample.com/pr1.ppt”)
Synonym1:
(“URL”, “synonym”, “URI”)
Synonym2:
(“location”, “synonym”, “URI”)

 ISBN: 978-1-61208-100-7

TABLE 2. STATE-OF-THE-ART SUMMARY

Criteria	Corresponding Principles
Agent anonymity	Taken into account in the matching approach
Information/knowledge representation format	Connected graph-based RDF-triples (subsec. C) Fullness and consistency of the description of the knowledge processors possible actions (subsec. E) Taken into account in the matching approach
Ontologies aligned	Fullness and consistency of the description of knowledge processors possible actions (subsec. E) Usage of ontology patterns (subsec. G)
Automation	Triples for determining format for values (subsec. F) Taken into account in the matching approach
Algorithm complexity	Taken into account in the matching approach
Precision	Taken into account in the matching approach
Matching method class(es)	Synonyms (contextual methods) (subsec. A) Abbreviations shouldn't be used (linguistic methods) (subsec. B) Connected graph-based RDF-triples (contextual methods) (subsec. C) Combined methods are used in the matching approach
Usage of synonyms	Synonyms (subsec. A) Abbreviations shouldn't be used (subsec. B) Taken into account in the matching approach
Ontology element matching	Homogeneous ontological description of elements (subsec. D) Taken into account in the matching approach
Internet usage	Taken into account in the matching approach

B. *Abbreviations shouldn't be used*

Abbreviations of the ontological description terms complicate matching of the ontologies. They should be avoided in ontological description.

Instead of using triple:

("START_D", "is", "06+11+2004") – VCal

or

("DTSTART", "is", "20041106") – iCal

The following rule should be used

("start date", "is", "2004-11-06")

If it is not possible avoiding using abbreviations (e.g., when they are standardised like in the examples above) than it is recommended to resolve this potential problem by adding appropriate synonyms for these abbreviations, for example:

("START_D", "synonym", "start date")
("DTSTART", "synonym", "start date")

C. *Connected graph-based RDF-triples*

RDF-triples have to be joined with each other. This allows to process such ontological descriptions as a graph and to apply graph-based methods of ontology matching.

("URI", "is", "http://myexample.com/pr1.ppt")
("slide", "is", "5")

It is important to connect the above triples, by complimenting database by the following additional triples:

("slide", "part_of", "presentation")
("presentation", "property", "URI")

D. *Homogeneous ontological description of elements*

Ontological description elements have to be homogeneous. For example, if an element is a subject or object in one rule it cannot be a predicate in another rule.

Instead of using triple:

("presentation", "URI", "http://example.com/pr1.ppt")

The following two rules should be used:

("URI", "is", "http://example.com/pr1.ppt")
("URI", "part_of", "presentation")

The triples describing synonyms are specific triples and they can contain predicates as subjects and objects.

E. *Fullness and consistency of the description of the knowledge processors possible actions*

Ontological descriptions of the knowledge processors have to include all their possible actions and relations, but at the same time shouldn't include any unnecessary information.

An example of projector ontology RDF triples is presented as a graph in Figure 2 and as a list of corresponding triples under the picture.

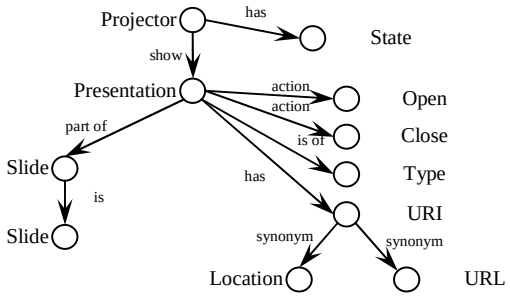


Figure 2. Ontology of the projector.KP.

("projector", "has", "state")
("projector", "show", "presentation")
("presentation", "action", "open")
("presentation", "action", "close")
("slide", "part of", "presentation")
("slide", "is", "Number")
("presentation", "has", "URI")
("presentation", "is of", "Type")

Synonyms:

("URL", "synonym", "URI")
("location", "synonym", "URI")

F. *Triples for determining values format*

In case of complex formats of data values it is needed to add special triples into the ontology, which describe value format. For example, vCal has description:

"Start_T", "value", "07+29+30")

To let the system recognize this time format it is needed to add the following triple:

("Start_T", "format", "hh+mm+ss")

G. *Usage of ontology patterns*

Most of knowledge processors are task-oriented and it is expected that there will be many knowledge processors

performing the same functions (e.g., knowledge processors representing functionality of LCD projectors). As a result utilizing reusable ontology patterns for ontology creation is proposed. This would enable unification and standardization of the ontologies and significantly simplify ontology matching. Such patterns should be problem-oriented as knowledge processors themselves. As an example of such a pattern the projector knowledge processor ontology presented in Figure 2 can be considered.

IV. MULTI-MODEL APPROACH FOR ON-THE-FLY MATCHING KPS AND SIB ONTOLOGY

The below proposed approach allows matching of KPs and SIB ontology for the interoperability purposes and is based on the ontology matching model illustrated by Figure 3. The approach takes into account that the matching procedure has to be done "on-the-fly" by mobile devices with limited resources and remembering the fact that knowledge processors are responsible for performing certain concrete and well-described tasks, which means that the corresponding ontology generally should be small–to–medium size and describe only very limited domains.

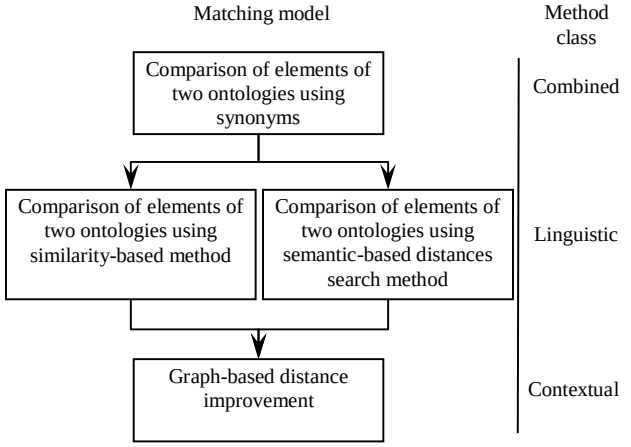


Figure 3. Multi-model approach to on-the-fly matching KP and SIB ontology.

Ontology is represented as RDF triples, consisting of the following ontology elements: subject, predicate, object. Degree of similarity between two ontology elements is in the range [0, 1]. The approach consists of the following steps:

1) Compare ontology elements taking into account synonyms of both ontologies. The degree of similarity between equal elements is set to 1 (maximum value of the degree of similarity).

2) Compare all elements between two ontologies and fill the matrix M using *similarity-based model* described in Section V. Matrix M is of size m to n, where m is the number of elements in the first ontology and n is the number of elements in the second ontology. Each element of this matrix contains the degree of similarity between the string terms of two ontology elements using the fuzzy string comparison method described in Section III-B.

3) For knowledge processors, which can access Internet, e.g., WordNet or Wiktionary, the model of searching semantic distances was developed.

a) Compare all elements of two ontologies and fill the matrix M'. Matrix M' is of size *m* to *n*, where *m* is the number of elements in the first ontology and *n* is the number of elements in the second ontology. Each element of this matrix represents the degree of similarity between two ontology elements.

b) Update values in matrix M, where each new value of elements of M is the maximum value of (M, M')

4) Improve distance values in the matrix M using the *graph-based distance improvement model* described in Section VI.

As a result the matrix M contains degrees of similarity between ontology elements of two knowledge processors. This allows determining correspondences between elements by selecting degrees of similarities which are below than the pre-selected threshold value.

The next sections describe major elements of the proposed approach in details.

V. SIMILARITY-BASED MODEL FOR ONTOLOGY MATCHING

The similarity-based model for the ontology matching is presented in Figure 4. It contains a stemming procedure to normalize words, improved fuzzy string comparison procedure, and normalization procedure. The normalization procedure makes it possible to reduce the resulting similarity for its easier interpretation and is not considered here in detail.

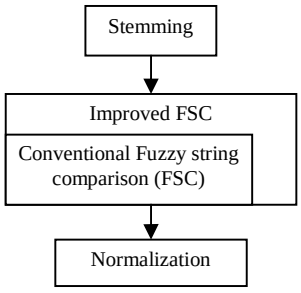


Figure 4. Similarity-based model for the ontology matching.

A. Stemming

To improve the matching quality the application of the stemming procedure is proposed. This operation makes it possible to identify ontology elements even if they are written in different forms. The following conversions can be done: "looking" → "look", "device" → "devic", "vertical" → "vertic", and "horizontal" → "horizont". This procedure is uniquely tuned for each supported language.

B. Fuzzy string comparison

The basis of the string comparison algorithm is the well-known conventional algorithm that calculates occurrence of substrings from one string in the other string.

 ISBN: 978-1-61208-100-7

1. Perform the comparison based on the above algorithm twice: FC_1 = FuzzyCompare(Element$_1$, Element$_2$) and FC_2 = FuzzyCompare(Element$_2$, Element$_1$).

2. Calculate the result as an aggregation of the above results in accordance with the following formula:

$Re'=n*FC_1+(1-n)*FC_2$, where n is a weight, $n \in [0;1]$.

$n = 0.5$ sets the same weight to the both strings, $n = 0$ searches only Request within Class, and $n = 1$ searches only Class within Request. It is proposed to set $n = 0.5$.

VI. GRAPH-BASED DISTANCE IMPROVEMENT MODEL

The graph-based improvement model for propagation similarities from one ontology element to another is presented in Figure 5. The main goal of this model is to propagate the degree of similarity between closely matching ontology elements to ontology elements related to them through RDF triples.

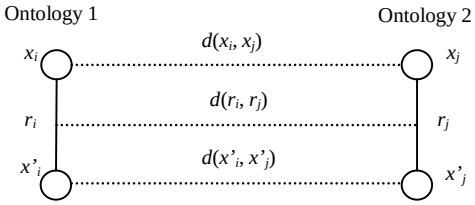


Figure 5. Matching of two ontology model.

Set $X=(x_1, x_2, ..., x_n)$ is the set of subjects and objects in the ontology of two knowledge processors. Set $D_x = (d(x_i, x_j), ...)$ is a degree of similarity between x_i and x_j. Set $R = (r_1, r_2, ..., r_n)$ is a set of predicates in the ontology of two knowledge processors. Set $D_r = (d(r_i, r_j), ...)$ is a set of degrees of similarity between r_i and r_j. Constant Tr is a threshold value that determines whether two ontology elements mapped to each other or not.

The following algorithm allows propagating similarity distance to RDF subjects and objects.

$d(x_i, x_j)$ = maximum(D_x)
while ($d(x_i, x_j) > Tr$) do
  for each $d(x'_i, x'_j)$ as $x_i\, r_m\, x'_i$ and $x_j\, r_l\, x'_j$ do
    if $d(r_m, r_l) > Tr$ then

$$d(x'_i, x'_j) = \sqrt[2]{d(x_i, x_j) * d(x'_i, x'_j)}$$

    endif
  endfor
  Exclude $d(x_i, x_j)$ from D_x
  $d(x_i, x_j)$ = maximum(D_x)
endwhile

The following algorithm allows propagating similarity distance to RDF predicates.

for each $d(x_i, x_j) > Tr$ do
  for each $d(x'_i, x'_j) > Tr$ as $x_i\, r_m\, x'_i$ and $x_j\, r_l\, x'_j$ do

$$d(r_m, r_l) = \sqrt[3]{d(x_i, x_j) * d(x'_i, x'_j) * d(r_m, r_l)}$$

  endfor
endfor

VII. SMART-ROOM CASE STUDY

This is an extended use case scenario originally proposed in one of our previous publications [6]. A meeting takes place in an "intelligent room" that is equipped with LCD projector, whiteboard, and access to Internet-based translation service.

Users that are planning to make presentation have special knowledge processor (called User Knowledge Processor or UKP) installed on Nokia MAEMO device and it implements the required functionality as described below. Upon event of user entering to the smart space meeting room, at least the following information from users' mobile devices become accessible for other smart space UKPs:

- user profile information (name, photo, domain of interests, e-mail, and phone number, etc.);
- presentation information (title, keywords, URI).

Before the meeting starts the agenda is automatically built and shown on the whiteboard including speakers' names, photos, and presentation titles. The current presentation data is highlighted on the screen. All meeting participants can see the detailed agenda on the screens of their personal mobile devices.

Users can change their user profile items. For this purposes the appropriate GUI has been implemented. When the user changes the information about his/her presentation in the profile, UKP changes the appropriate rules in the smart-space. A GUI interface for visualizing the detailed agenda on the screen of participants' MAEMO devices has been implemented as well.

At the scheduled time the appropriate presentation starts automatically, i.e., the LCD projector is switched ON and the appropriate presentation is shown. The user can control the slideshow directly from the mobile device. Five minutes before the presentation ending time and when the presentation time is about to be over the whiteboard reminds the speaker about the time restrictions.

The overall architecture of this case study is presented in Figure 6. It includes the following knowledge processors:

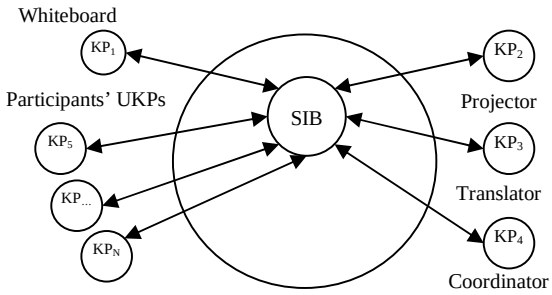


Figure 6. The case study architecture.

- KP1 – whiteboard (PC)
- KP2 – projector (PC)
- KP3 – translator (PC + Internet service)
- KP4 – coordinator (PC)
- KP5… N – UKPs (e.g., Nokia N810)

 ISBN: 978-1-61208-100-7

The application of the matching procedure is shown in Figure 7. Let us consider the following example. The user having mobile device with a knowledge processor (User KP or UKP) running on it and he/she is going to give a lecture. The corresponding ontology is presented in Figure 8.

The following ontology describes the user in respect of giving a lecture:

(“user”, “gives”, “lecture”)
(“current slide”, “part of”, “lecture”)
(“current slide”, “is”, “Number”)
(“lecture”, “has”, “Location”)

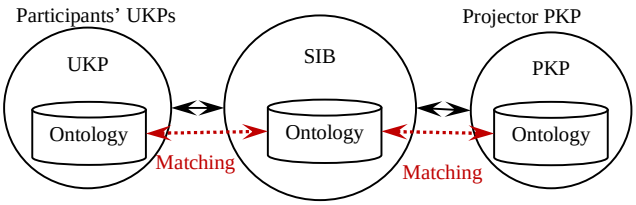


Figure 7. The extended case study architecture.

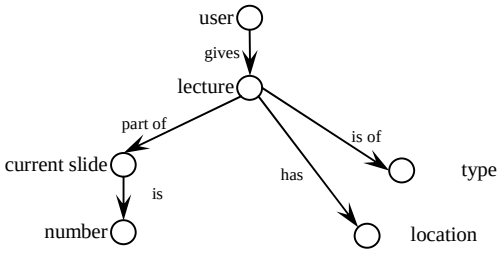


Figure 8. Ontology of user KP.

(“lecture”, “is of”, “Type”)

The LCD projector represented in the smart space by Projector KP (PKP) is capable of this function, as it is shown by ontology of the projector KP in Figure 9. This ontology currently located in the smart space. In order for the presentation to be shown, the UKP has to share the information about the presentation location (URI) with PKP.

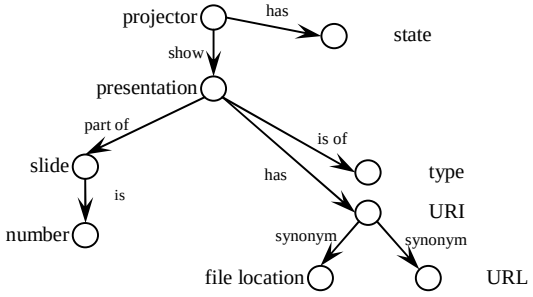


Figure 9. Ontology of projector KP.

The description of corresponding projector ontology is presented below:

(“projector”, “has”, “state”)
(“projector”, “show”, “presentation”)
(“slide”, “part of”, “presentation”)
(“slide”, “is”, “Number”)
(“presentation”, “has”, “URI”)
(“presentation”, “is of”, “Type”)
(“URL”, “synonym”, “URI”)
(“file location”, “synonym”, “URI”)

For this purposes the UKP and PKP ontology have to be merged. The element “*lecture*” from UKP is merged with the “*presentation*” from PKP as a result of the semantic-based distances search method. Distance between these elements is 0.3. The element “*location*” from UKP is merged with the “*URI*” from PKP as a result of checking the synonyms rules of the PKP ontology.

The element “*Current slide*” is merged with the element “*Slide*”. The degree of similarity between these elements estimated via the fuzzy string comparison method is 0.58.

As a result of this matching UKP knows that the following rule has to be added to the smart space to start the lecture:

(“URI”, “is”, “http://example.com/presentation1.ppt”)

VIII. CONCLUSION

The paper proposes the multi-model approach to on-the-fly ontology matching in smart spaces. The approach has been developed by integrating the most efficient techniques applicable to Smart-M3 and based on analysis of the state of art. It takes into account that the matching has to be done “on-the-fly” by mobile devices with limited capacities and uses the fact that knowledge processors are responsible for performing certain concrete and well-described tasks. The experiments showed that the matching procedure based on the proposed algorithm takes less than a second on Nokia N810 mobile device.

In this study we could not find fully automated solution for defining threshold value, but we are continuing to think in this direction and your ideas on how it can be solved are very welcome.

ACKNOWLEDGEMENT

The presented work is a result of the joint project between SPIIRAS and Nokia Research Center. Part of the implementation work has been also supported by Open Innovations Framework Program FRUCT – www.fruct.org.

REFERENCES

[1] SourceForge, Smart-M3, http://sourceforge.net/projects/ smart-m3. Retrieved: 10.5.2010.

[2] Semantic Web, www.semanticweb.org. Retrieved: 10.5.2010.

[3] Smart-M3 Wikipedia page, http://en.wikipedia.org/wiki/Smart-M3. Retrieved: 10.4.2010.

[4] Oliver, I and Honkola, J. Personal Semantic Web Through a Space Based Computing Environment. Middleware for Semantic Web 08 at ICSC'08, Santa Clara, CA, USA (2008).

[5] Oliver, I, Honkola, J and Ziegler, J. Dynamic, Localized Space Based Semantic Webs. WWW/Internet Conference, Freiburg, Germany (2008).

[6] Smirnov, A., Kashevnik, A., Shilov, N., Oliver, I., Lappetelainen, A., Boldyrev, S. Anonymous Agent Coordination in Smart Spaces: State-of-the-Art. Smart Spaces and Next Generation Wired/Wireless Networking (ruSmart 2009) conference, Springer, LNCS 5764, pp. 42-51. St-Petersburg, Russia (2009).

 ISBN: 978-1-61208-100-7

Mobile Augmented Reality System for Interacting with Ubiquitous Information

Andriamasinoro Rahajaniaina Jean-Pierre Jessel
Institut de Recherche en Informatique de Toulouse
Université de Toulouse – Paul Sabatier
Toulouse, France
{Andriamasinoro.Rahajaniaina, Jean-Pierre.Jessel}@irit.fr

***Abstract*—In this paper we describe a mobile collaborative Augmented Reality system which allows multiple users to visualize and to interact with a distributed parcel's information in a mobile device using wireless network. This information augmented the live images of the real environment surrounding the user. The main goal of our work is to use the replication of the message dispatcher server with message filter in order to allow each client allowing load balancing between them and to permit the message dispatcher server dispatch all updates messages to the concerned clients only. Our system explores a collaborative Augmented Reality application across several hardware device, and multi-user interface adaptation.**

Keywords - Ubiquitous collaborative computing; mobile collaborative augmented reality; distributed Geographical Information System.

I. INTRODUCTION

Augmented Reality (AR) is among the more popular techniques used to perform user interface for ubiquitous applications. AR presents information in its context within a 3D mixed environment.

The problem of mobile AR is the diversity of the mobile devices (visualization and interaction). In this case tools for visualizations and interaction must take account of this constrained. Most of the previous researches [1][2][3] that used Geographical Information System (GIS) database relating for building, streets in order to enhance the experiences of users (e.g., tourists, visitors). These previous works are dedicated for one or few platforms and applications. Instead, the system presented here tries to surmount this problem. We propose a collaborative augmented reality system to visualize and change the parcel's characteristic depending on location and context. As GIS database stores numerous data, so retrieving information from it is among of one critical point in a distributed AR system because it may generate latency during exchange. To overcome this problem, we show in detail our approach about retrieving information related to a Parcel from the GIS database, and the adequate metaphor for visualization and for interaction.

Outdoor AR systems have traditionally been reserved to use GIS databases related to buildings and streets in order to provide help to users.

First, the Mobile Augmented Reality Systems (MARS) project [1] allowed user to arrange the multimedia information according to chronological order. This system used a campus database to overlay labels on buildings seen through a tracked head-worn display.

The Archeoguide project [4] was designed to increase real images of user's environment with virtual story information related to them.

Next, in [5], the authors presented a prototype of an interactive visualization framework specifically designed for presenting geographical information in both indoor and outdoor environments. Participants can visualize 3D reconstructions of geographical information in real-time.

Then, ARscouting system [3] allows the mobile client (scout) takes several images for instance of a target building and transmitted them to a custom database. After that, the reconstruction engine gets a notification and triggers the reconstruction process. Once the reconstruction task is over, the server stores the virtual object and transmits it to the scout in order to increase user interface.

The claimed Mobile Augmented Reality Applications (MARA) [2] allows users to interact with their surrounding environment using the standard mobile device inputs. The users could share or exchange all data with others connected users.

This paper is organized as follows. Section 2 presents a detailed description of our *ARGisUbiq* system. After that, we give experimentation and result in Section 3. Section 4 presents conclusion and future works.

II. ARGISUBIQ SYSTEM

ARGisUbiq system is an improved version of our preview work [6]. The main goal of *ARGisUbiq* system is to propose a collaborative application AR GIS in agronomic domain that shows all information about a parcel according to the user's location.

The main difference between this work and preview ones is the use of the message distributed server for managing the collaboration between clients and to maintain the coherence of the database on all the clients and servers.

A. Message dispatcher server

This subcomponent is dedicated to dispatch all messages between the clients and the servers. It uses IceStorm service from Internet Communication Engine (ICE) [7] and applies publisher and subscriber principles. As IceStorm allows the replication of the message dispatcher via IceGrid, we create three identical ones in order to avoid that the message dispatcher become a bottleneck. In this case, the publisher

 ISBN: 978-1-61208-100-7

could make a load balancing between these dispatchers of messages. Each dispatcher has a topic manager which manages one or more topics activities. According to the client's number, it is possible adding a new message dispatcher dynamically by changing the server's configuration file.

Each topic can have several publishers and subscribers. In this scheme of communication, the slave servers and clients are the subscribers and they must subscribe on a topic in order to receive all message from this one. A topic is identified by its name in the gateway and it is the responsible of the message dispatching deposited by the publishers to all subscribers on this topic.

As all Remote Protocol Communiction (RPC), a publisher has a proxy of the publisher on the server in order to deposit its message to the topic. When receiving the message, the servant on the topic dispatches it to its subscribers.

a) Type of message: With aim of having the best exchange, we use two kinds of messages: update messages and synchronization messages.

An update message is deposited by a publisher after an user interaction which changes the state of the additional object in the augmented space. This message will be dispatched to the subscribers.

A synchronization message is distributed in order to maintain the coherence of the data on all clients and servers, the master server deposits a synchronization message on the dispatcher of message in a regular time interval. This one contains all last updates and which will distributed towards in all participants including the slave servers.

With this method of communication, the distributed messages are occupied the network's bandwidth when several publishers deposited their messages in the same time and generated a problem. As ICE does not envisage this kind of problem, we add our message filter on the distributed message server in order to distribute this one only to the concerned subscribers. We describe this technique in the following section.

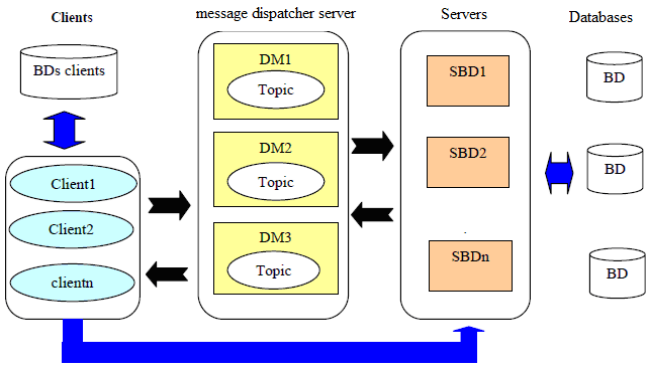


Figure 1. The general architecture of communication

b) Message filter: We have two possibilities to install this filter: at the time of the creation of the topic or in the servant. In the first case, it is enough to set up a rule or naming the topic. In other word, the topic's name will be followed by a suffix whose value is the bounding box where the creator of topic is located. Consequently, the messages will be distributed only to the subscribers which are in this bounding box. The disadvantage here is that the publisher should create a new topic with each time it change a bounding box. It can involve a waste of time since it should disconnect from the old topic and be established a connection to the new topic. This waste of time will have an impact on the fluidity of the user's interaction.

For the second possibility, we use only one topic and the value of the bounding box is deposited by the publisher with its message. After that, the servant will be distributed this one only to the subscribers which have the same value of bounding box. In this case, the publishers do not need to change topic with each time they change a bounding box. We chose this second type of filter in order to have more fluid interaction. This one is useful only for the update messages because only the users close to the initiator of change will be interested. We use message dispatcher server to manage the user's collaboration which we will see in the section below.

B. *User's collaboration*

This work allows users to collaborate. The main goal is to manage the update of the juridical situation of a parcel. Each participant could change this one only in *hybridview* mode and the flag value is equal to zero. On the other hand, the other users are not obliged in *hybridview* mode. Once the process of modification is engaged, the flag value becomes to 1 (parcel locked).

The type of collaboration depends on the location of each participant with respect to the initiator of modification one:

1) Synchronous collaboration: this mode of collaboration proceeds between initiator of modification and the other participants who are in the same bounding box. Indeed, the message of updated juridical situation is immediately sent to these participants only in order to update in real-time their data and their interfaces of visualization (see Figure 2).

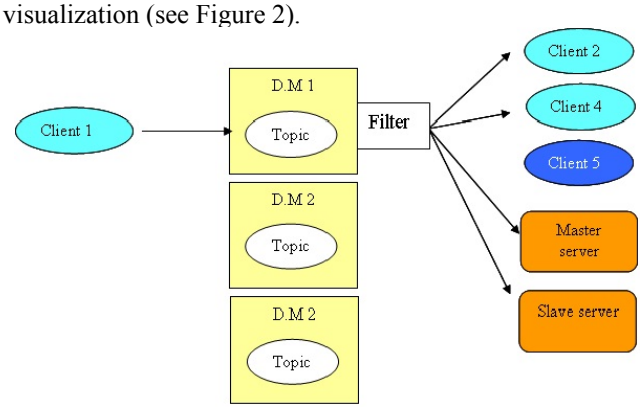


Figure 2. The update message dispatching

2) Asynchronous collaboration: this case arises between initiator of the modification and the others participants who are located in other bounding box. In this case, modifications are distributed asynchronously to these participants in order to update their databases. This principle is adopted because these ones are not directly concerned with the modification. During the modification, the other users cannot make modification.

C. Selection Information source

Like others outdoor AR systems, *ARGisUbiq* system uses GIS data as data source. As parcel's information related to agronomic and type of plant is unavailable on producer's map, we create our own database inspired from parcel database (using a vector format formed by shape files, index one and dbf one).

Each polygon in the file shape is delimited by a bounding box. We exploited these limit in order to know the parcel in which the operator is located. Thus, we tested two methods: first is based on the calculation of the barycenter and the second is called Winding number.

1) The barycenter method: This method is used especially when there are several parcels inside the bounding box. We calculate the barycenter of each parcel. We take account only the parcel having its barycenter nearby the user's position. Once the found parcel, one will recover other information by using the index of the parcel. The problem with this method, it is that the operator can be inside the bounding box but it is not always inside a parcel because the fact of having the smallest distance compared to the barycenter does not signify that it is inside. To solve this problem, we use the Winding number method which we will see in the section below.

2) The winding number method: As the parcels stored in the shp file format (shape file) are convex polygons, then we can apply the Winding Number method [8].

This method makes possible to know the inclusion of a point in a convex polygon. The algorithm is summarized as follows: we take an infinity ray R starting at a point P and we calculate the number of intersection where R crosses the segments which form the polygon. If R crosses a segment of the polygon whose has the clockwise orientation, the counter value is equal to -1. On the contrary, it is equal to +1.

We do not take account of the segments parallel with R. If the sum of these numbers is equal zero that means that the point P is outside the polygon. In contrast, it is inside.

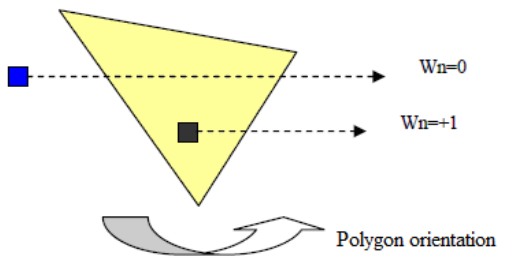


Figure 3. The application of Winding number method

By replacing the point P by the position of the operator and the polygon by one of the parcels in the database, we applied this algorithm after having tested the inclusion of the position of the operator in the bounding box. This test of inclusion in the bounding box enables us to avoid the useless execution of the algorithm for the points outside the bounding box. When we found the parcel, we will recover other information by using *the index of the parcel.*

D. Visualization metaphor

To visualize the GIS information selected as described in the preview section, we proposed two metaphors of visualization: the metaphor of visualization in *textview* mode and the metaphor of visualization in *hybridview* mode. We added audio augmentation to these metaphors not to obstruct the sight of the operator on the one hand and on the other hand in order to provide the details of information by overcoming the limits imposed by the size of the screen.

In this case, additional information essential to the sight is used for the visual augmentation and the others are dedicated to the audio augmentation.

The combination of these various modes of augmentation enables us to make a multimodal augmentation instead of limiting to the usual visual one. This one offers us opportunity of providing an identical metaphor of visualization for the various types of used devices.

In the *textview* mode, the scene is augmented by virtual text and aural information related to a parcel and the position of user. In the *hybridview* metaphor, user interface is enhanced by virtual text, parcel map and audio information.

The blue point on the map is the user's position. We combine landscape and portrait mode with *textview* and *hybridview* when users use lightweight hardware as visualization tools.

The transition between the two metaphors depends on the way which the user holds his PDA. The *hybridview* is the default visualization metaphor for UMPC and Personal Computer (PC) notebook, and the *textview* metaphor is for PDA. Each juridical situation is associated with a color of which green for *public property*, yellow for *public property (request in progress)* and red for the juridical situation *private property*.

Figure 4. The textview mode on an UMPC

We decide to use classical interaction tools like menu, stylus, and button because these are available on each device that we use as visualization tools. When the user selects one of both menus using his stylus, the user interface changes according to the menu item selected. After that, the menu item changes to another one.

E. Interaction metaphor

We propose the possibility for user to choose visualization metaphor using *textview* and *hybridview* menus. As described above, when using a Personal Digital Assistance (PDA), the transition between the two metaphors depends on how the user holds his PDA and the value of pitch angle (Θ) from the inertial sensor Xsens MTi (Θ value between -5.0° and 0.0° for portrait mode and landscape mode for others values) see Figure 5.

When the current mode of visualization is the *hybridview* mode, the operator can modify the juridical situation of the parcel visited while clicking on the map. According to the situation of the parcel, it can pass from the juridical situation *public property* to the situation *private property* while passing by the intermediate situation *public property (request in progress)*. Only the parcel in *public property* or *public property (request in progress)* situation accepts the modifications of juridical situation (see Figure 6).

To manage the user's collaboration, we attribute for each parcel a flag which designate its modification state. An operator can change the juridical situation of a parcel only if its flag value is equal zero. When a parcel is in an intermediate phase (an operator interacts with it to change its state) the flag value becomes equal to 1.

III. EXPERIMENTATION AND RESULTS

This first prototype was tested with three platforms: the first client has used Q1 Samsung with 800 Mhz Celeron M ULV processor, 256 Mo RAM and the two other clients have used a Pocket PC Dell Axim x51v with 624Mhz Intel Xscale processor, 64Mo RAM. We have used a database of common formed by 180,000 parcels and each parcel is formed by 10 up to 20 vertices. We have tested two different scenarios: first, we have used a local database: as we have loaded the database in the memory at the first time, the Q1 client has run after 5 s of the database loading and 22 s for the two PDA clients. They do not need to establish a connection to the database server but they must to connect to the dispatcher of the message server. This last one is formed by three dispatcher of message running on a Linux Ubuntu operating system. This connection has spent 0.3 s. After this step, the Q1 client was able to achieve 25-30 fps (frame per second) and 17-20 fps for the PDA clients during the exchange with the data in memory.

In the second test, the database is duplicated on three replica servers: one master registry and two slave ones. Each slave registry has had its own node which monitors two applications servers. The locate request from the client to the registry has spent 0.3 s for Q1 client and 0.7 s for PDA clients. After that, the Q1 client was able to 23-28 fps and 15–20 fps for the PDA clients during the exchange with the replica server. As describe above, all update message must deposit on the dispatcher of the message server.

As we saw, the difference between the results using the local database and replicate database was small. It's not surprising because we have added to the client a functionality to reduce the number of exchange with the replica server. It's also due to the performance of our replica servers: 50% of users only are satisfied for hardware device ergonomic because most of users prefer using wireless inertial sensor instead of using MTi.

As we describe above, our work could overcome the network problem because we are working in memory to reduce the exchange with the database servers. Consequently, our system permits the users to collaborate in real-time. Moreover, it does not consume more memory time.

From the experimentation in the long run, the pitch values have presented any drift but it has not impact to our system because this one does not need exact pitch values for running. In other hand, the pitch values must be in the interval as we describe in [6].

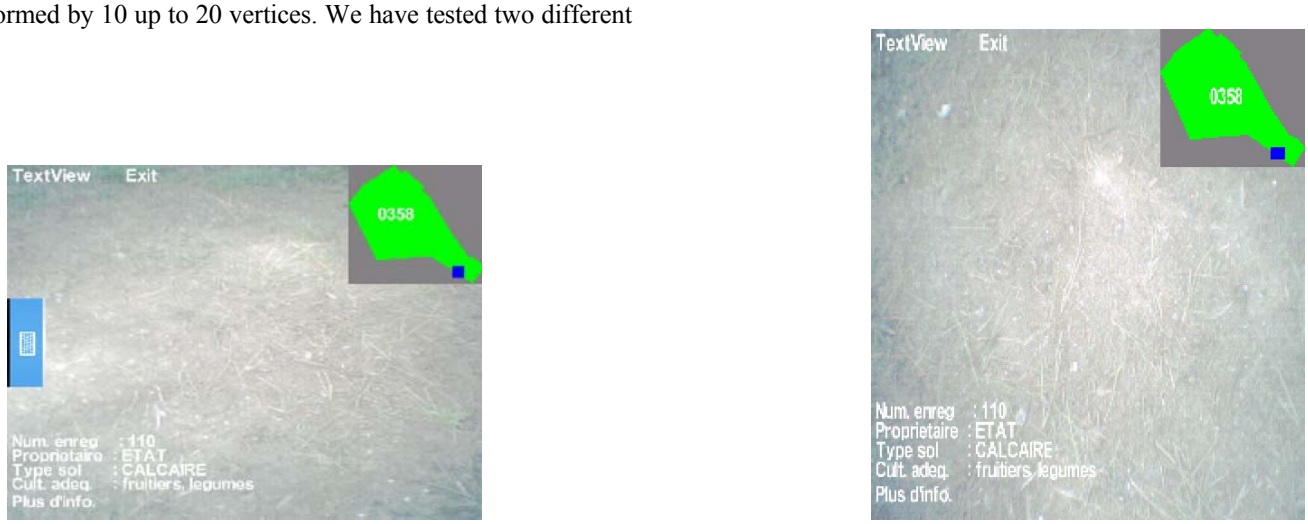


Figure 5. The landscape hybridview mode (left) and the portrait hybridview mode (right)

 ISBN: 978-1-61208-100-7

Figure 6. The portrait hybridview mode before change (left) and after change (right)

From these results we can deduce design guidelines to choose hardware device in future AR applications.

IV. CONCLUSION AND FUTURE WORK

In this paper, we addressed the problem of enhancing user's contextual perception of the real word using GIS data on several hardware and software platform. To tackle this, we have proposed the *ARGISUbiq* multiplatform architecture which exploits Mobile collaborative Augmented Reality principles to improve user's interaction with GIS data. As we use specifically built GIS data, we described our database's structure and how to select appropriate information related to user's position. Our distributed application is based on Internet Communication Engine, an object-oriented middleware, used to ensure the connection between database servers and clients. To avoid eventual problem with database server, we duplicate our database on several servers and we use IceGrid services to provide load balancing between all servers. Some clients are able to access concurrently to a selected server.

We use IceStorm services on the message dispatcher server to manage the dispatching of the update and synchronization message to the subscribers and the slave servers.

We are entirely satisfied with our results. In the future work, instead using MTi sensor we plan to use low cost or embedded inertial sensor and computer vision based techniques to compute the user's orientation.

REFERENCES

[1] T. Höllerer, S. Feiner, T. Terauchi, G. Rashid, and D. Hallaway, "Exploring MARS: Developing Indoor and Outdoor User Interfaces to a Mobile Augmented Reality System", Computers and Graphics, 23(6), Elsevier Publishers, 1999, pp. 779-785.

[2] M. Kähäri and D. J. Murphy, "MARA-Sensor based Augmented Reality System for Mobile Imaging Device", http://research.nokia.com/projects/MARA/, 18 July 2010.

[3] B. Reitinger, C. Zach, and D. Schmalstieg, "Augmented RealityScouting for interactive 3D reconstruction," unpublished.

[4] T. Gleue and P. Daehne, Augmented Reality-based Cultural Heritage On-site Guide, http://archeoguide.intranet.gr, 18 July 2010.

[5] F. Liarokapis, I. Greatbatch, D. Mountain, A. Gunesh, V. Brujic-Okretic, and J. Raper, "Mobile Augmented Reality techniques for GeoVisualisation," Proc. 9th International Conference on Information Visualisation, IEEE Computer Society, London, 2005, pp. 745-751.

[6] A. Rahajaniaina and J. Jessel, "Visualization of Distributed Parcel's Information on Mobile Device," Proc. International Conference on Advanced Geographic Information Systems, Applications, and Services (GEOProcessing 2010), St. Maarten, Netherlands Antilles, IEEE Computer Society, February 2010, pp. 27-32.

[7] M. Henning, M. Spruiell, D. Boone, B. Eagles, B. Foucher, M. Laukien, M. Newhook, and B. Normier, Distributed Programming with Ice, ZeroC: http://www.zeroc.com/Ice-Manual.pdf, 18 July 2010.

[8] D. Sunday, Fast Winding Number Test for Point Inclusion in a Polygon. http://softsurfer.com/algorithms.htm, 18 July 2010.

 ISBN: 978-1-61208-100-7

Performance Analysis of Receive Collaboration in TDMA-based Wireless Sensor Networks

Behnam Banitalebi, Stephan Sigg, Michael Beigl
TecO, Karlsruhe Institute of Technology (KIT)
Karlsruhe, Germany
{behnam, sigg, michael}@teco.edu

Abstract— **This paper presents a general framework to enable the implementation of receive collaboration in wireless sensor networks (WSN). The framework also allows the evaluation of collaborative channel equalization (CCE) performance as one aspect of receive collaboration. Our analysis shows that the use of receive collaboration to receive signals from remote sources is energy efficient with reasonable computational load and memory consumption. However, the increased time delay is a restricting parameter, which limits the application of receive collaboration to a small number of cooperating nodes and to rather short data streams.**

Keywords-*receive collaboration; wireless sensor network; channel equalization; TDMA*

I. INTRODUCTION

Due to the limited and nonrenewable batteries of the sensor nodes in WSNs, energy efficiency is one of the most important issues of the network and protocol design in WSNs.

The application of array processing schemes [1-4] in wireless communication applications is beneficial in terms of energy efficiency and reliability. In WSNs, array processing schemes are applicable during a cooperation of a group of neighbouring nodes. Obviously, synchronization, sharing the signals via local communications and processing capabilities to process the transmitted/received signals are some necessities of the sensor nodes.

Due to the limitations of WSNs, array processing schemes are applicable by cooperation of the sensor nodes which increases the amount of inter-node communications. In [5] and [6], a general framework for implementation of array processing schemes is developed. According to this idea, a group of sensors cooperate to improve signal reception. Collaborative channel equalization is also suggested as an aspect of receive collaboration. The use of array processing schemes in transmit mode is considered as distributed beamforming in [7-9]. It is shown that both receive and transmit collaboration are effective methods to decrease and distribute the energy consumption.

Despite of its energy efficiency, the increased computational load and memory consumption of receive collaboration limit the applications. However, increase in its applications is expected due to future advances in hardware design.

In [5], receive collaboration is introduced as a way to increase the energy efficiency of WSNs in receive mode. According to this new idea, a group of neighboring nodes cooperate to improve the reception of a signal by applying array processing schemes like channel equalization. Based on this idea, it is applicable to improve the quality of the received signal at a fixed value of transmitted SNR or to reduce the transmitted power without any decrease in the quality of received signal. The implementation of receive collaboration and CCE in CDMA based WSNs are introduced in [6]. It is shown that due to the large amount of spreading and despreading during CDMA based local communications, the computational load and memory requirements of receive collaboration increase more than linearly. Therefore, the processing capability and available memory of the nodes are the two restricting parameters to develop receive collaboration in CDMA based networks.

In TDMA-based protocols, a TDMA frame is divided into time slots and each node is assigned one. The transmission schedule allows nodes to send and receive without collision. In TDMA based MAC protocols the interference between adjacent wireless links is guaranteed to be avoided. Thus, the energy waste coming from packet collisions is diminished. In [10-12] some TDMA based MAC protocols for sensor networks are studied. Although the use of TDMA in local communications requires exact synchronization of cooperating nodes, it is an efficient way of mitigating the limitations of CDMA based networks.

The objective of this work is to develop and evaluate receive collaboration in TDMA based WSNs. To do so, in the next section, we propose a general framework for implementation of receive collaboration in TDMA based WSNs. This framework is a primary step to implement the array processing schemes. To evaluate the performance, advantages and disadvantages of receive collaboration, CCE is considered in Section III as an aspect of receive collaboration. According to CCE, after aggregation of the received signals at the processing node, a channel equalization scheme is applied on the signals to decrease undesired effects of the transmission channel. In this section, energy efficiency, computational load, time delay and memory requirements of the proposed framework are evaluated. It also includes some simulations. Finally, Section IV concludes this paper.

 ISBN: 978-1-61208-100-7

II. RECEIVE COLLABORATION FOR TDMA BASED WSNS

In the following subsections, a general framework for implementation of receive collaboration in TDMA based WSNs is developed.

A. Data Reception

The first step of receive collaboration is the reception of an impinging signal from the remote node. In TDMA based networks, various methods may be considered to manage the reception and transmission timing of the nodes.

In the following steps, it is assumed that the cooperating nodes transmit their signals only in their corresponding time slot whereas they are ready to receive the impinging signals both in their time slots and in the common time slot.

B. Announcement

After reception of the impinging signal, one of the cooperating nodes is selected as the processing mode to handle the receive collaboration. New processing node is selected by a reference node which can be the previous processing node or cluster-head in the cluster based networks. An announcement message containing the ID of the new processing node is broadcasted through the other cooperating nodes in the common time slot.

C. Synchronization

Random distribution of the cooperating nodes causes random time delays in local communications. Therefore, in advance of the aggregation of the received signals, cooperating nodes should be synchronized. Similar to the CDMA based networks [5], during the synchronization step, the processing node estimates the time delays of local communications and assigns new time slots to the cooperating nodes to increase the efficiency of the receive collaboration. Proper time delays should be applied to the new time scheduling to avoid overlapping.

To do so, processing node broadcasts a synchronization message via the common time slot. It can both detect the time slots of the cooperating nodes and estimate the time delays of local communications based on the feedbacks from cooperating nodes. To properly estimate the time delays, the synchronization message should be short enough to avoid overlapping due to different time delays. Finally, the processing node computes a new scheduling for the cooperating nodes and informs them via their previous time slots. Reception of the acknowledgements from the cooperating nodes confirms proper time slots allocation.

Various time slot scheduling methods can be suggested to avoid overlapping due to different time delays of local communications. As the simplest method, some guard bands are considered among the time slots. Assuming the length of the guard bands to be equal to the maximum time delay of local communications overlap-free local communication is guaranteed. Despite of its simplicity, this method is not time-efficient. In another method, non-processing nodes are sorted according to their time delays and the node with smaller time delay would belong to the first time slot and so on. Although this method increases the efficiency of time scheduling, it is not completely efficient due to gaps among the received time slots. As an efficient method, it is possible to designate the time slots such that they have no overlap at the processing node, meanwhile there is no gap among the received time slots by the processing node. The length of the time slots depends on the method used in the aggregation step. It is discussed in more detail in the next subsection.

D. Aggregation

In this step, received signals by the cooperating nodes are aggregated at the processing node. Proper array processing scheme is applied mostly by combination of these signals after applying weighting coefficients which are generated recursively based on the aggregated signals. In this paper, the least squares constant modulus algorithm (LS-CMA) [13-14] is considered as the channel equalization scheme. LS-CMA estimates optimum weighting coefficients by minimizing the following cost function with respect to w_k, the $1\times M$ vector of weighting coefficients, which corresponds to k-th sample of the aggregated signal

$$J(\mathbf{w}_k) = E\left[\left(|y_k|^2 - 1\right)^2\right] \tag{1}$$

here, $E[.]$ denotes the expected value and y_k, the channel equalizer output is in the form

$$y_k = \mathbf{w}_k \cdot \mathbf{x}_k^H \tag{2}$$

where x_k is the $1\times M$ vector containing the k-th sample of received signals. The operator H is the conjugate transpose operator.

According to the stochastic gradient methods [15], the weight vector in each time instant is updated based on its previous value and the gradient of the cost function. In practice, w_k is updated recursively as follows

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \mu \cdot \mathbf{x}_k \cdot \left[|y_k|^2 - 1\right] \cdot y_k \tag{3}$$

In this equation, μ is the step size of the algorithm, a constant parameter to control the convergence rate of the algorithm. After convergence of the algorithm to its optimum weighting coefficient, w_{opt}, it has small variations due to the variation of the effective parameters on the aggregated signals. It enables the processing node to apply w_{opt} for some parts of the aggregated signals.

To start the aggregation step, processing node broadcasts a request for aggregation. After receiving the aggregation request, each non-processing node sends its signal to the processing node via its corresponding time slot. Assuming to be enough samples at the first part of aggregated signals for convergence, w_{opt} is achieved and used to generate the output of the channel equalizer.

Depending on the criticality of the power supply and computational capability of the processing node and also energy consumption of the local communications, it is possible to distribute the computational load. To do so, processing node sends the coefficients to the corresponding nodes. Each node sends its signal to the processing node after applying the weighting coefficient.

Due to the simple and low cost construction of sensor nodes, it is not applicable to implement exact synchroniza-

 ISBN: 978-1-61208-100-7

tion modules for the cooperating nodes. Therefore, especially for higher data rates, there are some synchronization errors during local communications.

The question is that how much is the ability of CCE in the case of imperfect synchronization. Despite the performance degradation of CMA based channel equalizer due to random distribution of the cooperating nodes, it is shown that CMA is still helpful for such applications [5]. Since channel equalization is not based the nodes' positions, it is possible to model the synchronization errors as some changes in the nodes' position. However the synchronization error should be small enough to have correlated signals at the processing node in data aggregation step.

III. PERFORMANCE EVALUATION

CCE is an aspect of receive collaboration in which the processing node applies a channel equalization method to decrease undesired transmission channel effects. In this section, the effect of CCE on some critical parameters of WSN such as energy efficiency, computational load, time delay and memory consumption are considered.

To better visualization of the analysis of this section, some simulations are also presented. The assumptions of this section are listed below:

- Duration of the transmitted signal: 1 *ms*
- Number of the cooperating nodes: 50
- The distance between remote and cooperating nodes: 2 *km*
- Primary time slot duration in local communications: 100 μs
- The number of primary time slots in the primary scheduling: 100
- New time slot duration: 50 μs
- Number of the time slots in the new scheduling: 50
- Radius of the disk containing the nodes: 50 *m*
- Node density: $6.4 \cdot 10^{-6}$ *nodes*/m^2
- Carrier frequency of the remote node: 20 *MHz*
- BER of interest at the processing node: 0.01

These parameters are constant unless it is mentioned. To focus on CCE, we avoid using any coding technique in our simulations.

A. Energy Efficiency

Energy efficiency is the relative improvement of using CCE reception method which is defined as follows

$$e = \frac{E_{noCCE} - E_{CCE}}{E_{noCCE}} \tag{4}$$

where E_{noCCE} and E_{CCE} are energy consumption of non-CCE and CCE based reception method.

In the case of non-CCE based reception method, one of the cooperating nodes receives the impinging signal from the remote node. According to the Friis equation, the transmission loss is

$$L_{LR} = 20\log(L) + 20\log(f_d) + 32.44 \tag{5}$$

where L (*km*) is the distance between remote and cooperating nodes and f_d (*MHz*) is the carrier frequency of the remote node. Without the use of CCE, to achieve a fixed BER at the receivers, received SNR should be higher than a threshold SNR_{noCCE} which is estimated in [5]. Therefore, the transmitted power is

$$P_{noCCE} \geq L_{LR} + SNR_{noCCE} + N \tag{6}$$

where, N is the noise power at the receiver. All parameters in equation (6) are in dB. If the duration of the transmitted signal is T_d, the total energy consumption in the case of using no CCE is

$$E_{noCCE} = P_{noCCE} \cdot T_d \tag{7}$$

In the case of using CCE, to meet the BER of interest, the SNR at the receiver should be equal to or greater than SNR_{CCE}. Similar to the discussion above, the energy consumption at the transmitter is equal to

$$E_{rec} = P_{rec} \cdot T_d \tag{8}$$

where

$$P_{rec} \geq L_{LR} + SNR_{CCE} + N \tag{9}$$

In the second step, a reference node selects the new processing node and introduces it to the other cooperating nodes by broadcasting an announcement message. The energy consumption of this step is

$$E_{ann} = P_m \cdot T_m \tag{10}$$

where T_m is the duration of the managing messages like announcement or synchronization. For simplicity, we assume the same length for these messages. P_m is also the required transmission power for local communications. Assuming the required SNR at the receivers during local communications to be SNR_{SR}, we have

$$P_m \geq L_{SR} + SNR_{SR} + N \tag{11}$$

In (11), all of the parameters are in dB. L_{SR} is the maximum transmission loss in local communications which corresponds to the maximum inter-node distance in the virtual array. It is calculated similar to (5).

During the synchronization step, the processing node broadcasts a synchronization message and receives some feedbacks from the cooperating nodes. It contains M local transmissions. After estimation of the time delays of the local communications, the processing node generates new more time-efficient scheduling and informs the other cooperating nodes during M-1 local transactions. Reception of some acknowledgements from the cooperating nodes needs also M-1 local communications. Therefore, the synchronization step is performed during 3M-2 local communications and its energy consumption is

$$E_{syn} = (3M - 2) \cdot P_m \cdot T_m \tag{12}$$

 ISBN: 978-1-61208-100-7

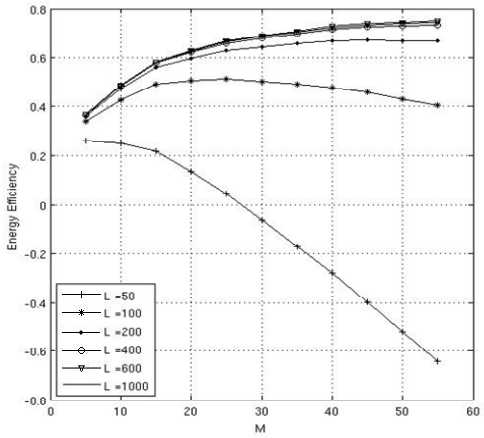


Figure 1. Effect of increasing the number of cooperating nodes on the energy efficiency for different values of L

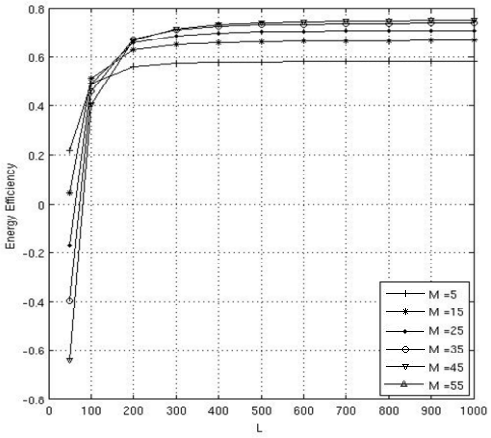


Figure 2. Effect of increasing the distance between remote and cooperating nodes on energy efficiency for different values of M

At the aggregation step, all of the non-processing nodes send their data to the processing node. Depending on the length of data stream and the time slot intervals, this step is performed in several time slots. The energy consumption of this step is

$$E_{agg} = (M-1)\cdot P_m \cdot T_d \quad (13)$$

Therefore, the energy consumption of CCE based reception method is

$$E_{CCE} = [(3M-1)\cdot T_m + (M-1)\cdot T_d]\cdot P_m + T_d \cdot P_{rec} \quad (14)$$

By substitution of (14) in (4) and some simplifications, energy efficiency is achieved as (15).

Fig. 1 illustrates the effect of increasing the number of cooperating nodes on energy efficiency. In all situations, the area in which the cooperating nodes are distributed is the same. Therefore, increasing the number of nodes is the same as increasing the node density. In this figure, energy efficiency is calculated for different values of L (the distance between remote and cooperating nodes). As seen in this figure, when $L = R$, increasing of M has negative effect on the energy efficiency such that CCE is inefficient for $M \geq 30$. Although CCE decreases the transmitted power by the remote node, the energy consumption of local communications grows up by increasing the number of cooperating nodes. In CCE, energy consumption can be divided into two parts; energy consumption by the remote node (E_{RN}) and that of local communications (E_{LC}). For small values of L, due to the small values of transmission loss, both E_{RN} and E_{noCCE} are rather small. Therefore E_{LC} plays key role in energy efficiency.

Moreover, Fig. 1 shows that by increasing of L (in the case of proper transmission range) the energy efficiency improves. It is because of the higher increment rate of E_{RN} rather than E_{LC}. Finally E_{LC} can be neglected for higher values of L and therefore the curves are saturated.

Increasing of L has no effect on the computational load and memory consumption and its effect on the time delay is neglectable. Therefore, it can be said that the application of CCE in long range communications is more beneficial than that of in short distances.

Fig. 2 shows the effect of increasing the distance of the remote node on the energy efficiency for different values of the number of cooperating nodes. All of the curves saturate by increasing of L, but increasing of M increased both the distance in which saturation happens and the final value of the energy efficiency. As mentioned before, by increasing of L, E_{RN} increases whereas E_{LC} remain constant such that after some increase in L, E_{LC} become neglectable and the curves approach to their final value. Increasing of M increases E_{LC}. Therefore, saturation happens in larger values of L. On the other hand, for small values of L, $E_{LC} > E_{RN}$. Therefore energy efficiency descends significantly. Since increases by M, using less cooperating nodes yields better results.

B. Time Delay

In the case of using no CCE, one of the cooperating nodes receives the impinging signal from the remote node. Therefore, the needed time for no CCE based reception method is $T_{noCCE} = T_d + L/C$ where T_d is the length of transmitting data sequence by the remote node, L is the distance between receiver node and the remote node and C is the free space wave propagation speed. In CCE-based reception method, the time delay of the first step is equal to T_{noCCE}. Neglecting the time delay of selecting the processing node, needed time for announcement is $T_{ann} = T_m + d_{max}/C$, where d_{max} is the maximum inter-node distance among the cooperating nodes.

The transmission of the announcement message is postponed until the next common time slot which at the worst case, it causes a time delay of $M'T_S'$, where M' and

$$e = \frac{(SNR_{noCCE} - SNR_{CCE})\cdot T_d - [(3M-1)\cdot T_m + (M-1)\cdot T_d]\cdot P_m}{20\log L + 20\log f_d - 32.44} \quad (15)$$

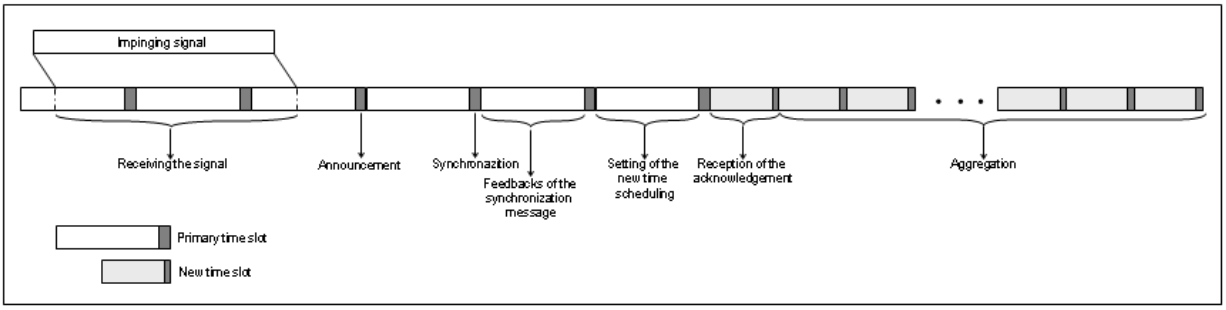


Figure 3. Time delay of the receive collaboration in TDMA based WSNs

T_S' are the number and duration of the time slots of each frame in the primary scheduling.

At the synchronization step, the processing node broadcasts the synchronization message in the common time slot and receives the feedbacks from the cooperating nodes in their corresponding time slots. The processing node sends new time scheduling information to the cooperating nodes via their primary time slots. Finally, the processing node receives some acknowledgements from the cooperating nodes. Therefore the total time delay of the synchronization step is $T_{syn} = 3M'T_S' + MT_S$, where M and T_S are the number and duration of the time slots in the new time scheduling. Finally, at the aggregation step, all of the nodes send their data to the processing nodes, which is performed in $M \cdot T_d$.

Fig. 3 represents the consuming time of different steps of receive collaboration. According to this figure, the time delay of CCE is equal to

$$T = \frac{L}{C} + 4M' \cdot T_S' + (M+1) \cdot T_d + M \cdot T_S \qquad (16)$$

In equation (16), the time delays of local communications are neglected.

Fig. 4 illustrates the time delay of CCE. It is shown that the time delay increases linearly by increasing the duration of the received signal from the remote node. This increase is mostly because of the serial aggregation of the signals in the aggregation step due to the use of TDMA in local communications. It is why the increment rate of the time delay curves increase by increasing of the number of cooperating nodes.

C. Memory Consumption

Fig. 5 shows the memory consumption of receive collaboration. As expected from the Table 1, the curves are linear. It shows that increasing of M has approximately linear effect on the memory consumption. Generally, it can be said that memory consumption is not a critical issue in TDMA based WSNs.

According to receive collaboration steps, memory consumption of the non-processing nodes is low. The memory consumption of the processing node depends on the length of the signal transmitted by the remote node.

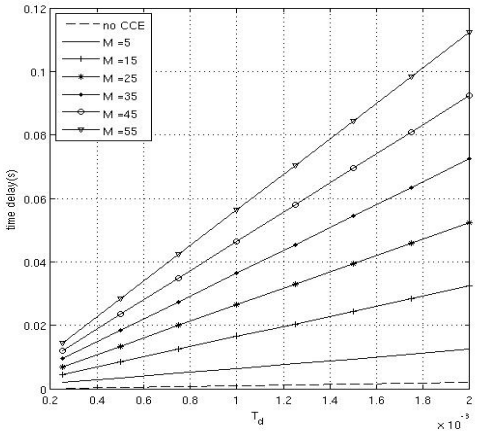


Figure 4. Effect of increasing the duration of the received signal on the time delay

D. Computational Load

In TDMA based networks computational load is not a critical issue. Assuming no encoding at the cooperating nodes, the computational load is limited to the generation of proper scheduling for the cooperating nodes (computational load of the selection of new processing node and estimation of the time delay of local communications are neglected). Therefore, the computational load of receive collaboration is:

$$O_{RC}^{A} = (M-1) \cdot L_d \qquad (17)$$

and

$$O_{RC}^{M} = M \cdot L_d \qquad (18)$$

where O_{CCE}^{A} and O_{CCE}^{M} are the number of additions and multiplications and L_d is the number of data symbols. The computational load of the array processing scheme should be considered. Assuming the number of iterations for convergence of LS-CMA is L_{con} and after convergence the weighting coefficients are valid for all of received signal. According to (2) and (3), the number of additions and multiplications of LS-CMA is

$$O_{LS-CMA}^{A} = M \cdot L_{con} + (M-1) \cdot L_d \qquad (19)$$

and

$$O_{LS-CMA}^{M} = (M+3) \cdot L_{con} + M \cdot L_d \qquad (20)$$

 ISBN: 978-1-61208-100-7

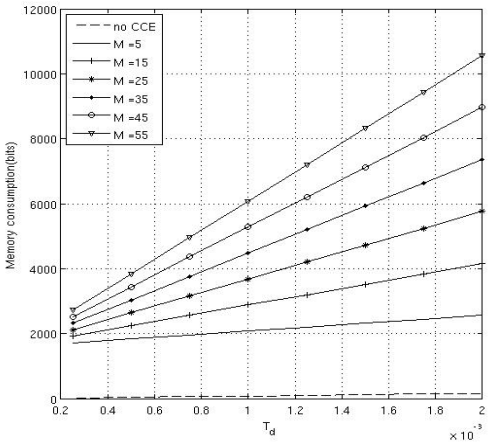


Figure 5. Effect of increasing the duration of the received signal on memory consumption

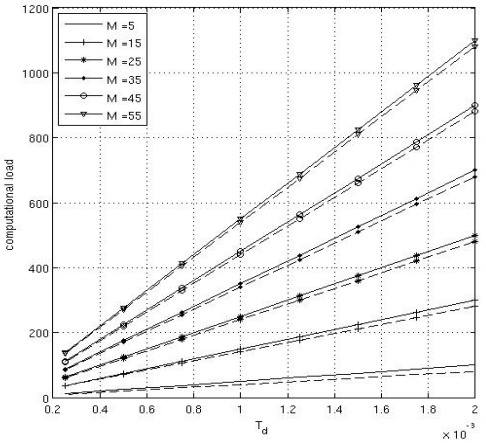


Figure 6. Effects of increasing L on the computational load for different values of M

Our simulations show that increasing the length of received signal (T_d) has no considerable effect on the energy efficiency. But it increases the computational load. Fig. 6 shows that the number of additions and multiplications are approximately the same. These parameters increase linearly by increasing the duration of the received signal. The highest value of this parameter corresponds to the longest signal duration and largest number of cooperating nodes which is less than $3 \cdot 10^5$ addition and multiplication.

IV. CONCLUSION

The performance of receive collaboration and collaborative channel equalization in TDMA based WSNs are investigated. It is shown that when the transmitter node is at the short distance (smaller than the virtual array radius) from the virtual array, the energy efficiency of CCE is negative, but it grows up by increasing the distance until 10 times of the virtual array radius. Although the computational load and memory consumption of CCE is very low, increasing of the time delay by increasing the number of cooperating nodes or by increasing the length of received signal is a limiting parameter of CCE.

In TDMA based WSNs, receive collaboration is completely useful to receive short data streams. Reasonability of receive collaboration to receive rather long data streams is due to the acceptable time delay at the cooperating nodes. Moreover, the efficiency of receive collaboration improves by increasing the distance between cooperating nodes and the remote node.

ACKNOWLEDGMENT

The authors would like to acknowledge partial funding by the European Commission for the ICT project CHOSeN ”Cooperative Hybrid Objects Sensor Networks” (Project number 224327, FP7-ICT-2007-2) within the 7th Framework Program. We would further like to acknowledge partial funding by the ’Deutsche Forschungsgemeinschaft’ (DFG) for the project ”Emergent radio” as part of the priority program 1183 ”ORGANIC COMPUTING”.

REFERENCES

[1] H. L. Van Trees, Detection, Estimation and Modulation theory, Part IV, Optimum array processing, John Wiley & Sons, 2002.

[2] O. Gay-Bellile, X. Merchal, G. Burns, and K. Vaidyanathan, “A reconfigureable superimposed 2D-mesh array for channel equalization,” IEEE International symposium on Circuits and systems, Aug. 2002.

[3] K. L. Bell, Y. Ephraim, and H. L. Van Trees “A bayesian approach to robust adaptive beamforming,” IEEE transaction on signal processing, vol. 48, no. 2, Feb. 2000.

[4] K. Yao, R. E. Hudson, C. W. Reed, D. Chen, and F. Lorenzelli, “Blind beamforming on randomly distributed sensor array system,” IEEE Journal on Selected reas in communications, vol. 16, no. 8, Oct. 1998.

[5] B. Banitalebi, S. Sigg, and M. Beigl, “On the feasibility of receive collaboration in wireless sensor networks,” accepted in PIMRC, Sep. 2010.

[6] B. Banitalebi, S. Sigg, D. Gordon, and M. Beigl “Analyzing the energy efficiency and computational load of collaborative channel equalization in wireless sensor networks,” is submitted to SENSYS, Nov. 2010.

[7] R. Mudumbai, G. Barriac, and U. Madhow, “On the feasibility of distributed beamforming in wireless network,” IEEE Transactions on Wireless Communications, vol. 6, no. 5, May 2007.

[8] H. Ochiai, P. Mitran, H. Vincent Poor, and V. Tarokh, “Collaborative beamforming for distributed wireless Ad Hoc sensor networks,” IEEE Transaction on Signal Proceeding, vol. 53, no. 11, Nov. 2005.

[9] R. Mudumbai, D. Richard Brown III, U. Madhow, and H. Vincent Poor, “Distributed Transmit Beamforming: Challenges and Recent Progress,” IEEE Communication Magazine, Feb. 2009.

[10] W.B. Heinzelman, A.P. Chandrakasan, and H. Balakrishnan, "An application-specific protocol architecture for wireless microsensor networks," IEEE Trans. on Wireless Communications, vol. 1, no. 4, Oct 2002.

[11] G Pei and C. Chien, "Low power TDMA in large wireless sensor LEACH networks," MILCOM2001, vol. 1, Oct 2001.

[12] Z. Chen and A. Khokhar, "Self organization and energy efficient TDMA MAC protocol by wake up for wireless sensor networks," IEEE SECON2004, Oct 2004.

[13] D. N. Godard, “Self recovering equalization and carrier tracking in two dimensional data communication systems,” IEEE Transactions on Communications, vol. COM-28, 1980.

[14] B. G. Agee, “The least squares CMA: a new technique for rapid correction of constant modulus signals,” ICASSP 1986.

[15] S. Haykin, *Adaptive filter theory*, 4th Edition, Prentice Hall, 2002.

 ISBN: 978-1-61208-100-7

Anonymous Agents Coordination in Smart Spaces

S. Balandin, I. Oliver, S. Boldyrev
Ubiquities Architectures team, Nokia Research Center
Itämerenkatu 11-13, 00180, Helsinki, Finland
{Sergey.Balandin, Ian.Oliver,
Sergey.Boldyrev}@nokia.com

A. Smirnov, A. Kashevnik, N. Shilov
Computer-Aided Integrated Systems laboratory, SPIIRAS
14-th Liniya 39, 199178, St.-Petersburg, Russia
{smir, alexey, nick}@iias.spb.su

***Abstract* – Rapid developments of communication, data processing and storage technologies, and continuing proliferation of consumer devices that surround user have created an opportunity for creation of a new generation of services based on smart spaces concept. The current approach for expanding mobile devices functionality is integration of new physical components. But this approach is bounded by the physical device size limits, dissipation of heat and the limited scalability of user experience due to small displays and incapability to produce high-end experience (e.g., audio) to the user. The smart spaces maximize the user benefits by utilizing capabilities of all available devices. This leads to a shift in the concept when instead of putting new functionality into the devices, all consumer electronics become a building blocks of the common information and service spaces. The smart spaces also provide another level of handling the user data. However, development of the smart spaces where a number of devices can use a shared view of resources and services is related to a number of problems. One of such problems is how to resolve possible conflicts arising from attempts of simultaneous access to the shared information. This paper describes an approach for coordination of anonymous agents, which solve this problem for the Smart-M3 smart space.**

***Keywords:* Smart Spaces; Use cases for consumer electronics; Smart-M3; Agents coordination; Shared information; Anonymous agents.**

I. INTRODUCTION

Modern device usage is moving towards so called “smart spaces” where a number of devices can use a shared view of resources and services [1], [2]. Smart spaces can provide better user experience by allowing the user easily integrate new devices into personal information infrastructures and allow seamlessly access all information distributed over the multi-device system from any of the devices. Examples of smart spaces can be found in [3], [4], [5]. One of the essential features assumed by such environment is the information sub-system that provides permanent robust infrastructure for storing and retrieving the information of different types from the multitude of environment participants.

Based on the analysis of earlier studies one can conclude that development of the Smart Spaces methodologies and techniques is a key requirement for creating attractive use case studies and building efficient developer eco-systems in the future. However, development of robust and efficient Smart Spaces solution is related to a need of addressing a number of practical problems. One of the problems to solve is coordination between the smart space participants, e.g., for resolving conflicts of simultaneous access to the shared data resource. To some extend this problem looks similar to the well known problem addressed in the database management systems, but after deeper study a lot of key differences could be identified. In computer science, the Atomicity, Consistency, Isolation, Durability (ACID) [6] is a set of properties that guarantee that database transactions are processed reliably.

The database modification procedure must follow the atomicity states, which implements “all or nothing” principle and refers to an ability to guarantee that either all of the transaction tasks are performed or none of them. Each transaction is said to be “atomic”, when if one part of the transaction fails, the entire transaction fails and the original state is preserved.

The consistency property ensures that the database remains in a consistent state before the start of the transaction and after its end (whether successful or not). It guaranties that only valid data could be written to the database. If for some reason, a transaction that violates the database consistency is executed, the entire transaction will be rolled back and the database will be restored to the last consistent state. On the other hand, every successfully executed transaction takes the database from one consistent state to another state that is also consistent.

The isolation refers to the requirement that other operations cannot access or see the data in an intermediate state during the transaction. This constraint is required to ensure good performance and guaranty inter-transactions consistency.

The durability is a guarantee that once the user has been notified about the success of the transaction, this state will persist. This means that the database must survive system failures and that the system already has checked the integrity constraints and won't need to abort the transaction. Many databases implement durability by writing all transactions into a transaction log that can be played back to recreate the system state right before a failure. In this case the new transaction can only be deemed committed after it is safely loaded into the log.

The well known and widely used in programming solution for restricting access to the shared resources is use of semaphores. The semaphore operations must be atomic, which means that no process may ever be preempted in the middle of one of those operations to run another operation on the same semaphore. There is a number of different implementation of semaphore principles, starting from a simple protected variable that locks/unlocks a certain resource and up to

 ISBN: 978-1-61208-100-7

counting semaphores which are the counters for a set of available resources, rather than a locked/unlocked flag of a single resource [7]. The semaphore value is a number of units of the resource that are free. If there is only one resource, a "binary semaphore" with values 0 or 1 is used.

Another solution used in concurrent programming is a monitor. The monitor is an object intended to be used safely by more than one thread [8]. The defining characteristic of a monitor is that its methods are executed with mutual exclusion. So for each point of time, at most one thread may be executing any of its methods. This mutual exclusion greatly simplifies reasoning about the implementation of monitors compared to the code that may be executed in parallel. The monitors also provide a mechanism for threads to temporarily give up exclusive access in order to wait for some condition to be met, and after that regain exclusive access and resuming their task. Monitors also have a mechanism for signaling to other threads that such conditions have been met.

However studying of the available solution has discovered that all of them are not suitable for the anonymous agent coordination in smart spaces, so the new solution has to be defined, which had been defined as a main target for this study. The next section provides an overview of the use case scenario that has been used as a main reference for studying the proposed solution. In Section 3 we present basic reference model of the discussed smart space. The method of resolving possible conflicts arising from simultaneous access to the shared information is described in Section 4. The following Section 5 gives a description of the developed demo prototype of the proposed solution for the reference use case scenario. The main results and findings of our study are summarized in Conclusion section.

II. SMART SPACE USAGE SCENARIO

The reference use case scenario describes a meeting taking place in a “smart room”, equipped with an intelligent whiteboard and a projector. The meeting participants have mobile devices (smartphones, PDAs, laptops, etc.) that store the appointments of the participants and their personal data, e.g., contact information, areas of interests, etc. Those meeting participants that are planning to make presentations have their presentations available on the mobile devices or accessible via internet/intranet (most important that the mobile devices always “know” how to access them).

In extension of the use case scenario defined in the previous works [9], this scenario is targeted in demonstrating the coordination function for resolving problems that arise due to possible simultaneous access to the shared information.

When the meeting participants entering to the room, their mobile devices discover the available smart space facilities, e.g., the whiteboard, and engage the handshaking protocol. If a participant wants to make a presentation, his/her mobile device is sharing the following information about the user: name, photo, domain of interests, e-mail, and phone number; and the presentation information: title, keywords, URI.

It is also necessary to schedule the presentations and create the meeting agenda. In this scenario the scheduling is done in the following simple way. There are several time slots covering whole time of the meeting. When a user comes, his/her presentation is scheduled into a free time slot. This is done by updating appropriate information units in the meeting room smart space, like it is illustrated by Figure 1.

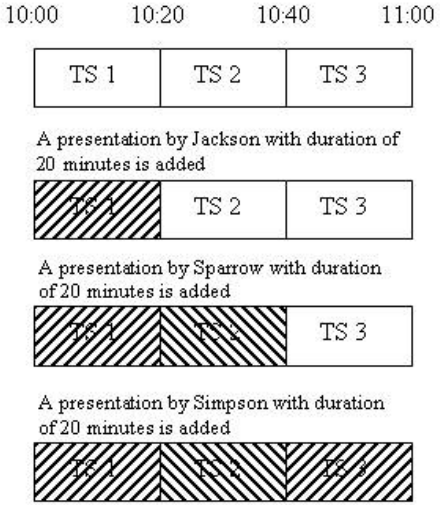


Figure 1. Scheme of scheduling presentations to the available time slots.

But the schedule conflicts can occur if two or more users simultaneously trying to schedule presentations (within resource request transmission and processing time delay) to the same time slot as it is shown in Figure 2.

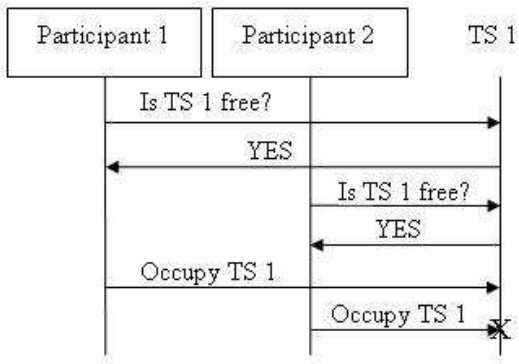


Figure 2. Possible conflict due to simultaneous access to shared information.

As a result, before the meeting starts the agenda is shown on the whiteboard including the speakers’ names and presentation titles. However, the same time slots can be occupied by different presentations or even some presentations will be lost from the list. The meeting participants can see the detailed agenda on the screens of their mobile devices, but agenda

 ISBN: 978-1-61208-100-7

might look differently for different people. The case can be even further complicated when some additional services are implemented, e.g., the presentation keywords could be translated to the preferred language (using the translator KP, which is also a part of the smart space), when the preferred language is taken from the user profile (the translator KP implements an interface to one of the Internet translation services). And the result M3 implementation will look like it is shown in Figure 3, where KP1 is a whiteboard, KP2 is a projector (PKP), KP3 is a translator and KP4…N are KPs of users' mobile devices (UKPs).

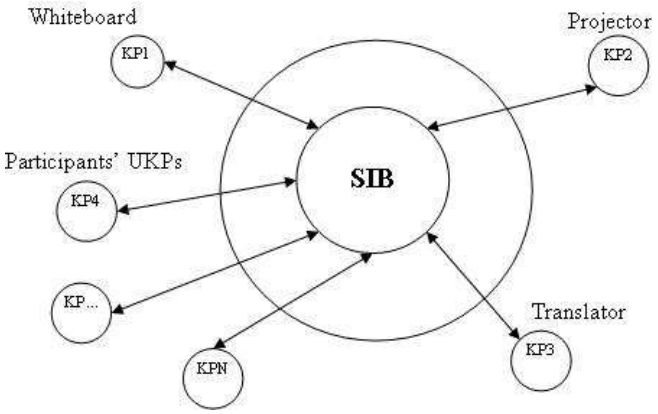


Figure 3. Current view of the proposed use case scenario.

Later in the paper we will show how this reference use case scenario can be implemented using the proposed coordination solution.

III. SMART SPACE REFERENCE MODEL

The general reference model of the discussed smart space could be illustrated by Figure 4.

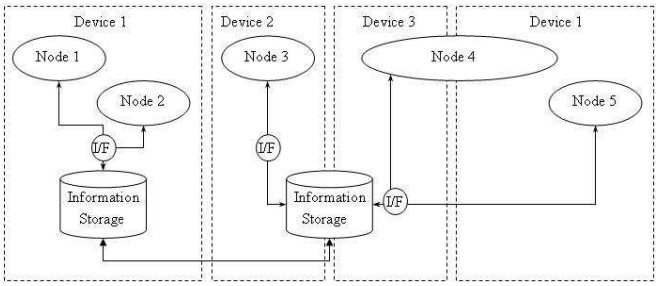


Figure 4. The reference model of discussed smart space.

Where:

Nodes - are logical elements capable to perform certain actions. One node can be distributed over several physical devices and several nodes can be located at the same device.

Information storages - also are logical units that store users information and can be distributed over several devices and several information storages can be located on at the same device.

I/F is an *interface* - that provides information exchange between the nodes and information storages. The interface is considered to be fully reliable and does not create additional delay and energy overheads. In this reference model the interface performs a technical function of connecting nodes to information storages. It does not implement logical functions and does not affect information transfer costs. For this reason the interface is not considered in the mathematical model.

Information is described by *information units* (IU) - represented as logical expressions: "subject"-"predicate"-"object" = [true | false], where *subject* is an actor (human or node that performs certain actions), *predicate* is an action that is being performed or supposed to be performed (e.g., "playing music") and *object* is what the action is performed with (e.g., a song being played). The nodes have predefined *behavior rules* defining their actions in line with the received information units.

From the implementation point of view the smart space can be illustrated as is shown in Figure 5.

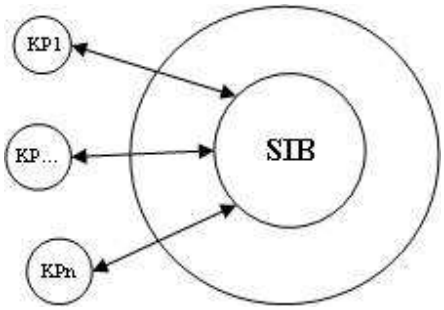


Figure 5. The Smart Space from implementation point of view.

The smart space itself consists of one or several Semantic Information Brokers (SIBs). The rules of information usage (applications) are implemented in knowledge processors (KP) connected to the smart space via SIBs. The SIBs are responsible for storing smart space information and its sharing: as soon as an information unit becomes available for the SIB, it becomes available for every KP. The knowledge processors are responsible for information processing.

IV. COORDINATION FOR CONFLICT RESOLUTION

So let's assume that we have a space that is used by 2 users. The users interact with the space by using their standard Knowledge Processors (KP), e.g., "u1" and "u2" correspondingly. If the "u1" needs to occupy certain resource R1, it currently only has to check that the statement: {"R1", "is_occupied_by", None} is valid, and if it is true then the KP "u1" can submit the triple: {"R1", "is_occupied_by", "u1"} to occupy it.

However, this works fine as long as we can guaranty that u1 and u2 will not try to simultaneously get access to the same resource, where term simultaneously is defined by the time interval from the moment when u1 has executed the first triple

 ISBN: 978-1-61208-100-7

and till the moment when it executes the second triple. But if during this time interval the node u2 will try to do the same, it also will get information that R1 can be occupied, which will result in resource access collision, as both nodes will have logical permission to occupy resource R1. As a consequence handling of the second triple becomes very complex procedure and independently of what tricks and fixes we will introduce at this stage with high probability it will lead to the logical errors and inconsistencies.

So in order to overcome the described above problem we introduce a special type of KP – the Coordinator KP. The Coordination KP acts as a kind of resource access manager. However, unlike classical resource manager solutions, which assume presence of a centralized application, to which all other application should send their resource requests, the functionality of Coordination KP is done based on principles described in the previous chapter. Most important that other applications do not need to know about presence of the Coordination KP in the space.

The coordination is performed seamlessly, automatically and anonymously by introducing a special set of RDF triples for handling access to the critical resources. Also the Coordination KP is subscribed to special triples that monitor all "resource access requests" and handles these requests on behalf of SIB, so that other KPs will not notice it. Below is the explanation how it works:

The Coordinator KP is subscribed to the information unit (RDF triple): {None, "check-insert", None}, where "None" logically means "any".

As a result, with the Coordinator KP the above scenario is changed as follows: the KP "u1" inserts the following rule into the smart space: {"R1, is_occupied_by, u1", "check-insert", "None"}, and subscribes to {"R1, is_occupied_by, u1", "check-insert-result", None}. The Coordinator KP checks the existence of the triple: {"R1", "is_occupied_by", None}. If it exists, the Coordinator KP inserts the triple {"R1, is_occupied_by, u1", "check-insert-result", "failure"}, the KP "u1" receives the result "failure" since it is subscribed. If the triple does not exist the Coordinator KP inserts the triple {"R1", "is_occupied_by", "u1"} and the triple: {"R1, is_occupied_by, u1", "check-insert-result", "success"}. The KP "u1" receives the result "success" since it is subscribed.

In case of simultaneous insert of rules by two KPs (u1 and u2): ("R1, is_occupied_by, u1", "check-insert", "None") and ("R1, is_occupied_by, u2", "check-insert", "None"), the Coordinator KP inserts the rule for the first KP and doesn't insert it for the second one. After that the KP u1 will occupy the resource R1 and the KP u2 will have to try to occupy some other resource, which can be offered to it by the Coordinator KP or defined internally by the KP u2. The result M3 smart space architecture with the Coordinator KP is presented in Figure 6, where the dotted lines show the information flow coming via the Coordinator KP.

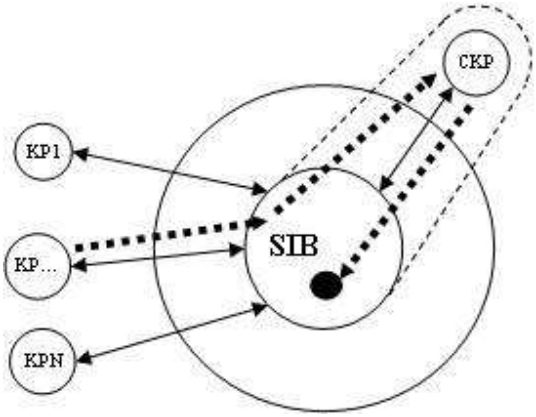


Figure 6. Organization of the information flow via Coordinator KP.

Further in-deep description of how the proposed solution can be implemented and used for the reference use scenario and the role of the Coordinator knowledge processor are discussed in the next chapter. Please also note that the same principle of coordination can be implemented completely inside the SIB.

V. IMPLEMENTATION OF THE SCENARIO

This scenario has been implemented using 6 personal computers (PC controlling the whiteboard, PC controlling the projector, PC controlling the Coordinator, PC controlling the Translator) and one Nokia N810 Internet Tablet emulating the user's mobile device. The other two user mobile devices were emulated on PCs. A proprietary M3/Piglet toolkit has been used for the prototype development. The knowledge processors were implemented using Python programming language.

Figures 7–10 are the screenshots for different knowledge processors at different stages of the scenario, e.g., Figure 7 shows work of the KP installed on the MAEMO device of the first meeting participant, which presentation was assigned to the first time slot. Figure 8 reflects work of the KP of the third meeting participant. The presentation was assigned to the third time slot. One can also see the translations of the presentation keywords, which were translated into Finnish language. The execution log of the Coordinator KP is shown in Figure 9. It lets the UKP of the first participant to occupy the time slot TS 1, the second participant to occupy the time slot TS 2, and the third participant to occupy the time slot TS 3.

The result output for the whiteboard KP is shown in Figure 10. One can see how the agenda is built based on users entering the room and occupying presentation time slots. Then, during the meeting the current presentation is highlighted (with three asterisks).

 ISBN: 978-1-61208-100-7

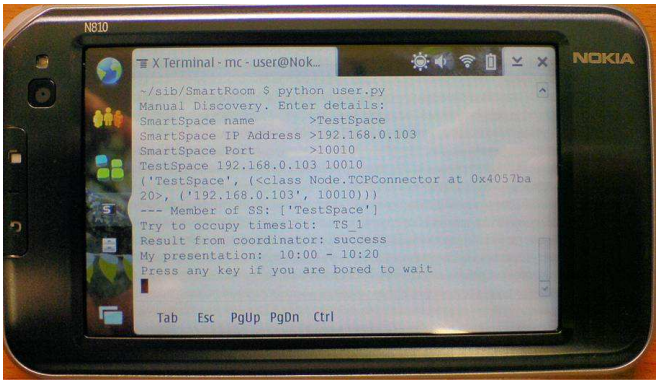


Figure 7. Screenshot of UKP running on Nokia N810 MAEMO device.

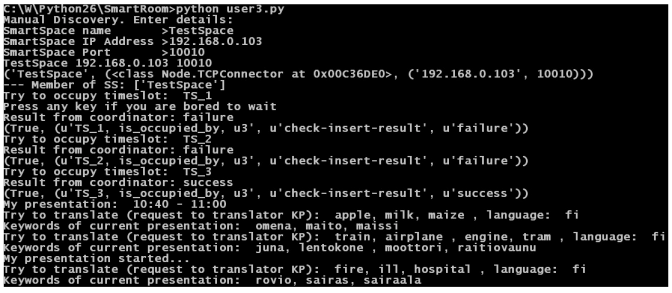


Figure 8. The status window of KP on PC of the third meeting participant.

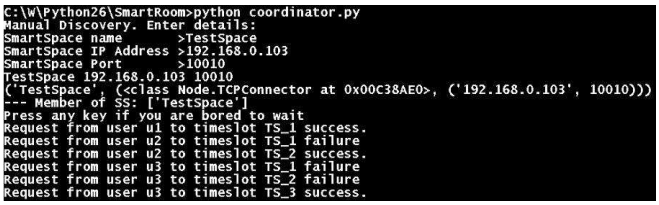


Figure 9. Log-output window of the Coordinator knowledge processor.

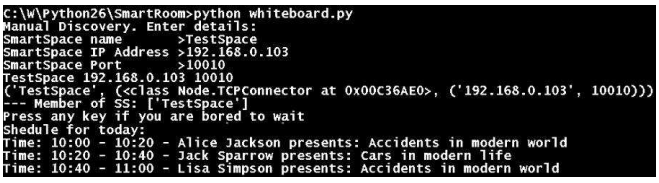


Figure 10. Log-output window of the Whiteboard knowledge processor.

As a result, nowadays we have full implementation for all of the above described elements that allow performing anonymous agents' coordination in Smart Spaces.

VI. CONCLUSIONS

The paper describes a solution for anonymous coordination of the agents, which allows addressing and solving a huge set of problems arising from a possibility of simultaneous access to the shared information.

The existing mechanisms for solving similar problems, such as transactions (used in database management systems), semaphores and monitors (used in programming) could not be applied directly. As a result an additional coordinator knowledge processor implementing the required functionality was introduced and described in the paper.

The paper gives detailed description of this knowledge processor work principles, which are also illustrated using an example implementation of the proposed principle for the reference case study scenario.

ACKNOWLEDGMENT

This paper is done within scope of the joint project between St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS) and Nokia Research Center. Some of the results are due to research carried out as a part of the project funded by grants # 09-07-00436-a and 08-07-00264-a of the Russian Foundation for Basic Research, and project # 213 of the research program "Intelligent information technologies, mathematical modelling, system analysis and automation" of the Russian Academy of Sciences. The authors would like to thanks Nokia, Russian Foundation for Basic Research and Russian Academy of Sciences for the provided financial support.

REFERENCES

[1] Oliver, I. and Honkola, J. *Personal Semantic Web Through A Space Based Computing Environment*, Middleware for Semantic Web 08, proceedings of ICSC'08, 2008.

[2] Oliver, I., Honkola, J., and Ziegler, J. *Dynamic, Localised Space Based Semantic Webs*, WWW/Internet Conference, Freiburg, Germany, 2008.

[3] Oliver, I. *Design and Validation of a Distributed Computation Environment for Mobile Devices*, proceedings of European Simulation Multiconference: Modelling and Simulation, 2007.

[4] Oliver, I., Nuutila, E., and Seppo, T. *Context gathering in meetings: Business processes meet the Agents and the Semantic Web*, proceedings of the 4th International Workshop on Technologies for Context-Aware Business Process Management, 2009.

[5] Jantunen, J., Oliver, I., Boldyrev, S., and Honkola, J. *Agent/Space-Based Computing and RF memory Tag Interaction*, proceedings of the 3rd International Workshop on RFID Technology - Concepts, Applications, Challenges, 2009.

[6] Wikipedia: http://wikipedia.org/wiki/ACID. Retrieved: 16.7.2010.

[7] Dijkstra, E., *Cooperating Sequential Processes*, Technical Report EWD-123, Technological University, Eindhoven, Netherlands, 1965.

[8] Hoare, C. *Monitors: an operating system structuring concept*, Communications of the ACM, Vol. 17 No. 10, pp. 549–557, 1974.

[9] Oliver, I., Nuutila, E., and Törma, S. *Context gathering in meetings: Business processes meet the Agents and the Semantic Web*, proceedings of the 4th International Workshop on Technologies for Context-Aware Business Process Management (TCoB 2009), 2009.

 ISBN: 978-1-61208-100-7

Coordination and Control in Mobile Ubiquitous Computing Applications Using Law Governed Interaction

Rishabh Dudheria, Wade Trappe
WINLAB
Rutgers University
North Brunswick, NJ 08902, USA
Email: {rishabh,trappe}@winlab.rutgers.edu

Naftaly Minsky
Department of Computer Science
Rutgers University
Piscataway, NJ 08854, USA
Email: minsky@cs.rutgers.edu

***Abstract*—This paper introduces a mechanism for regulating the interactions between the members of an ad hoc, heterogeneous and mobile multi-agent system, in order to ensure reliable and secure coordination between them. We demonstrate this mechanism, and its importance, by describing its application to a police team whose mission is to manage (i.e., monitor and control) the traffic in an area, by operating on a set of traffic-related devices, such as draw bridges, traffic lights, and road blocks. In particular, we demonstrate how the following critical aspects of the working of such a team are provided for: a) reliable coordination between the team members; b) the ability of the leader of the team to steer its subordinates; c) reliable auditing of the operations of the team; and d) robustness of the team under certain unexpected adverse conditions, such as the unpredictable failure of the team leader. Beyond developing suitable formalisms for local regulation of actions and communications, performance tests have been conducted with the proposed implementation on the ORBIT testbed and the results presented show the viability of this approach.**

***Keywords*-Law Governed Interaction; ad hoc coordination; decentralized enforcement; security**

I. INTRODUCTION

Current mobile ubiquitous technology supports reasonably reliable and secure communications between pairs of agents. It does not, however, provide adequate support for ad hoc, heterogeneous, multi-agent systems, whose members need to coordinate dynamically with each other in order to carry out their function—whether they operate in a collaborative or competitive mode, or some mix of the two. Such dynamic coordination is required in many application domains, such as in *law-enforcement*, and *military* applications, where an ad hoc team of diverse individuals is assembled to carry out a complicated and open-ended mission; coordination is also required in an *impromptu marketplace* where consumers may interact with each other via their wireless devices to share content and trade various digital tokens; and in various applications involving *vehicular communications*.

Effective and trustworthy coordination, however, requires participants to conform to a common coordination protocol. For example, for car drivers to survive their passage through an intersection, they must coordinate with other car drivers. But this requires all drivers to comply with certain traffic laws, like the one that requires stopping at a red light.

The goal of this paper is, therefore, a reliable and secure mechanism for establishing common coordination protocols over ad hoc multi-agent systems, whose members interact with each other via wireless communication. We illustrate the importance of such a mechanism, and some of its required characteristics, via the following example.

An Ad Hoc Police Team Mission: Consider a team of police officers, whose mission is to manage (i.e., monitor and control) the traffic in a certain region. In particular, the team is responsible for operating a set of traffic-related devices, such as draw bridges, traffic lights, or road blocks. This they can do via a collection of sensors and actuators distributed in their domain. Moreover, suppose that the team is managed by a leader, who assigns the team members to various tasks, monitors their progress, and exerts control over what each team member can do.

For such a team to operate effectively and safely, it must operate according to an appropriate protocol—which we shall denote by P—that regulates the interaction among the team members, and between them and the various actuators. This protocol should facilitate effective coordination between the team members, so that, for example, it would never happen that two policemen attempt to raise or lower a draw bridge at the same time. P should also regulate the interaction between the team members and their leader, providing the leader with a degree of control over the behavior of the team members, and ensuring that the leader gets from each member the information it needs to manage the team. Moreover P must facilitate proper handling of various exceptions, such as the disappearance of the team leader, which would require the employment of a careful leader election procedure.

Ensuring that such distributed agents operate properly is difficult as such a protocol cannot, practically, be hard-wired into the communication devices in the police cars, because a single police car may be required to participate in various missions, subject to different protocols. We need a far more flexible technique for establishing a given coordination protocol over ad hoc multi-agent systems. In this paper we

 ISBN: 978-1-61208-100-7

have explicitly modeled the types of control that are needed in a wireless ad hoc network and, using the police example as motivation, we have devised a flexible technique for coordinating an ad hoc collection of agents. Our approach leverages Law Governed Interaction (LGI) [1], which was originally developed for regulating transactions over the Internet. Under our version of LGI, each wireless device would have a built-in generic *controller* that can interpret an arbitrary interaction protocol, written in a special protocol-language. With such a generic controller, addressing the challenges of the police-mission becomes easy as all we must do is: (a) write our protocol P in a language recognized by the controllers; and (b) load this protocol into all the controllers built into the team member cars, and into the various actuators on the road.

The rest of this paper is organized as follows. Section II describes related work. Section III presents an overview of LGI, which provides the mechanism for implementing such applications. In Section IV, we provide a motivating example of an ad hoc team of traffic police officers whose mission is to monitor the traffic-related situation along with its implementation using the concept of LGI. The architecture of our proposed solution is described in Section V. We report various performance tests of our implementation conducted on the ORBIT testbed [2] in Section VI. Finally, we conclude and provide directions for future work in Section VII.

II. Related Work

DRAMA [3] is a policy-based network management system for mobile ad-hoc networks. The policies are represented by event-condition-action rules concerned with configuration, monitoring, and reporting of management events in a network. DRAMA policies are enforced in a distributed manner by Policy Agents that are co-located with the managed network elements. Policy operations– such as enabling, disabling, or introducing new policies– are propagated between Policy Agents in a peer-to-peer manner. DRAMA, however, is not concerned much with controlling the communication between managed network elements, and has only a rudimentary and stateless access control capability.

Xu et al. proposed SATEM (Service-Aware Trusted Execution Monitor) [4], which is a partial realization of LGI at a lower layer running on a TPM. Notably, this implementation did not include statefulness. However, this work suggests the validity of our approach as they have shown the feasibility of implementing such enforcements using trusted platforms. Further, the authors have enhanced this work to provide a distributed mechanism that allows trusted nodes to create protected networks in [5]. Only nodes that can demonstrate their trustworthiness by proving their ability to enforce policies are allowed to become members of the protected MANET. This avoids attacks from untrusted nodes as well as prevents attacks from member nodes due to enforcement of network policy.

In [6], Viterbo et al. have proposed a system that applies regulatory mechanisms to coordinate the interaction among applications in ubiquitous computing. A Domain Regulation Service regulates the interaction between client and server applications based on an explicit set of rules and contextual data. This service in turn acts as a centralized entity and may become a bottleneck besides being a single point of failure. Their system does not support stateful policies.

Rei [7] is a policy language for pervasive computing applications that includes constructs for rights, prohibitions, obligations and dispensations (deferred obligations). Rei includes a representation of speech acts (delegation, revocation, request and cancel) that are used to decentralize control and support dynamic modification of policies. Rei is a flexible and an expressive policy language that allows various kinds of policies (such as security, privacy, management, conversation etc.) to be specified. However, to our knowledge, Rei does not provide any support for handling communication faults and stateful policies.

III. An Overview of LGI

We have used the LGI paradigm to define the regulation policies as *laws*. The most salient aspects of LGI laws are their *strictly local formulation* and the *decentralized nature* of their enforcement. In this section, we provide an overview of the LGI mechanism. The implementation of LGI for ad hoc networks is similar to the LGI implementation for the Internet by the *Moses toolkit* [8] with some modifications as described in Section V.

LGI is a mode of interaction that allows an *open* group of distributed heterogeneous *agents* to interact with each other with confidence that the explicitly specified policies, called the *law* of the open group, is complied with by everyone in the group [1]. The messages exchanged under a given law $\mathcal{L}$ are called $\mathcal{L}$-messages, and the group of agents interacting via $\mathcal{L}$-messages is called a *community* $\mathcal{C}$, or more specifically, an $\mathcal{L}$-community $\mathcal{C}_{\mathcal{L}}$.

The concept of "open group" has the following semantic: (a) the membership of this group can be very large, and can change *dynamically*; and (b) the members of a given community can be *heterogeneous*. LGI does not assume any knowledge about the structure and behavior of the members of a given $\mathcal{L}$-community. All such members are treated as black boxes by LGI. LGI only deals with the interaction between these agents. Members of a community are not prohibited from non-LGI communication across the Internet, or from participation in other LGI-communities.

For each agent x in a given $\mathcal{L}$-community, LGI maintains the *control state* $\mathcal{CS}_x$ of this agent. These control states, which can change dynamically subject to law $\mathcal{L}$, enable the law to make distinctions between agents, and to be sensitive to dynamic changes in their states. The semantic of the control state for a given community is defined by its law, and could represent such things as the role of an agent in this community, its identity, its privileges, or reputation, etc. The

 ISBN: 978-1-61208-100-7

$\mathcal{CS}_x$ is viewed as a collection of objects called *Terms*. For instance, under the $\mathcal{L}$ law (to be introduced in Section IV), a term with the value *role(officer)* in the control state of an agent denotes that the agent has been authenticated to be a genuine officer.

In the rest of this section we discuss the concept of law, its local nature, and describe the decentralized mechanism for law enforcement. The interested reader is referred to [1] for more detail regarding LGI.

A. The Concept of Law and Its Enforcement

The law of a community $\mathcal{C}$ is defined over certain types of events occurring at members of $\mathcal{C}$, mandating the effect that any such event should have; this mandate is called the *ruling* of the law for a given event. The events subject to laws, called *regulated events*, include (among others): the *sending* and the *arrival* of an $\mathcal{L}$-message; the *coming due* of an *obligation* previously imposed on a given agent; and the submission of a *digital certificate*. The operations that can be included in the ruling of the law for a given regulated event are called *primitive operations*. They include: operations on the control state of the agent where the event occurred (called, the *home agent*); operations on messages, such as *forward* and *deliver*; and the imposition of an obligation on the home agent. The ruling of the law is not limited to accepting or rejecting a message, but can mandate any number of operations, like the modifications of existing messages, and the initiation of new messages and of new events, thus providing the laws with a strong degree of flexibility. More concretely, LGI laws are formulated using an *event-condition-action* pattern. In this paper we will depict a law using the following pseudo-code notation:

upon <event> **if** <condition>
do <action>

where the <*event*> represents one of the regulated events, the <*condition*> is a general expression formulated on the event and control state, and the <*action*> is one or more operations mandated by the law. This definition of the law is abstract in that it is independent of the language used for specifying laws. Concretely, we used Java but note that Prolog is also a viable language for writing the laws. However, despite the pragmatic importance of a particular language being used for specifying laws, the semantics of LGI is basically independent of that language.

Thus, a law $\mathcal{L}$ can regulate the exchange of messages between members of an $\mathcal{L}$-community, based on the control state of the participants; and it can mandate various side effects of the message exchange, such as modification of the control states of the sender and/or receiver of a message, and emission of extra messages.

1) The Local Nature of Laws: Although the law $\mathcal{L}$ of a community $\mathcal{C}$ is *global* in that it governs the interaction between *all* members of $\mathcal{C}$, it is enforced locally at each member of $\mathcal{C}$, by the following properties of LGI laws:

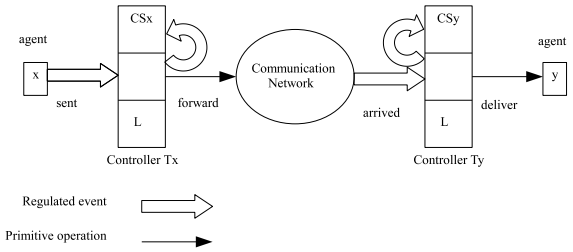


Figure 1. LGI framework achieves regulation of agents through controllers.

- $\mathcal{L}$ only regulates local events at individual agents.
- The ruling of $\mathcal{L}$ for an event e at agent x depends only on event e and the local control state $\mathcal{CS}_x$ of x.

The ruling of $\mathcal{L}$ at x can mandate only local operations to be carried out at x, such as an update of $\mathcal{CS}_x$, the forwarding of a message from x to some other agent y, and the imposition of an obligation on x. The fact that the *same law* is enforced at all agents of a community gives LGI its necessary global scope, establishing a *common* set of ground rules for the members of $\mathcal{C}$ and providing them with the ability to trust each other, in spite of the heterogeneity of the community. Furthermore, the locality of law enforcement enables LGI to scale with the size of the community.

2) Distributed Law-Enforcement: The law $\mathcal{L}$ of community $\mathcal{C}_\mathcal{L}$ is enforced by a set of trusted agents, called controllers that mediate the exchange of $\mathcal{L}$-messages between members of $\mathcal{C}_\mathcal{L}$. Every member x of $\mathcal{C}$ has a controller $\mathcal{T}_x$ assigned to it ($\mathcal{T}$ here stands for trusted agent), which maintains the control state $\mathcal{CS}_x$ of its client x. All these controllers, which are logically placed between the members of $\mathcal{C}$ and the communication medium as illustrated in Figure 1 carry the same law $\mathcal{L}$. Every exchange between a pair of agents x and y is mediated by their controllers $\mathcal{T}_x$ and $\mathcal{T}_y$, so that this enforcement is inherently decentralized.

3) The basis of trust between members of a community: For members of an $\mathcal{L}$-community to trust its interlocutors to observe the *same law*, one needs the following assurances: (a) Messages are securely transmitted over the network; (b) The exchange of $\mathcal{L}$-messages is mediated by controllers interpreting the same law $\mathcal{L}$; and (c) All these controllers are correctly implemented. If these conditions are satisfied, then it follows that if agent y receives an $\mathcal{L}$-message from agent x, this message must have been sent as an $\mathcal{L}$-message; in other words, that $\mathcal{L}$-messages cannot be forged.

We assume messages transmitted over the network are secured through proper cryptographic authentication and integrity mechanisms. To ensure that a message forwarded by a controller $\mathcal{T}_x$ under law $\mathcal{L}$ would be handled by another controller $\mathcal{T}_y$ operating under the same law, $\mathcal{T}_x$ appends the one-way hash [9][10] H of law $\mathcal{L}$ to the message it forwards to $\mathcal{T}_y$. $\mathcal{T}_y$ would accept this as a valid $\mathcal{L}$-message if and only if H is identical to the hash of its own law. As to the correctness of controllers, we assume here that every $\mathcal{L}$-community is willing to trust the controllers certified by a given $\mathcal{CA}$, which is specified by the law $\mathcal{L}$. In addition,

 ISBN: 978-1-61208-100-7

every pair of interacting controllers must first authenticate each other by means of certificates signed by this $\mathcal{CA}$.

B. Additional Features of LGI

Some of the further features of LGI are now discussed. For additional information, the reader is referred to [11].

1) The Treatment of Certificates: Certificates may be required by a given law $\mathcal{L}$ to certify the controllers used to interpret this law. Certificates may also be submitted by an actor x to its controller $\mathcal{T}_x$. The effect of such certificates is subject to the law in question. Typically, such submitted certificates are used to authenticate the identity of the actor, or the role it plays in the environment in which the community in question operates.

LGI currently supports the SPKI/SDSI model [12] for certificates. Under LGI, a certificate is a four-tuple *(issuer, subject, attributes, signature)*, where *issuer* is the public-key of the $\mathcal{CA}$ that issued and signed this certificate, *subject* is the public-key of the principal that is the subject of this certificate, *attributes* is what is being certified about the *subject*, and the *signature* is the digital signature of this certificate by the *issuer*. The *attributes* field is essentially a list of (attribute, value) pairs. For example, the attributes of a certificate might be the list [name(Joe), role(officer)], asserting that the name of the subject in question is Joe and its role in this community is that of a officer.

2) Enforced Obligation: Informally speaking, an *obligation* under LGI is a kind of *motive force*. Once an obligation is imposed on an agent (generally, as part of the ruling of the law for some event), it ensures that a certain action (called a *sanction*) is carried out at this agent, at a specified time in the future, when the obligation is said to come due, and provided that certain conditions on the control state of the agent are satisfied at that time. Note that a pending obligation incurred by agent x can be *repealed* before its due time. The circumstances under which an agent may incur an obligation, the treatment of pending obligations, and the nature of the sanctions, are all governed by the law of the community.

3) The Treatment of Exceptions: Primitive operations that initiate messages, like *deliver* and *forward*, may end up not being able to fulfill their intended function. For example, the destination agent of a *forward* operation may fail by the time the forwarded message arrives at it. Such failures can be detected and handled via a regulated event called *an exception*, which is triggered when a primitive operation that initiates communication cannot be completed successfully. It is up to the law to prescribe what should be done to recover from such an exception. The syntax of an exception event is *exception(op, diagnostic)* where *op* is the primitive operation that could not be completed, and *diagnostic* is a string describing the nature of the failure. The home of the exception event is the home of the event that attempted to carry out the failed operation. For instance, if a message m, forwarded by an agent x to an agent y operating under law $\mathcal{L}$ cannot reach its destination, then an event *exception(forward(x,m,[y,$\mathcal{L}$]),'destination not responding')* would be triggered at x. Commonly, exceptions are triggered by the *forward* and *deliver* primitive operation, as well as other communication primitives.

IV. An Ad Hoc Police Team Mission

This case study involves the police team mission introduced in Section I. We now elaborate on the structure and operations of this team. The team, whose purpose is to manage traffic in a given region by operating on a set of traffic-related devices (sensors and actuators), involves the following participants: (a) the *officers* who query the various sensors on the road, and operates on the various actuators; (b) a *leader* who monitors the activities of the officers participating in the mission and grants them access control rights to various devices; (c) a *supervisor* who maintains the information about the current leader, and has the ability to appoint a new leader, if the current leader fails, and to notify all team members of the new leader; (d) an *auditor* who maintains a log of messages sent to the various devices and provides this information to the leader whenever it requests for it.

We classify the messages sent by the officer into the following: *control messages* sent to the various devices (such as a command to raise a draw bridge, or to change the color of a traffic light), and *conversation messages* sent to any team member.

The members of the team and the sensors and actuators managed by them—collectively referred to as *agents*—operate according to protocol P specified informally below:

1) *Authentication of Identity*: For an agent to participate in the mission it must authenticate itself and its role via a certificate issued by a specific certification authority (CA) known as admin.
2) *Steering of the team*: The team of officers can be regulated by the leader in the following way: (a) the leader can grant and withdraw permission to an officer to access a particular device; (b) the leader has the right to query any required information from any of the officers taking part in the mission; (c) the leader has the power to stop the officer from taking part in the mission; and (d) the leader can assign a conversation message budget to any officer, which would restrict the number of arbitrary messages circulating in the network, thus reducing the possibility of congestion.
3) *Control Messages*: An officer is allowed to issue control messages to a device to which it has access rights. The copy of such a control message must be sent to the auditor.
4) *Fault tolerance*: The supervisor has the power to appoint a new leader, if the current leader fails to send him heart-beat messages.
5) *Control State Content and Conversation Messages*: Any member taking part in the mission should be able to access its control state to know the various terms

 ISBN: 978-1-61208-100-7

stored in it. The leader, auditor and supervisor are able to send arbitrary messages to each other, and to the plain officers. Each officer can also send an arbitrary conversation messages, to any team member, provided that it has sufficient budget for such messages.

A. Realization of Policy P via an LGI Law $\mathcal{L}$

To ensure that our mission team operates as required, we will have all team members, and all traffic related devices to be managed, operate under an LGI law $\mathcal{L}$ that realizes the policy P described above. They are, accordingly, called $\mathcal{L}$-agents, or simply agents. (Note that such agents can recognize each other as bona fide $\mathcal{L}$-agents.)

Before we get to law $\mathcal{L}$ itself we make the following preliminary comments: First, terms in each agent's control state are used to represent the role played by this agent. In particular, the control state of the current leader should contain a term, *role(leader)*. Likewise, the presence of term *budget(B)* in the control state of an officer x means that x has a budget of amount *B* and is entitled to send *B* conversation messages. An acting leader is forced to announce its identity periodically to the supervisor after every T_{report} seconds. If the current leader does not report to the supervisor within a predetermined time T_{fail} seconds, then it is assumed that the current leader has failed and the supervisor appoints a new officer as the leader of the mission.

Law $\mathcal{L}$ itself consists of two parts namely the *preamble* and the *body*. The preamble of $\mathcal{L}$ consists of the following clauses. First, there is the law clause that identifies the name of this law and the $\mathcal{CA}$ *admin* whose public key is used for the authentication of the controllers that mediate the messages of this system. Second, there is an *authority clause* that identifies the $\mathcal{CA}$ *admin* (represented by the keyed hash of its public key) for certifying the roles played by the different actors in this community. Third, the *initialCS clause* defines the initial control state of all actors in this community—it is empty in this case. Finally, the two *alias clauses* provide shorthand for the identifier (id) of supervisor and auditor respectively.

The law is now presented as a list of fragments along with their pseudo code, and explained in English.

1) Authentication of Identity: The fragment of the $\mathcal{L}$ law in Figure 2 shows how the authentication of identity takes place. When a participant engages in the system, it does so by sending an *adoption* message to its LGI controller, a message that can carry its certificate. When the message arrives at the controller, it invokes an *adopted event*. If an actor submits a certificate, then the controller verifies it with the public key of the $\mathcal{CA}$ *admin* and challenges it with the private key of the subject as shown by rule $\mathcal{R}1$. If the subject is not the one who presented the certificate, or if the issuer is not the $\mathcal{CA}$ *admin*, then no role and no identity is assigned to the actor and it is forced to quit. If the attributes of the certificate contain the role of *supervisor, leader, device or officer*, then this role of the

```
P reamble:
      law(name(L),authority(admin)).
      authority(admin,keyHashOfAdmin).
      initialCS( ).
      alias(supervisor,"supervisor@192.168.10.1").
      alias(auditor,"auditor@192.168.10.2").

R1)  upon adopted(Self,Issuer,Subject,Attributes)
        if(Subject!=Self or Issuer!=Admin)
           do Quit
        if(Attributes.role = supervisor)
           do Add(role(supervisor))
           do ImposeObligation(failure,600)
        if(Attributes.role = leader)
           do Add(role(leader))
           do Forward(Self,currentLeader,supervisor)
           do ImposeObligation(report,300)
        if(Attributes.role = auditor or device)
           do Add(role(Attributes.role))
        if(Attributes.role = officer)
           do Add(role(officer))
           do Add(budget(10))

R2)  upon adopted(Args)
        do Quit
```

Figure 2. Authentication of Identity: Fragment of the $\mathcal{L}$ Law

actor is extracted from the attributes and saved in the control state maintained by the controller on behalf of the actor. The leader reports its identity to the supervisor and an *obligation* is imposed on the controller of the leader to come due after T_{report} period (for example, we use a reporting time of 300 seconds). Also, an initial budget of $B_{initial}$ messages (say 10 messages) is provided to all the officers for initial arbitrary communication. The controller of the supervisor keeps a check on the status of the current leader through the obligation *failure*, which comes due after every T_{fail} period (assumed to be 600 seconds) since the last time a successful reporting was made. On the other hand, if no certificate is provided in the adoption message, then the actor will be automatically forced to *quit* as shown by rule $\mathcal{R}2$.

2) Steering of the Team: Figure 3 show the fragment of the $\mathcal{L}$ law that handles this process. Policy $\mathcal{P}$ allows the leader to *steer* the messaging activity of all the officers by suitably modifying their budgets. This provision is implemented by rules $\mathcal{R}3$ to $\mathcal{R}6$, which allows the leader to send a message of the form *incrementBudget(Amount)* or *decrementBudget(Amount)* to an officer, resulting in an increase or reduction in their corresponding budget by the specified amount. The leader can grant any participating officer the right to access a device pertaining to the mission (such as bridge, traffic lights, cameras, etc.) through rule $\mathcal{R}7$. According to rule $\mathcal{R}8$, when the controller of an officer obtains the right to access a device, the corresponding *permission* is added to its control state and the message is then delivered to the officer. Similarly, the leader can cancel any officer's right to access a particular device, which results in the removal of the corresponding permission term from the control state of the officer (rules $\mathcal{R}9$ and $\mathcal{R}10$).

The leader is authorized to request any desired information from an officer by sending a *requestInfo* message as shown by rule $\mathcal{R}11$. By rule $\mathcal{R}12$, an officer is obligated to

```
R3)  upon sent(C,incrementBudget(Amount),X)
          if (CS has role(leader))
             do Forward

R4)  upon arrived(C,incrementBudget(Amount),X)
          if(CS has role(officer))
             do Replace(budget(B),budget(B+Amount))
             do Deliver

R5)  upon sent(C,decrementBudget(Amount),X)
          if (CS has role(leader))
             do Forward

R6)  upon arrived(C,decrementBudget(Amount),X)
          if(CS has role(officer))
            if (B>Amount)
              do Replace(budget(B),budget(B-Amount))
            else
              do Replace(budget(B),budget(0))
            do Deliver

R7)  upon sent(C,grantAccess(Device),X)
          if (CS has role(leader))
             do Forward

R8)  upon arrived(C,grantAccess(Device),X)
          if (CS has role(officer))
             do Add(permission(Device))
             do Deliver

R9)  upon sent(C,repealAccess(Device),X)
          if (CS has role(leader))
             do Forward

R10) upon arrived(C,repealAccess(Device),X)
          if (CS has role(officer))
             do Remove(permission(Device))
             do Deliver

R11) upon sent(C,requestInfo(I),X)
          if (CS has role(leader))
             do Forward

R12) upon arrived(C,requestInfo(I),X)
          do ImposeObligation(requestInfo(C),180)
          do Deliver

R13) upon sent(X,replyInfo(I),C)
          if (CS has obligation(requestInfo(C)))
             do RepealObligation(requestInfo(C))
             do Forward(X,reply(replyInfo(I),
                controlState(Terms)),C)
          else
             do Deliver("Info_not_requested_by_this_
                destination")

R14) upon arrived(X,reply(replyInfo(I),controlState(
          Terms)),C)
          do Deliver

R15) upon obligationDue(requestInfo(C))
          do Forward (Self,notResponding(
             controlState(Terms)),C)

R16) upon arrived(X,notResponding(controlState(Terms)),
          C)
          do Deliver

R17) upon sent(C,stop,X)
          if (CS has role(leader))
             do Forward

R18) upon arrived(C,stop,X)
          do Deliver
          do Quit
```

Figure 3. Steering of the team: Fragment of the $\mathcal{L}$ law

respond to the request made by the leader within $T_{request}$ period (say 180 seconds). If the officer responds to the query posed by the leader within $T_{request}$ period, then the obligation is repealed, the control state terms are appended to the reply and forwarded to leader (by rule $\mathcal{R}13$). Such a reply message is simply delivered to the leader as per rule $\mathcal{R}14$. According to rules $\mathcal{R}15$ and $\mathcal{R}16$, if an officer does not respond to the obligation within $T_{request}$ period, then a *notResponding* message containing the control state terms is sent to the leader. The actions to be taken by the leader in such a circumstance are left to the discretion of the law of the mission at hand. Our law simply provides the ability to inform the leader of such a non responsive officer. The leader can dismiss any officer from taking part in the operations of the mission by issuing a *stop* message via rule $\mathcal{R}17$. By rule $\mathcal{R}18$, when a *stop* message arrives at the controller of the officer, the message is delivered to the officer and it is forced to quit.

```
R19) upon sent(X,operation(Parameters),D)
          if(CS has role(officer))
            if(CS has permission(D))
              do Forward
              do Forward(X,message(X,operation(
                 Parameters),D),auditor)
            else
              do Deliver("do_not_have_permission_to_
                 access_this_device")

R20) upon arrived(X,operation(Parameters),D)
          do Deliver

R21) upon arrived(X,message(X,operation(Parameters),D),
          auditor)
          do Deliver

R22) upon sent(C,query(Device),auditor)
          if(CS has role(leader))
             do Forward

R23) upon arrived(C,query(Device),auditor)
          do Deliver

R24) upon sent(auditor,queryResponse(R),C)
          do Forward

R25) upon arrived(auditor,queryResponse(R),C)
          do Deliver
```

Figure 4. Control Messages: Fragment of the $\mathcal{L}$ law

3) Control Messages: The monitoring function is carried out via the fragment of the $\mathcal{L}$ law shown in Figure 4. An officer can issue a *control* message (such as *operation(bridge(raise,speed)))* to operate on one of its accessible device, as shown by rule $\mathcal{R}19$. The necessary action to be carried out in response to this message is left up to the destination device. Our law simply provides the ability to deliver such a message to the device through rule $\mathcal{R}20$. According to rule $\mathcal{R}21$, a copy of such a control message is delivered to the auditor fulfilling the monitoring requirement. The leader can query the status of any device by sending a request to the auditor via rule $\mathcal{R}22$. By rule $\mathcal{R}23$, such a query message is simply delivered to the auditor. The auditor's response to the query is delivered to the leader

 ISBN: 978-1-61208-100-7

through rules $\mathcal{R}24$ and $\mathcal{R}25$.

```
R26) upon obligationDue(report)
          if (CS has role(leader))
             do Forward(Self,currentLeader,supervisor)
             do ImposeObligation(report,300)

R27) upon arrived(C,currentLeader,supervisor)
          do RepealObligation(failure)
          if(CS has leader(A))
             do Replace(leader(A),leader(C))
          else
             do Add(leader(C))
          do ImposeObligation(failure,600)

R28) upon obligationDue(failure)
          do Deliver(Self,appoint,Self)
          do Add(readyToAppoint)
          do Remove(leader(A))
          do ImposeObligation(failure,600)

R29) upon sent(supervisor,appoint,N)
          if (CS has readyToAppoint)
             do Remove(readyToAppoint)
             do Add(leader(N))
             do Forward

R30) upon arrived(supervisor,appoint,N)
          if(CS has role(officer))
             do Remove(role(officer))
             do Remove(budget(B))
             do Add(role(leader))
             do ImposeObligation(report,300)
             do Deliver

R31) upon exception(supervisor,appoint,N)
          do Remove(leader(N))
          do Add(readyToAppoint)
          do Deliver(Self,exception(appoint),Self)
```

Figure 5. Fault Tolerance: Fragment of the $\mathcal{L}$ law

4) Fault Tolerance: Figure 5 introduces the fault tolerance fragment of the law $\mathcal{L}$, which would allow our police team to recover from an unpredictable failure of its leader. We will consider the failure of the leader along with its controller to be of a *fail-stop* kind. We also assume that the supervisor and auditor do not fail. A broader perspective on such treatment of failures as part of self-healing under LGI can be obtained by referring to [13].

We adopt the concept of a *guardian* originally proposed by Tripathy et al. [14] to handle the failure of the leader. We assume that the supervisor acts as a guardian for the mission and is responsible for appointing an officer to the post of the leader whenever the current leader fails.

The leader is forced to report its status to the supervisor after every T_{report} period via an *obligation* as shown by rule $\mathcal{R}26$. The supervisor suitably updates the current leader information stored in its control state on receiving such an update (by rule $\mathcal{R}27$). In the absence of such a timely reporting, the obligation *failure* comes due at the controller of the supervisor as shown by rule $\mathcal{R}28$. The supervisor is then asked to appoint a new leader. This state of supervisor is characterized by the presence of the term *readyToAppoint*. According to rule $\mathcal{R}29$, when the supervisor sends a message to appoint some officer as the new leader, a new leader term for this appointee is inserted in its control state and the message is sent to the appointee. By rule $\mathcal{R}30$, when the forwarded *appoint* message arrives at an officer, it becomes the new leader. If an *exception* occurs while the supervisor is trying to appoint a new officer to the position of leader, then the corresponding leader term is once again removed from the control state of the supervisor and the term *readyToAppoint* is added back to the control state. Then, the supervisor is prompted again to appoint a new leader as shown by rule $\mathcal{R}31$.

```
R32) upon sent(X,getCS,Y)
          do DiscloseCS(all)

R33) upon sent(X,AnyOtherMessage,Y)
          if(CS has role(leader or supervisor or
            auditor))
           do Forward
          if (CS has role(officer) and budget(B))
           if(B > 0)
             do Replace(budget(B),budget(B-1))
             do Forward
           else
             do Deliver("Message_blocked_due_to_
                insufficient_budget")

R34) upon arrived(X,AnyOtherMessage,Y)
          do Deliver

R35) upon exception(E,D)
          do Deliver(Self,exception(E,D),Self)
```

Figure 6. Control State Content and Conversation Messages: Fragment of the $\mathcal{L}$ law

5) Control State Content and Conversation Messages: The participants of the system can check the terms stored in their control state and exchange various other messages via the rules given in Figure 6. Any participant can check the terms stored in its control state by sending a *getCS* message to its controller (by rule $\mathcal{R}32$). According to rule $\mathcal{R}33$, any participant (except the devices) can send any conversation message to another participant in the community. An officer can send a conversation message only if it has sufficient budget; the cost of which will be deducted from its budget. On receiving such a conversation message, the controller of the recipient simply delivers it to the actor as per rule $\mathcal{R}34$. If any other *exception* is raised, then the corresponding message and the reason for its failure is delivered to the sender by rule $\mathcal{R}35$.

B. Discussion

The law can be extended to achieve coordination in such a way that it would never happen that two officers issue contradictory control messages at the same time (for example, two officers should not be able to raise and lower the bridge at the same time) without knowing about each other. It is also possible to impose a restriction of changing the traffic light in front of a bridge to red before lowering the bridge. Further, it may be desired for certain missions to have the various devices (such as bridges, cameras, traffic lights etc.) work under their own law so that they can be operated independently by the officers (irrespective of the

 ISBN: 978-1-61208-100-7

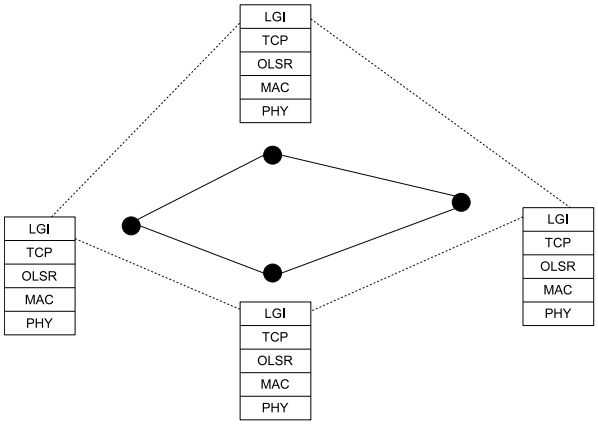


Figure 7. Our system runs LGI on a wireless ad hoc network protocol stack using TCP and OLSR.

law under which they are working). LGI supports this feature by allowing different policies to cross-interact, in a regulated manner, giving rise to *interoperability* between different LGI-communities. If a given community (like our police team) is required to operate in a context that imposes some global constraints on all wireless communication taking place in it, then the LGI policies can be organized in a *conformance hierarchy* [15]. We do not address these issues due to lack of space. Finally, a failure of the supervisor or the auditor can be addressed by replicating them. If we assume that the failure of the auditor or supervisor is very rare, then such replication would not adversely affect the scalability of the system.

V. LGI Architecture for Manet Applications

An example ad hoc network with LGI is shown in Figure 7. The LGI application runs on top of TCP. TCP is used instead of UDP for reliable delivery. The routing protocol used in the current implementation is Optimum Link State Routing (OLSR) [16]. We have chosen a proactive routing protocol such as OLSR over reactive routing protocol like Ad hoc on demand Distance Vector (AODV) routing [17] in order to minimize the message transmission delay. It should be noted that LGI as a mechanism is independent of the network routing protocol. We assume that each node has a trusted implementation of the LGI controller. Such an implementation requires the use of Trusted Platform Module (TPM) as specified by the Trusted Computing Group (TCG) [18], which we intend to do in future.

An actor adopts its controller with a particular law to become a member of the corresponding community. Laws can be made freely available on a server so that a user can download the appropriate law whenever it has internet connection. Another option is that the host can obtain the required law from a *certification authority* ($\mathcal{CA}$) when it applies for a certificate. The actor then communicates with the other participants in the community via its controller. We assume that mechanisms are in place to ensure that the nodes participating in the network can form a protected community so that outsider attacks can be prevented. This can be achieved using a secured routing protocol [19][20]. We disregard physical layer attacks as such consideration is beyond the scope of this paper.

To deploy LGI in an ad hoc network environment and test the implementation on the ORBIT grid [21], it was necessary to introduce subtle changes from the Moses Toolkit that operates over Internet via TCP/IP. The *preamble clause* can no longer specify the URL of the law in the *portal clause* in this implementation of LGI for MANET. Similarly, the *addPortal primitive* can no longer specify the URL of the law. Certifying Authorities can no longer be included in the law by specifying the URL of their public key. Human actors, controller service pool, and controller manager are no longer supported in this implementation. Features of the Moses toolkit that employed URL are no longer supported in the current implementation of LGI for ad hoc networks.

VI. Performance of LGI in a Wireless Ad Hoc Environment

The first part of this section introduces a model for the relative overhead of LGI-regulated communication based on a performance model published in [1]. The second part of this section reports on the *event evaluation time*, the *actor to controller communication time* and computes the *relative overhead* for messages by evaluating the *unregulated message transfer time* and the *LGI-regulated message transfer time* for a 5-hop topology. The results reported here are for laws written in Java rounded off to the nearest integer wherever appropriate. The experiments have been conducted on the ORBIT grid with nodes having a processor speed of 1 GHz on a Linux 2.6.12 platform. We have used the OLSRD 0.4.10 [22] implementation and 802.11a radio for communication. Further, a 5-hop topology is created using ORBIT tools to estimate the overhead due to LGI.

A. A Model for the Relative Overhead of LGI

Consider an LGI message m sent by an actor x to a destination actor y. This message is mediated by the controller of x ($\mathcal{C}_x$), which sends the message to controller of y ($\mathcal{C}_y$). (We denote controllers by the letter C here instead of the letter T used before, in order to avoid confusing it with the notation for time). Therefore, this message is converted to three consecutive messages: (1) from x to $\mathcal{C}_x$, (2) from $\mathcal{C}_x$ to $\mathcal{C}_y$, and (3) from $\mathcal{C}_y$ to y. The overhead $o_{x,y}$ due to the extra messages and the law-evaluations involved, is given by the following formula:

$$o_{x,y} = t_{com}^{x,C_x} + t_{eval}^{sent} + t_{com}^{C_x,C_y} + t_{eval}^{arrived} + t_{com}^{C_y,y} - t_{com}^{x,y} \quad (1)$$

where t_{eval}^{e} is the time it takes a controller to compute and carry out the ruling for the event e, and $t_{com}^{a,b}$ is the communication time from a to b. The relative overhead

$ro_{x,y}$ of an LGI message from x to y (as compared to the unregulated transmission of such a message) is defined as:

$$ro_{x,y} = o_{x,y}/t_{com}^{x,y}. \quad (2)$$

B. Measurements

1) Event evaluation time (t_{eval}): This experiment measures the time required by the controller to evaluate an event under LGI written as Java law. In this experiment, an actor adopts its controller with a sample law and then sends a message to its controller. The controller on evaluation of this sample law forwards the message to the same actor generating an *arrived* event, which in turn loops 100,000 times before getting delivered to the actor. The event evaluation time was averaged over 10 experiments, and gave:

Avg t_{eval} = 118 μs, Standard deviation = 2 μs.

This event evaluation time includes the time needed for local communication from actor to controller and back from controller to actor (within the same host), which is ignored as the experiment consists of 100,000 loops of actual event evaluations. Further, we have ignored the dependency on the events corresponding to different rulings of any LGI law. The variance of the event evaluation time for the different rules of the police team law introduced earlier is negligible.

2) Actor to Controller Communication time (t_{local}): This experiment finds the time it takes for a message sent by an actor to reach its own controller. A message is sent by the actor to its own controller, which on evaluation of the law, simply delivers the message back to the actor . This process is executed 100,000 times. The time obtained is averaged over 10 such experiment to get an accurate value. The average t_{delay} is 500 μs with a standard deviation of 4 μs. The delay measured in this experiment consists of the time taken for the actor to send a message to its controller (t_{local}) within the same host, event evaluation time (t_{eval}) corresponding to the ruling of the law at the controller and the time it takes for the message to be delivered back to the actor. Thus,

Avg t_{local} = (500 - 118)/2 = 191 μs.

On average, when an actor communicates with its controller on the same host, it takes t_{local}+ t_{eval} = 191+118 = 309 μs to receive (or to send a message) and to handle the associated event. Thus, the average throughput rate for the controller is 3236 events per second.

We now apply the model in [1] to evaluate the relative overhead of LGI-communication for a 5-hop topology.

3) Unregulated Message Transfer time ($t_{unregulated}$): In this experiment, we calculate the average time it takes for two hosts (separated by a 5-hop topology) to communicate with each other, i.e., the time required to transfer an unregulated message. A simple client is run on the sending node and a server application is run on destination node (both written in Java). The client program sends a message to the server program. This process is repeated 10,000 times. The resulting unregulated message transfer time is averaged over 10 such experiments to get an accurate value.

Avg $t_{unregulated}$ = 1.97 ms,
Standard deviation = 0.0015 ms

The unregulated message transfer time depends on many factors, such as message length, communication protocol, and distance between nodes. The distance between the nodes could not be varied much as these tests were run on the ORBIT indoor grid where nodes are spread over a distance of 80 ft by 70 ft [21]. In general, the delay caused by the message length and the distance between the nodes is negligible and has been neglected in these calculations. This unregulated message transfer time measurement does, however, take into account the overhead caused by the routing protocol (OLSR in our case).

4) Regulated Message Transfer time ($t_{regulated}$): This experiment used the same 5-hop topology. The actors on the sending and receiving nodes adopt their respective controllers with a Java law that simply forwards any message that is being sent and delivers any message that is received. The actor on the sending node sends a message to the actor on the receiving node. The in between nodes of the 5-hop topology act as routers and forward the message to the destination.

Avg $t_{regulated}$ = 2.4 ms, Standard deviation = 0.1 ms

The regulated message transfer time consists of the event evaluation for the ruling of the current law and local communication delay at the two end nodes along with the message transmission delay. Thus, the relative overhead is

$ro_{x,y}$ = (2.4 - 1.97)/1.97 = 0.22.

This overhead is far from prohibitive for most applications.

VII. Conclusion and Future Work

We have introduced a model of interaction control for the *regulation of wireless communication* in ad hoc networks using LGI to regulate the dynamic behavior of the interacting wireless agents. The power of the proposed mechanism resides in its ability to handle *statefulness*, *obligations*, *exceptions* and *locality*. There are many practical applications of such a system (e.g., police personnel at a sports event, medical personnel at an accident scene, emergency responders to a natural disaster, secure electronic commerce [23], manageable and robust multi-agent systems [24], etc.), yet little prior work exists that addresses these scenarios. We have prototyped an example based on a team of police officers in an ad hoc mission to control traffic. We have considered the critical elements of management of such an ad hoc team to provide: a) the leader with the ability to steer its subordinates and b) monitor relevant operations of its subordinates; and finally c) to provide robustness of the

 ISBN: 978-1-61208-100-7

agent-community under certain unexpected adverse conditions, such as unpredictable failure of the leader itself. We have shown that the overhead added due to LGI would not adversely impact performance. We plan to extend this work by implementing our mechanism on a TPM and extending it to support hierarchy of laws for ad hoc scenarios.

REFERENCES

[1] N. H. Minsky and V. Ungureanu, "Law-governed interaction: a coordination and control mechanism for heterogeneous distributed systems," *ACM Trans. Softw. Eng. Methodol.*, vol. 9, no. 3, pp. 273–305, 2000.

[2] M. Ott, I. Seskar, R. Siraccusa, and M. Singh, "Orbit testbed software architecture: supporting experiments as a service," in *Testbeds and Research Infrastructures for the Development of Networks and Communities, 2005. Tridentcom 2005. First International Conference on*, 23-25 2005, pp. 136 – 145.

[3] R. Chadha, H. Cheng, Y.-H. Cheng, J. Chiang, A. Ghetie, G. Levin, and H. Tanna, "Policy-based Mobile Ad Hoc Network Management," in *Policies for Distributed Systems and Networks, 2004. POLICY 2004. Proceedings. Fifth IEEE International Workshop on Policy for Distributed Systems and Networks*, Jun. 2004, pp. 35–44.

[4] G. Xu, C. Borcea, and L. Iftode, "Satem: Trusted service code execution across transactions," in *Reliable Distributed Systems, 2006. SRDS '06. 25th IEEE Symposium on*, 2-4 2006, pp. 321 –336.

[5] ——, "Trusted application-centric ad-hoc networks," in *Mobile Adhoc and Sensor Systems, 2007. MASS 2007. IEEE Internatonal Conference on*, 8-11 2007, pp. 1 –10.

[6] F. Viterbo, M. Endler, and J.-P. Briot, "Ubiquitous service regulation based on dynamic rules," in *Engineering of Complex Computer Systems, 2008. ICECCS 2008. 13th IEEE International Conference on*, march 2008, pp. 175 –182.

[7] L. Kagal, T. Finin, and A. Joshi, "A policy language for a pervasive computing environment," in *Policies for Distributed Systems and Networks, 2003. Proceedings. POLICY 2003. IEEE 4th International Workshop on*, 4-6 2003, pp. 63 – 74.

[8] C. Serban and N. H. Minsky, "The LGI website (includes the implementation of LGI, and its manual)," Rutgers University, Tech. Rep., Jun. 2005. [Online]. Available: http://www.moses.rutgers.edu_07.19.2010

[9] B. Schneier, *Applied Cryptography*. New York, NY, USA: John Wiley and Sons, 1996.

[10] R. Rivest, "The (MD5) message-digest algorithm," RFC 1321, MIT, Tech. Rep., Apr. 1992. [Online]. Available: http://www.ietf.org/rfc/rfc1321.txt_07.19.2010

[11] N. H. Minsky, "Law governed interaction (LGI): A distributed coordination and control mechanism (an introduction and a reference manual)," Rutgers University, Tech. Rep., Jun. 2005.

[12] C. M. Ellison, "The nature of a useable PKI," *Comput. Netw.*, vol. 31, no. 9, pp. 823–830, 1999.

[13] N. Minsky, "On conditions for self-healing in distributed software systems," in *Autonomic Computing Workshop, 2003*, 25 2003, pp. 86 – 92.

[14] A. Tripathi and R. Miller, "Exception handling in agent-oriented systems," in *Advances in exception handling techniques*. Springer-Verlag New York, Inc., 2001, pp. 128–146.

[15] X. Ao and N. H. Minsky, "Flexible regulation of distributed coalitions," in *LNCS 2808:the Proc. of the 8th European Symposium on Research in Computer Security (ESORICS) 2003*, Oct. 2003, pp. 39–60.

[16] T. Clausen and P. Jacquet, "Optimized link state routing (OLSR) protocol," RFC 3626, Oct. 2003. [Online]. Available: http://www.ietf.org/rfc/rfc3626.txt_07.19.2010

[17] C. Perkins, E. Belding-Royer, and S. Das, "Ad hoc on-demand distance vector (AODV) routing," RFC 3561, Jul. 2003. [Online]. Available: http://tools.ietf.org/html/rfc3561_07.19.2010

[18] *TPM Specification*, Trusted Computing Group (TCG) Std. 1.2, Rev. 103, Jul. 2007. [Online]. Available: http://www.trustedcomputinggroup.org/resources/tpm_main_specification_07.19.2010

[19] F. Hong, L. Hong, and C. Fu, "Secure OLSR," in *Advanced Information Networking and Applications, 2005. AINA 2005. 19th International Conference on*, vol. 1, 28-30 2005, pp. 713 – 718 vol.1.

[20] Q. Li, Y.-C. Hu, M. Zhao, A. Perrig, J. Walker, and W. Trappe, "SEAR: a secure efficient ad hoc on demand routing protocol for wireless networks," in *ASIACCS '08: Proceedings of the 2008 ACM symposium on Information, computer and communications security*. ACM, 2008, pp. 201–204.

[21] ORBIT Radio Grid Testbed. WINLAB, Rutgers University. [Online]. Available: http://www.orbit-lab.org/wiki/Tutorial/Testbed_07.19.2010

[22] A. Tonnensen. OLSR daemon 0.4.10. [Online]. Available: http://www.olsr.org/_07.19.2010

[23] C. Serban, Y. Chen, W. Zhang, and N. H. Minsky, "The concept of decentralized and secure electronic marketplace," *Electronic Commerce Research*, vol. 8, no. 1-2, pp. 79–101, 2008.

[24] N. H. Minsky and T. Murata, "On manageability and robustness of open multi-agent systems," in *SELMAS*, 2003, pp. 189–206.

 ISBN: 978-1-61208-100-7

Case Study of the OMiSCID Middleware: Wizard of Oz Experiment in Smart Environments

Rémi Barraquand, Dominique Vaufreydaz, Rémi Emonet and Jean-Pascal Mercier
PRIMA Team - INRIA
Zirst 655, avenue de l'Europe
38334 Saint Ismier cedex
{*remi.barraquand,dominique.vaufreydaz,remi.emonet,jean-pascal.mercier,omiscid-info*}*@inrialpes.fr*

***Abstract*—This paper presents a case study of the usage of OMiSCID 2.0, the new version of a lightweight middleware for ubiquitous computing and ambient intelligence. The objective of this middleware is to bring Service Oriented Architectures to all developers. After comparing to available solutions, we show how it integrates in classical workflow without adding any constraint on the development process. Developers only need to use a library available in widely used programming languages (C++, Java and Python). Then, the basics of OMiSCID and a brief technical description are described as its *User Friendly API*. This API makes it straightforward to expose, look for or send messages between software components over a network. The added value of the proposed middleware has already been experienced in international research projects. This paper demonstrates its effect in cutting down development time, improving software reuse and easing redeployment in the context of a Wizard of Oz experiments in intelligent environments.**

***Keywords*-Service Oriented Architecture, Ubiquitous Computing, Middleware, Wizard Of Oz, Smart Environments.**

I. INTRODUCTION

Today's vision of ubiquitous computing is only half way achieved. The multiplication of low cost devices along with the miniaturization of high performance computing units technically allow the design of environment widespread by cameras, motion detectors, automatic light controls, pressure sensors or microphones. Those devices, given the apparition of wireless networks, can communicate together with mobile and personal equipments such as cellular phones, photo frames or even personal assistants. On the other hand, the quiet and peaceful aspect of this vision where computing units can understand each others in order to collaborate is yet a research problem.

Build upon this network of devices, ambient intelligence tries to address the problem of making devices refer to user in an appropriate way by making them aware of his activity: current task, availability, focus of attention, etc. Such environments that sense user activity and act according to it are named "intelligent environments" or "smart environments".

However, activity understanding is yet a complex problem and relies on the ability to constantly aggregate information from an ever changing medium where devices come and go, break and evolve. Often used only to refer the access to information or applications from occasionally-connected devices, mobility is a key aspect of this vision.

For the user experience to be optimal anywhere anytime, ambient intelligence should be mobile. The intelligence can for example be embedded in a cell phone and dynamically adapt to the current environment and what services it provides. Ambient intelligence covers numerous fields and poses many challenges. Among these, handling dynamicity in software architecture is one that plays a central role.

This paper addresses the use of the OMiSCID [1] middleware that fits between the network of devices and ambient intelligence. It aims to ease the design of agile Service Oriented Architecture (SOA) and to solve constraints of pervasive computing and intelligent environments. OMiSCID orchestrates services in the environment, by providing cross-platform/cross-language tools for easy description, discovery and communication between software components.

The next section introduces the needs for such middleware. In section III, we present our approach focusing on key functionalities and concepts. Usage of OMiSCID is then illustrated by the design of a Wizard of Oz experiment. Finally, we draw some conclusions.

II. UBIQUITOUS COMPUTING REQUIREMENTS

The PRIMA team works on perception and perceptive spaces (using multiple cameras, microphones, wireless technologies, etc.), context awareness and ambient intelligence. In such research area, many services, developed by specialists using multiple techniques and languages, must interconnect in order to achieve a final goal. In the context of international researches, like DARPA or EU project funding, this problem is even more important because of different habits and of historical technical backgrounds within the involved research groups. In order to help non software-architect researchers to interact, we need a simple and usable solution that addresses a common problem: how to find, interconnect and monitor services within the context of cross-language cross-platform distributed applications?

To solve this problem, one can envision many different approaches. The first one is to decide, *a priori*, what will be the common way of programming. This solution can be

 ISBN: 978-1-61208-100-7

adopted in small groups and must be driven by underlying technologies. It may make people learn a new language or framework to interoperate and can changed every time involved scientists are replaced by new ones. The second possible scenario is to list all the software pieces and try to find a common way to do networking. In ubiquitous computing and ambient intelligence, this list will exhibit a variety of criterions:

- programming languages: C++ for video/audio processing, Java for lighter processing or user interfaces, scripting language.
- operating systems: Windows (for device support for instance), Linux, Mac OSX (cross-platform);
- data to exchange: simple texts, data structures or huge data like audio/video flows (messages or stream flows);
- configuration: peer to peer connection (IP address/port) or more complex interconnection scripts using service description and recruitment (dynamic service discovery);
- maintainability and sustainability.

Their are three main available solutions, the two most widely used are compared in the Table I. OSGi [2] is the first typical solutions to provide most of the requirements listed formerly. It permits construction of Java applications locally by recruiting components. Using iPOJO [3], it is possible to describe services and requirements in order to avoid writing dedicated code. Using specific adapters, like in R-OSGi [4] it is also possible to search for non local services. H-OMEGA [5] proposes also an alternative using UPnP for device discovery and a centralized server for code management. Nevertheless, even if it is always possible to use JNI for C/C++ application, OSGi is dedicated to Java. We do not retain this solution.

Web Services [6] are also a widely used solution for distributed applications. They permit using web technologies to construct distributed applications. Services are described with the *Web Services Description Language* (WSDL) and can be discovered using *WS-Discovery*. As said in Table I, *Web Services* are not designed for huge stream flows. Moreover, event if there are several alternatives to WSDL, like BPEL4WS[1] [7] or OWL-S[2] [8], they all provide service descriptions that are not easy to handle for a non specialist. We do not retain this solution.

The last possible solution is to use one of the specialized middlewares usually dedicated to a specific task and/or environment. We can illustrate them by focusing on *Smart Flow II* [9]. This middleware is very efficient in managing the data flow from many multimedia sources at the same time on several computers. But its force is also its weakness: it is difficult to configure and to manage other type of data.

From the previous sections, we can see that none of these solutions fulfills all the identified requirements. This assumption was the start of the OMiSCID middleware development. In the following sections, we will present the underlying concept and philosophy behind OMiSCID middleware solution.

[1] *Business Process Execution Language for Web Services*
[2] *Web Ontology Language for Services.*

III. OMiSCID Basics

OMiSCID stands for Opensource Middleware for Service Communication Inspection and Discovery [1]. It was designed and built to answer the problem of integration and capitalization of heterogeneous code inside augmented environment. OMiSCID is distributed with a non sticky MIT-like license, fully open source and available on a dedicated website: http://omiscid.gforge.inria.fr/.

A. Concepts

The OMiSCID middleware is built around 3 main concepts: services, connectors and variables. A service is a piece of software with a well defined way to access its provided functionalities. Services are composed of connectors and variables. At least, a service has the following read only variables: a name, a class, a unique service id and login/hostname information. Connectors are used to transfer data between services and are either input, output or both. Variables describe the service or its state. They may have local or remote write protection. Aggregating all these information, we obtain a service description that can be used to search and interconnect services.

B. Communications

Messages are atomic elements of all communications in OMiSCID. They are sent using a connector to a specific peer or to all listening services at once. The receiver will be notified that a new message is ready when it is fully available. Each message is provided with contextual information such as the service and connector it comes from. There are 2 main kinds of workflow for messages that can be mixed:

- A peer to peer approach. After receiving a message from a service and processing it, a response message is sent back to it;
- A data flow approach. After receiving a message on a connector and processing it, a message with the result is broadcasted on another connector in order to continue the processing chain.

Message can be sent as raw binary chunk or as text, which allows lot of flexibility for developers. Binary messages are often used to stream real time data such as video or audio. Text communication can be enhanced by the JSON format and allows for more advanced operation and extensions (see Section III-D).

C. Service discovery in dynamic context

Also known as service discovery, the ability to browse, find and dynamically bind running services, is one of the most important features of SOA moreover in ubiquitous

 ISBN: 978-1-61208-100-7

Table I
COMPARISON OF WIDELY USED SOLUTIONS

Description	cross-language	cross-platform	Messages	Huge data flows	Service discovery over the network
OSGI approach	No (Java)	Yes	Yes	Possible	Using *R-OSGi* for instance
Web Services	Yes	Yes	Yes	Not designed for	Using *WS-Discovery*

environments. It is common to filter services based on their current state, description or functionalities. OMiSCID provides the basic logical combination of predefined search criteria (variable value/name, connector properties, etc.) and since they are implemented as functor (function object), it is easy to extend it with user defined filters. Filters can be used in two different ways:

- An ask-and-wait approach asking for the list of services that match a certain criterion. This procedure will wait until at least one service match or that a timeout is reached rising an exception.
- An ask-and-listen approach notifying the application by the means of callback or listener whenever a service that matches the criterions appears or disappears.

Figure 2 and Section V-C gives clues about OMiSCID service discovery capabilities.

D. Serialization and remote procedure call

The philosophy of OMiSCID is to exchange information between services using either binary or textual messages without any standard on the format of those messages. OMiSCID 2.0 provides a simple way to marshal and unmarshal any object. This allow for easy communication between services without paying attention to serialization issues.

An OMiSCID service can expose some of its functionalities by the means of remote callable methods. Such distant calls can be done either in a synchronous or in an asynchronous manner.

E. OMiSCID Gui

OMiSCID provides a simple solution to declare, to discover and to interconnect services. Nevertheless, it needs, for maintainability and sustainability, an interface to visualize and to debug distributed services. OMiSCID Gui is a powerful tool built over the Netbeans platform and provides the developer with a graphical interface for multiple management tasks. It inherits many of the advantages from the Netbeans platform: portability, modularity, advanced window management, etc... OMiSCID Gui comes with light core modules and is extensible at infinite (see [10] for details). One of the core modules is a service browser that displays all the services present within the environment as well as their connectors and variables. This module also provides many contextual and extensible operations to be applied on the listed services and their variables/connectors. Among all the extensions available and easily installed using the Netbeans Plugin interface, one can find:

- A simple variable plotter than can dynamically create and display graphics of remote service variables.
- A family of plugins that allow displaying 2D information such as video stream or custom shapes representing for instance regions of interest of a 3D tracker.
- A plugin that displays a graph of the services in the environment along with their interconnections.
- A lot of other plugins such as: real time audio stream player, 3D visualization tools, cameras controls etc...

OMiSCID Gui comes with a public plugin repository already packed with visualization, controls, debugging plugins and can be extended by developers. All Netbeans platform plugin can also be integrated to our platform. Its ease of use make it a must-have tool for OMiSCID development, demonstration or service oriented application development.

IV. BRIEF TECHNICAL DESCRIPTION

One crucial requirement when designing a middleware for such heterogeneous research area is to make it usable by most of the people involved.

A. Multiplatform/Cross-Language

OMiSCID was designed with cross-platform/cross-language capabilities in mind. There are actually 3 supported implementation of OMiSCID: C++, Java and Python[3] (PyMiSCID). Moreover, the Java version can be used from Matlab and any other language running on the Java virtual machine (JavaFX, scala, groovy, javascript, etc.) We also provide an OSGI abstraction layer that exposes OMiSCID with standard OSGI paradigm. OMiSCID was developed more as a set of guideline rather than a specification, all the implementations are fully written in the target language, thus ensuring speed, reliability, close integration with data structure and programming paradigm.

All versions are fully cross-platform and works on Linux, Windows and OS X both 32 bits and 64 bits. The C++ version uses an abstraction layer that provide common system objects like sockets, threads, mutexes, etc. The Java version can be used on portable device like PDA. All implementations on all supported platforms can interoperate with each other.

B. User Friendly API

In order to simplify interpersonal communications between OMiSCID users, we developed a common *User*

[3]Formerly, Python version was a simple wrapper to the C++ source code.

 ISBN: 978-1-61208-100-7

Friendly API. It was defined to be easy to learn, easy to use and portable in several languages. Indeed, concepts, methods, parameters follow the same API in C++, Java and Python. However each implementation takes advantage of the language specificities and design patterns.

The API also provides simple callback/listener mechanisms. Thus, one can be notified of many different events: a new connection from a service, a disconnection, a new message, a remote variable changed, etc.

C. Performance and scalability using OMiSCID

For the discovery process, registering 100 services over the networks from 3 differents computer takes less than 1 seconde. Searching for a service among hundreds using a simple variable value is less than 20 ms long. More specific searches using user-defined search filter (processing video stream to select a camera for instance) can obviously spend much more time. Latency for sending messages is around 4 ms in average comparing to usual TCP/IP connections. Detailed tests and explanations can be found in [10] and on the OMiSCID Web site.

V. Case Study

For the past few years OMiSCID middleware has been used in different research projects [11], [12], [10], [13]. In [10], OMiSCID is used to redesign a complete 3D Tracking system as well as an automatic cameraman. The redesign reduces the number of software components and has the advantage to provide shared and reusable services. For instance, both architectures use the same video grabbing services. This service provides real time streaming of camera images, that is simultaneously accessible by multiple remote services such as visualization services or image processing ones: movement detector, person detector, posture estimator. OMiSCID middleware allows robustness for service discovery, reliable communication, connection and disconnection but also ease the federation of services. In [13], [12], OMiSCID is used for the realization of a smart agent. The perception of the agent is provided by services dynamically discovered in the environment allowing the agent to construct a situation model [14] of the current situation. The agent is yet another service and, according to its perception, it is able to perform actions in the environment by sending orders to actuator services. In [13], the knowledge of the agent is distributed and can be stored on remote database using a combination of OSGi and OMiSCID. Each service developed is a reusable piece of software that ensures a decrease of development time along the years. The Wizard of Oz (WOz) experiment we conducted is the perfect example.

A. Requirements for Wizard of Oz

WOz experiment is a research experiment, in which subjects interact with a computer system (fake mobile phone, remote controlled robot, etc.) that subjects believe to be autonomous, but which is actually being operated or partially operated by an unseen human: the wizard. The goal of such experiments is to evaluate system's functionalities without actually implementing them for real. The functionalities validated by the experiments can then be implemented while reconsidering others, saving money and time. In most settings, subjects are located in a room along with the system to evaluate while the wizard operates in another room. Both rooms can be separated by a beam splitter allowing the wizard to observe and react accordingly to the subject(s) actions.

Setting up a WOz experiment requires an important preparation, even more if the settings have to be mobile and re-deployable: to be carried in different places. The wizard must have access to a multitude of information in order to control the system as best as possible. Without the presence of a beam splitter, the environment must be equipped with cameras, microphones and speakers to record and stream the scene in real time. Among those devices, the wizard also needs the proper controllers to remotely manipulate the system, which requires an extensive use of wireless or wired communication between software components: controllers and controllees. In addition, the coupling between software components has to be able to change and to be easy to achieve. Allowing for instance to deploy debuggers, loggers or visualization tools at runtime.

B. Experimental Settings

On an ongoing research project we sought to evaluate the behavior of subjects immersed in a ubiquitous environment while asked to teach a smart agent how to control the space. Among few, the objectives were to validate hypothesis about human-machine interaction as well as to collect constructive outcomes that will help future design of ambient systems. Four kinds of actor are to be considered in this experiment: the subjects, the smart agent, the environment and the wizards.

The subjects by group of 2 or 3, were asked to teach the agent how to control the environment in order to organize a small meeting. For instance a common task is to teach the agent to switch on the light when people enter the room and to switch it back off when everybody is leaving.

The agent is embodied by a personal mobile phone with wireless capabilities, on which we deployed a learning software and a simple user interface. This interface allows collecting real-time feedbacks from the subjects (good, bad) during the session. The agent connects through the wireless network to a situation modeler service that provides a situation model [12] of the current situation. When requested by the subjects the agent takes action in the environment by sending orders to actuator services. Subjects can reward the agent whenever they agree or disagree with its action. Doing so the agent learns to control the environment using dynamically discovered services and feedbacks provided by users.

 ISBN: 978-1-61208-100-7

The environment is an office (Figure1) spread with many actuators and sensors. OMiSCID allows each of them to be accessible and controllable by services all-over the network (Figure2). Among those sensors we find cameras, microphones, thermometer and weather station. All the actuators are controllable by OMiSCID connectors and their states can be queried by those connectors or are exposed through variables. Among the actuator we list: a steerable projector, x10 controller, speaker or even shutter. For this experiment we needed two wizards.

The first wizard was in charge of simulating certain actuators in the environment such as pressure detectors under the chairs and sofas. Indeed, we didn't dispose of such device in our environmental facility, thus we emulated those actuators using a user interface plugged into OMiSCID Gui. The simulating interface was seen as yet another service that can be used by the situation modeler to build up a better situation model.

The second wizard was controlling the overall experiment using a master interface. This interface allowed writing real-time observations through an annotator service, as well as taking control over all the services in the environment. Such a master control for instance let the wizard speed up the experiment by helping the agent to guess better actions (when subjects got exhausted), or by putting back the environment settings in an appropriate state.

C. OMiSCID At Glance

For this experiment we deployed more that 20 services spread on 5 computers running different operating systems (Linux, Windows, MacOSX). Figure 2 presents some of the devices present in the environment as well as the interconnection of services. Due to the complexity of the schema some services have been removed. In the following we review some of the advantages of using OMiSCID in this experiments:

1) Multi-platform: 5 computers have been used during the experiments, two of them by the wizards. One of the wizards was using MacOS, on which we deployed the master control. Due to driver issue the sound recording system was using a Microsoft powered computer. The video streaming as well as all the other services (archiver, x10, etc) were running on Linux hosts.

2) Multi-language: To design the services we made use of different languages. C++ was used for performance reason such as for the video and sound processing/capture services. Python is a really powerful language for the rapid prototyping of application, we used python to quickly develop the x10 controllers or the PanTilt controller. Java has been used to develop the OMiSCID Gui modules but also to access the different web services present in the environment such as the weather service. JavaFX was used to develop the wizard control's interface, its script language make it easy to use for inexpensive user interface design.

3) Service Discovery: The simple but powerful service discovery system provided by OMiSCID has been used to dynamically connect services together. The best examples are the *situation modeler* and *archiver*. Using a service repository they were able to filter services present in the environment in order to connect to them. For instance the *situation modeler* was looking for all services having connector or variable exposing state information. Using that state information, it was able to provide a situation model on an output connector. The *archiver* was responsible to backup any information transmitted between services on a hard drive. The archiver was looking for all services having output connector. Thus it was easy for instance to deploy or shutdown services on the fly during the experiment.

4) Communication: Communication between services was achieved using different format. For video and sound services, data were raw binary information tagged with time stamps. Web services such as the weather provider were communicating information using XML on their connector. The PanTilt controller exposed its commands by the means of remote callable methods, and presented its internal state using readable variable.

5) OMiSCID Gui: OMiSCID Gui was used by the wizard for different purpose. Firstly, the streamed sound and video were played by the embedded player. Indeed, we have developed OMiSCID Gui modules to play video and listen to audio stream in real-time. Those modules were used to have a feedback of what was happening into the experimental facility disposed into another building. Secondly, OMiSCID Gui was used to control the archiver and other services.

6) Reusability: Each of the service used in this experiment is a reusable piece of software that can be carried and deployed easily. For a wizard of Oz experiment only the hardware and the equipments (cameras, microphones) have to be transported and reinstalled. Everything else is deployable instantly and can adapt to the configuration: number of computers, operating systems, number and nature of devices, etc...

VI. Conclusion

OMiSCID provides a complete but simple solution to declare, describe, discover and interconnect services as well as to manage their inter-communication. OMiSCID offers to researchers several facilities for ubiquitous computing with its multi-platform and cross-language capabilities. The underlying concepts as well as the API are user friendly and directed toward usability. Along with this middleware, OMiSCID Gui provides developers with an extensible, portable and modular platform that ease development and debugging, and improve maintainability of OMiSCID demonstrations and applications.

OMiSCID has successfully been used in several academic research projects and more recently in a wizard of Oz experiment. Such an experiment requires an important amount

 ISBN: 978-1-61208-100-7

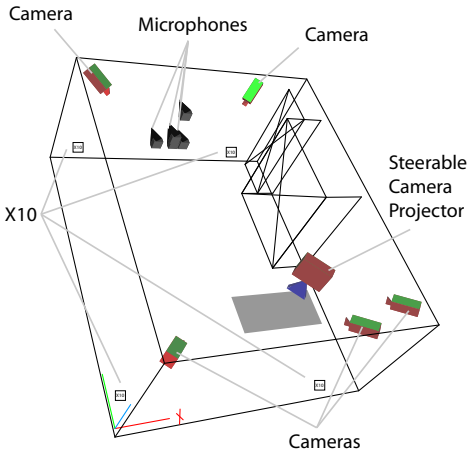


Figure 1. PRIMA's Smartroom.

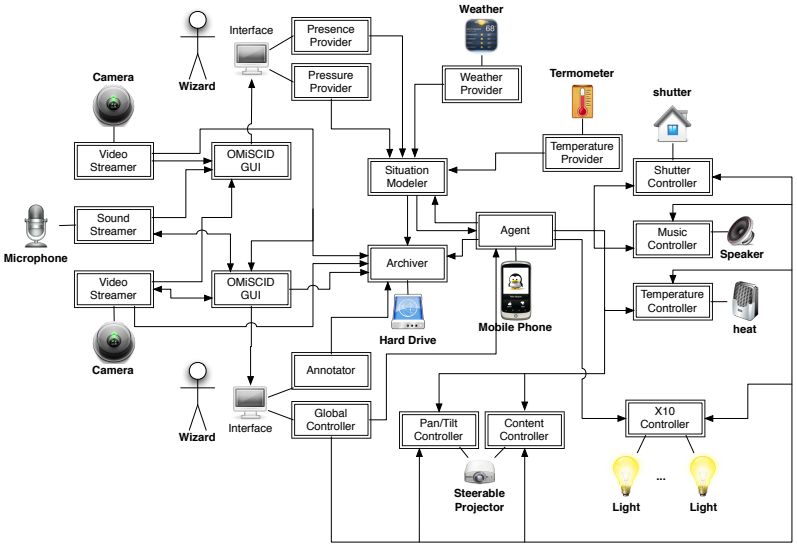


Figure 2. Wizard of Oz Services.

of resources and preparations, particulary when realized in smart environments. We have presented how the use of OMiSCID and OMiSCID Gui greatly reduces development time, maximizes reusability and eases redeployment. Furthermore, this solution does not rely on the number of computers, operating systems or network configuration.

REFERENCES

[1] R. Emonet, D. Vaufreydaz, P. Reignier, and J. Letessier, "O3miscid: an object oriented opensource middleware for service connection, introspection and discovery," in *1st IEEE International Workshop on Services Integration in Pervasive Environments*, Lyon (France), jun 2006.

[2] D. Marples and P. Kriens, "The open services gateway initiative: an introductory overview," *IEEE Communications Magazine*, vol. 39, no. 12, pp. 110–114, Dec. 2001.

[3] C. Escoffier, R. S. Hall, and P. Lalanda, "ipojo: an extensible service-oriented component framework," *Services Computing, IEEE International Conference on*, vol. 0, pp. 474–481, 2007.

[4] D. Wang, L. Huang, J. Wu, and X. Xu, "Dynamic software upgrading for distributed system based on r-osgi," in *CSSE '08: Proceedings of the 2008 International Conference on Computer Science and Software Engineering*. Washington, DC, USA: IEEE Computer Society, 2008, pp. 227–231.

[5] C. Escoffier, J. Bardin, J. Bourcier, and P. Lalanda, "Developing User-Centric Applications with H-Omega," in *Mobile Wireless Middleware, Operating Systems, and Applications - Workshops*. Springer Berlin Heidelberg, April 2009, pp. 118–123.

[6] M. Papazoglou, *Web Services: Principles and Technology*. Prentice Hall, September 2007.

[7] R. Khalaf, N. Mukhi, and S. Weerawarana, "Service-oriented composition in bpel4ws." in *WWW (Alternate Paper Tracks)*, 2003.

[8] D. Martin, M. Paolucci, S. Mcilraith, M. Burstein, D. Mcdermott, D. Mcguinness, B. Parsia, T. Payne, M. Sabou, M. Solanki, N. Srinivasan, and K. Sycara, "Bringing semantics to web services: The owl-s approach," in *SWSWPC 2004*, ser. LNCS, J. Cardoso and A. Sheth, Eds., vol. 3387. Springer, 2004, pp. 26–42.

[9] A. Fillinger, L. Diduch, I. Hamchi, M. Hoarau, S. Degre, and V. Stanford, "The nist data flow system ii: A standardized interface for distributed multimedia applications," in *World of Wireless, Mobile and Multimedia Networks, 2008. WoWMoM 2008. 2008 International Symposium on a*, 23-26 2008, pp. 1 –3.

[10] R. Emonet, "Semantic description of services and service factories for ambient intelligence," Ph.D. dissertation, Grenoble INP, sep 2009.

[11] J. L. Crowley, D. Hall, and R. Emonet, "Autonomic computer vision systems," in *Advanced Concepts for Intelligent Vision Systems, ICIVS 2007*, J. Blanc-Talon, Ed. IEEE, Eurasip,, Aug 2007.

[12] R. Barraquand and J. L. Crowley, "Learning polite behavior with situation models," in *HRI '08: Proceedings of the 3rd ACM/IEEE international conference on Human robot interaction*. New York, NY, USA: ACM, 2008, pp. 209–216.

[13] S. Zaidenberg, P. Reignier, and J. L. Crowley, "An architecture for ubiquitous applications," *Ubiquitous Computing and Communication Journal (UBiCC)*, vol. 4, no. 2, jan 2009.

[14] J. L. Crowley, P. Reignier, and R. Barraquand, "Situation models: A tool for observing and understanding activity," in *in Workshop People Detection and Tracking, held in IEEE International Conference on Robotics and Automation*, Kobe, Japan, 2009.

 ISBN: 978-1-61208-100-7

SmartBuilding: a People-to-People-to-Geographical-Places Mobile System based on Augmented Reality

Andrea De Lucia, Rita Francese, Ignazio Passero and Genoveffa Tortora
Dipartimento di Matematica e Informatica
University of Salerno
84084 – Fisciano (SA) – Italy
Tel:+39 (0) 89 963376
adelucia@unisa.it, francese@unisa.it, ipassero@unisa.it, tortora@unisa.it

***Abstract*— People-to-People-to-Geographical-Places systems connect places to communities. Generally, these systems require users to perform complex tasks using mobile phones. Thus, the creation of powerful interaction techniques and the design of effective and easy interfaces are key requirements. The system we propose in this paper, named SmartBuilding, is a mobile Augmented Reality system that supports the sharing of contextualized information in indoor environment, depending on the user location. This system shifts the interaction with the user from the classical desktop logical metaphors towards a more natural interaction style based on augmenting and annotating an indoor environment. It combines the world perceived by the phone camera with information concerning the user location and his/her community, enabling users to create several working areas and access to augmented content. SmartBuilding proposes an innovative interface that adopts the mobile device sensors to identify the content to be shown and to control the user interaction. Results on the usability of the proposed approach are positive.**

Augmented Reality; mobile user interfaces; context-awareness; People-to-People-to-Geographical-Places systems.

I. INTRODUCTION

Web 2.0 and social software are changing the way in which millions of users communicate. Users are not only receiver of content, but they contribute to the content creation in a bottom-up way, generating social networks and communities [3].

Changes in the people habits are also due to the great diffusion of mobile devices that enables users to be connected "anytime, anyway". The adoption of these devices enables also the diffusion of localization technologies. The most adopted is the Global Positioning System (GPS), for outdoor environments, while, in indoors environments recent location sensing systems range from RFID, requiring explicit installation, to standard wireless networking hardware, see [1][9] for example, or Bluetooth tags, suffering of several problems, such as long time for device discovering and passing through walls. More recently, WiFi triangulation and video detection approaches have been proposed.

According to Gartner, one of the ICT enterprise world leader, Location-Based services are at the second place in the first ten more required mobile devices applications in 2012 [7]. Moreover, according to Gartner, the request of this service will strongly increase in the next years. Gartner predicts that Location Based users will increase from 96 millions in 2009 to 526 millions in 2012. This kind of services is second in the top ten list for the high value that they have for users. In fact, they answer to several user necessities, from productivity to social network and entertainment needs.

Several People-to-People-to Geographical-Places systems have been proposed in literature [10], aiming at connecting social networks and communities to physical places. Actually, new powerful devices enable systems to incorporate place and people in new and powerful ways.

Augmented Reality (AR) is a technology that allows computer generated virtual imagery to exactly overlay physical objects in real time [22]. The integration of user localization, AR and of the feature offered by the top-of-the-range devices (on-board camera, accelerometers, compass, GPS etc.) enables the device to combine the camera preview with AR information in real-time. Thus, a mobile device can be seen as a window onto a located 3D information space, enabling "to browse, interact, and manipulate electronic information within the context and situation in which the information originated and where it holds strong meaning" [5]. Following this approach, information can be provided and created considering the context and the user profile. Mobile devices are small, thus researchers have to face the challenge of designing usable interfaces for device screens with limited dimensions and invent new interaction modalities.

In this paper, we propose a system, named SmartBuilding, which follows the metaphor of the "Cooperative Building" proposed in [19], .i.e. room elements with integrated information technology to support formal and informal communication. The approach we propose does not require specific hardware, except the user mobile devices, whose usage is largely diffused, and adopts innovative interfaces, which control the user interaction using the mobile device sensors.

II. RELATED WORKS

Several research projects, such as [4][14][16][19], investigate People-to-People-to-Geographically-Places systems displaying notes and messages [10].

E-graffiti [4] is a context-aware application, which detects the user's location and displays notes dependent on

 ISBN: 978-1-61208-100-7

that location. The application is evaluated and results show that location-specific notes were appealing to users. GeoNotes [16] is a location-based messaging system, which allows users to provide their contents in order to create a social and dynamic information space. It does not allow remote access to notes. All these systems provide a traditional menu-based interface.

Augmented reality has been adopted by the Augmented Reality Post-It (AR Post-It) messaging system, which, similarly to our approach, uses the mobile phone as an augmented reality (AR) interface allowing users to view electronic messages in an AR context [19]. There is the need of a paper marker in each specific location where messages are available, i.e. on the fridge.

Microsoft Research proposed Notescape, a tool that creates a "mixed reality" where virtual sticky notes appear to float in the physical space around the user's body [14]. The notes, using a mobile camera, follow the users as they move from place to place.

An AR interface has been adopted by CAMAR (Context-aware Mobile Augmented Reality), a system enabling users to interact with smart objects through personalized control interfaces on their mobile AR devices [15]. Similarly to our approach, it supports context-based contents augmentation and the sharing of contents among user communities. The main difference with the system we propose is that SmartBuilding does not need additional hardware: we associate the contents to a specific point of the room, while in CAMAR the interaction is limited to particular objects or markers.

The adoption of onboard sensors, like orienteer and accelerometer, to intercept user interaction, enables to implement novel and natural user interfaces. Even if not strictly related to the mobile technology, it is important to underline how Nintendo adopts low cost accelerometers in Wii controllers to enrich user experience and augment game usability. Wii controllers are equipped with onboard sensors and speakers to keep the user gaming experience analogically real. User movements reproduce the real actions and are captured and replicated on the screen, keeping the user involved in the experience; the controller speaker gives the player a better sense of immersion. Recently, mobile phones are providing the same interesting features, sometimes offering a more powerful technological platform: they are capable of detecting orienteering and acceleration and have the computational power to augment the preview obtained by the onboard camera. These devices may shift the user interaction from the classical keypad or button input towards on the phone movement [8]. As an example, Labyrinth Light [12], an Android application, controls a ball on a plane by reading the on board devices sensor. In a more complex way, Google Sky Map projects the user in an AR sky space. The device sensors are used to understand where the user is pointing his/her cellular phone and augments the screen with information about stars and planets.

Adopting the same technology, Layar [13] makes sets of data viewable on top of the camera of the mobile phone as one pans around a city and point at buildings. Layers are equivalent to web-pages in normal browsers. Real estate, banking and restaurant companies have already created layers of information available on the platform. The recent version 3.0 enables also to augment reality with 3D objects.

ARToolKit uses computer vision techniques to calculate the real camera position and orientation relative to marked cards, allowing the programmer to overlay virtual objects onto these cards [1]. Differently, in our approach the marked cards are adopted only to initially localize the user.

III. THE PROPOSED SYSTEM

SmartBuilding aims at augmenting a physical building with spatially localized areas in which users can share formal and informal multimedia documents and messages. The mobile devices are capable of projecting the user in an augmented world of information, controlling the camera interaction with on-board sensors (i.e. accelerometer and orientation). By moving his/her device in the surrounding space the user abandons the 2D desktop metaphor (i.e. folders and icons) and adopts spatial movements for exploring a new information space. The device screen becomes the touchable interface of this world and the user position and his/her profile provide the context.

SmartBuilding offers support for formal and informal communications. Indeed, it is possible to create administration areas, where formal communication is provided, as an example, describing some office procedure. Others information areas can be available for team work, or informal communication can be exchanged among specific user groups. Thus, the system enables each user to distribute his/her augmented content to specific colleagues, supporting selective content sharing and collaboration among people belonging to the same community. Voting and commenting features are also available and support information filtering.

A. The informative space

The device is used as a hand held lens giving a moving view on the AR scene. It is important to point out that the user needs to hold the device and, therefore, his/her maneuverability is quite reduced.

Basing on these observations, we decided to adopt the Azimuth orientation sensor, creating, as a consequence, a 360° space. This space constitutes a cylinder surrounding the user and ideally lays near the walls of his/her room. In our approach, the Azimuth is the main dimension of the augmented reality. The Pitch orientation sensor is adopted combined with the accelerometer, to detect how the camera is orientated in the space vertical dimension. This information enables to provide feedback to user by prospectively deforming the projected objects according to the device spatial orientation and is at the core of the area pagination system later presented.

Each environment of the building is equipped with a "Quick Response" (QR) code [21], having a twofold use: (i) it univocally encodes the room and (ii) it enables to locate the user position in the environment. The environmental setup of the system requires the user to direct the camera towards the QR code by pointing a viewfinder visualized on this/her camera preview. The obtained resolution of the room

 ISBN: 978-1-61208-100-7

marker enables us to deduct the shooting user-QR distance. In addition, the shooting angle, obtainable by comparing the shut QR side dimensions, allows us to determine more precisely the user position in the room. The devices also communicate the current state of the Wi-Fi signals from the various access points to the central server. Thus, it is possible to deduce each position variation by considering the azimuth variation, by integrating the detected acceleration variation and by interpolating the new Wi-Fi [18] signal, i.e. the strength of each access point carrier, of a user with his/her previous configurations and with those of the others.

B. The augmented interfaces

The environmental setup feature and the state variation of the device permit to reconstruct the user position. The augmented reality is corrected respect to the user orientation in the room obtaining the invariance to the user position for the projected objects in such a way that all users are able to find the information connected to the same physical place. This mechanism enables to associate each augmented information stream to a concrete area of the real environment.

As an example, Figure 1 shows a portion of the augmented space of the Software Engineering Lab. The information concerning the location is depicted in the lower left-hand part of Figure 1, while information on the Software School and on the events concerning the lab is depicted on the left and on the right of this figure, respectively. The user can access the information using the interaction styles proposed in the next sub-section. Let us note that in the right-hand lower corner of the screen depicts the position of the augmented areas inside the considered room [11]. This feature can be disabled.

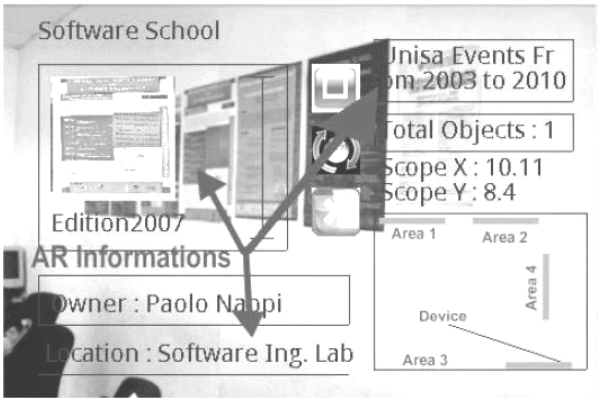


Figure 1. A screenshot of an augmented informative space

The interface in Figure 2 also displays the information concerning the owner of the area. A touch action on this area enacts the displaying of more details on the owner, if the user has the permissions, see Figure 2.

Selecting an icon in the lower part of Figure 2, the user communicates with the owner activating a telephone call or by email, he/she can see the owner web site or, using Google Maps find his/her position on the hearth. This interface allows the user a quick passage from the AR modality to other applications.

Another example of augmented area is shown in Figure 3, where the list of the users working in a given room is associated to their office door, on the external side. Thus, it is possible to verify the user state, if they are or not in the room and to leave them a message that they will receive when they enter their office.

Figure 2. Contact information of an AR area owner

A "Notes" area is available in each room registered with the system. This area represents a communication space related to the room and its occupants, displaying short messages painted directly on screen with fingers and stored as images, named Sticky Notes. Figure 4 depicts this augmented space populated by the users of a room concerning common interest facts. We noticed that the adoption of a paper background for messages provides a more realistic appearance.

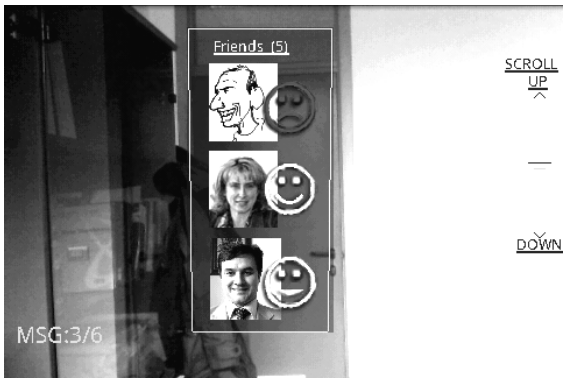


Figure 3. A Room Users' area

In addition to the "Room Users" and "Notes" areas, the system supports content sharing as follows: users can create areas to share documents with their colleagues, as an example to support group work. Each area can contain several contents: Sticky Notes, Text Notes and any other kind of files (including multimedia content) supported by the devices.

 ISBN: 978-1-61208-100-7

The creation of a new area is activated through a menu choice: the system exposes to user a viewfinder and a button, to localize the new area position. In addition, during the creation phase, the user is required to provide the area name. When an area is no longer used, its contents can be saved on the supporting web site for further consultation.

C. *The new interaction style*

SmartBuilding offers two different user interaction styles: it is possible to directly manipulate the AR contents or use an indirect interaction style, based on SDK list selection. In the former case, the augmented areas are visually paginated.

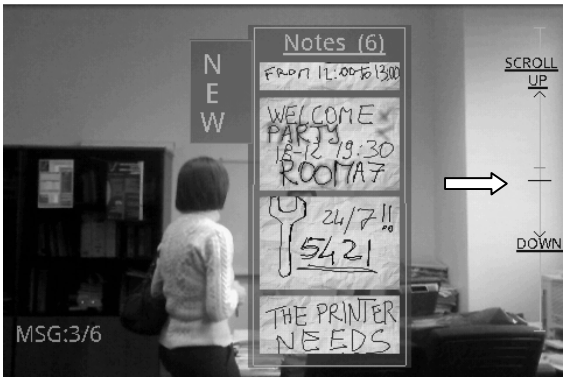


Figure 4. The Notes Area associated to the room white-board.

Because the users' movements are quite reduced when they hold the phone, there was the need of simplifying the user interaction. In particular, when the objects to show in the areas exceed the available size, a pagination mechanism is required. In the right-hand sides of Figures 3 and 4, the proposed pagination control is depicted. It is represented by a vertical segment and by a shorter horizontal segment, indicated by an arrow in Figure 4.

The main characteristic of this mechanism is that a simple gesture enables the user to change the visualized objects: the user can request the objects located in the next page of the selected area by tilting upward his/her device. The symmetrical gesture takes his/her navigation backwards. Exploiting the phone orientation sensor, it is possible to detect the Pitch movements of the device for controlling the pagination system. When the user changes the pitch direction, an analogous movement is induced on the paginated objects. In this way, the user perceives the objects to move according to the inclination given to the device. The area shape is also prospectively deformed according to the phone inclination.

Up and down movements cause the segment to accordingly move. The scrolling mechanism is activated only when the device inclination exceeds the thresholds represented by the "SCROLL UP" and "DOWN" markers.

When a user clicks on an icon, the contextualized action is fired. As an example, in the case of "Notes" and user defined "Document" areas, the click on an object causes the corresponding editing application to be launched. When the user needs to create a new note or document, he/she is required to select a distribution list choosing among the users of the system. The creation of these objects is directly supported at interface level by the button "NEW", as shown in Figure 4 in the case of "Notes".

If the indirect manipulation is preferred, the classical list mechanism, exposed by the Android SDK, is adopted to drive the user action in choosing the objects. When the user touches an area, a scrolling list is presented.

In addition to visual feedback, user involvement in the proposed experience may be enhanced by audio and haptic feedbacks. According to Henrysson et al. [8], we adopt the device speaker and its vibration to add multi sensorial feedbacks to user actions.

IV. EVALUATION

To assess the usability of the proposed system we carried out an evaluation study. In particular, following the traditional approach proposed in [11], we analyzed the user reactions to the functionalities provided by SmartBuilding.

To evaluate a context-aware AR system it is also important to consider that AR representations combine rendered graphics with the real world environment and require a specific type of interaction among virtual artifacts and the real world.

For example, user localization can be difficult if the user moves too fast, and, therefore, the system has problems in showing the appropriate contents in the appropriate location. Thus, we also evaluate the usability of the system following the directions proposed in [6], adapting it to the case of mobile phones.

In the following, we illustrate the techniques used in these evaluation studies and we present the obtained results.

A. *Preparation and User tasks*

After a short introductory session on using the device (mainly focused on menu, back button and SDK menus) and the proposed system, the subjects, provided with the appropriate paper documentation, were required to accomplish the following tasks:

- Task 1 – Users had to leave a text note on the Room Area on a door of the building.
- Task 2 – Users had to read and comment the event information in the Software Engineering lab.
- Task 3 – Collaboration. The subjects were structured in groups; each group owns a virtual area. Each group performs a simple collaborative session in which the group members upload on the Group Wall material concerning a lecture they have taken part (1.5 hour). Each group comments and votes the contribution of their peers of the other groups (1 hour).

At the end of each task, we collected feedback from users about their experience with SmartBuilding by submitting them a task evaluation questionnaire. We followed the template After Scenario Questionnaire (ASQ) proposed by IBM [11]. It consists of three questions aiming at determining user satisfaction concerning the task completion, evaluating their satisfaction regarding the ease of

 ISBN: 978-1-61208-100-7

completing, the time taken and the support information available during the task. It is designed using a 7-point Likert scale anchored at *1* by Strongly Agree and at *7* by Strongly Disagree.

Once the tasks were accomplished, the participants filled the Computer System Usability Questionnaire (CSUQ) [11] consisting in *19* questions evaluating user overall system usability focused on *3* subscales: System Usefulness, Information Quality and Interface Quality. Also in this case a 7-point Likert scale anchored at *1* by Strongly Agree and at *7* by Strongly Disagree has been adopted, but each answer contains an open "Comment" space to collect deeper details about user impressions.

To integrate the evaluation of mobile augmented characteristics, we added to CSUQ questionnaire the additional questions proposed in [11], concerning the user perception of: System Lag, Image Disparity, Resolution, Rendering Quality, Maneuverability and Environmental Conditions. We also added some specific question tailored on the nature of our system. To evaluate in detail some aspects of the proposed interface, we formulated the following additional questions scored on a Likert scale anchored from *1*, Very good to *7*, Poor:

1. The horizontal movement of augmented objects is …
2. The vertical scrolling control is …
3. The vertical inclination feedback is …
4. The environment lighting affects my performance.

In particular, we added question 4 to investigate if in case of a system mixing on the same screen an environmental preview with synthetic objects, the room lighting may disturb the user sight.

B. Participants

Subjects of the evaluation were eighteen students attending to the fundamentals of Computer Science course of the Environment Evaluation and Control program (University of Salerno).

A user profile questionnaire has been proposed to the participants before the evaluation started. Ten participants have a good experience in the usage of mobile devices and no more than 30% of the sample had good computer skills.

The participants were located in a didactic laboratory.

The system was setup localizing the "Notes" area on the white-board of the laboratory, the "Room" area on the door of the course teacher and three separate group areas have been created in the lab for Task 3.

C. Results

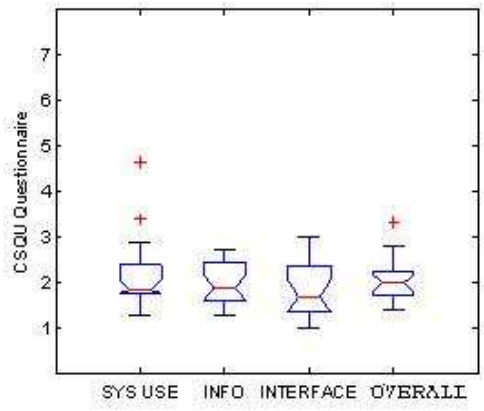
Analyzing the result provided by ASQ questionnaires concerning Task 1 (Avg. score 3.33), Task 2 (Avg. score 3.11) and Task 3 (Avg. 2.66) we noticed a diffused user satisfaction. In particular, subjects preferred Task 3, probably for its collaborative nature.

In general, we noticed that the worst perception about the easiness of completing the tasks has been signaled by less computer skilled users and by less mobile expert ones. We separated the CSUQ questionnaire results from the Augmented Mobile specific ones, since the former is suitable to determine the overall system usability, while the latter provides a specific evaluation of usability aspects concerning the proposed augmented reality interface. Figure 5 reports the CSUQ results aggregated in three categories.

The participants diffusely perceived the system as useful (average of SYS USE was 1.83). Analyzing the single question scores, it was evident that the great part of the subject sample felts confident to be able to accomplish the assigned tasks in an effective, quick and comfortable way. Just two subjects, not particularly technical skilled, expressed negative judgments.

The system is also evaluated in terms of the quality of the information provided and robustness (INFO factor in Figure 5). Also in this case the evaluation is positive. A specific question group of CSUQ questionnaire is devoted to evaluate the interface quality (INTERFACE). In particular, two specific questions required the users to score how the interface is pleasant and how much they liked it. Analyzing the individual answers, we noticed that, also the number of subjects not liking the interface was limited.

The OVERALL usability factor provides the overall system rating and is defined by the full set of items of the three sub-factors. As depicted in Figure 5, this factor reaches a satisfying score of 2.



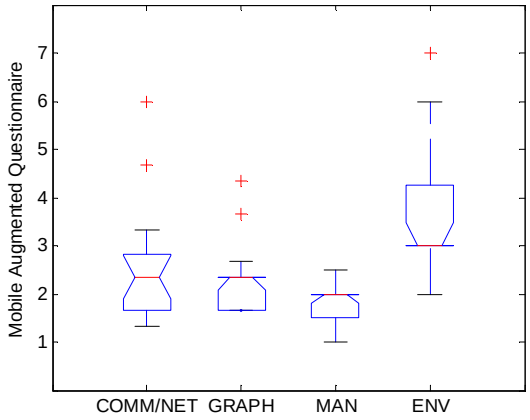
Legenda:
SYS USE="system usefulness", INFO="information quality", INTERFACE="interface quality", OVERALL="overall usability score"

Figure 5. The CSUQ questionnaire results.

Concerning the Mobile Augmented Questionnaire, we aggregated the various questions proposed in [6] with the four additional ones proposed in Section 4.1, in the three categories described in the Legenda of Figure 6. Also in the case of Mobile Augmented evaluation, the user perception was positive. In particular, best results were obtained by the "GRAPH" category that evaluates the perceived graphical resolution and rendering qualities. Image disparity is a critical factor in augmented reality interfaces. Indeed, if the AR content reproduced on the camera is offset from the real world view, the user can be disoriented. User impressions on the proposed augmented reality interface were good for the most of subjects. The system lag (COMM/NET factor) did not affect the perceived system quality and did not impact on

 ISBN: 978-1-61208-100-7

the maneuverability (MAN), which is the movement of the user is not limited to match the augmented information.



Legenda:
COMM/NET="system lag", GRAPH="image disparity, resolution and rendering quality", MAN="maneuverability", ENV="environmental conditions"

Figure 6. The Mobile Augmented questionnaire results

Analyzing the "environmental conditions" questionnaire result category, it was also evident that the system is affected by changes in the environmental lighting. Indeed, we observed that, the proposed interface is suitable for typical working room lighting.

V. CONCLUSION

In this paper, we have presented SmartBuilding, a People-to-People-to-Geographical-Places system aiming at creating an informative space surrounding the user where contents are created and displayed using Augmented Reality interfaces. The SmartBuilding Augmented Reality interface tries to improve the user interaction providing pagination mechanisms based on the device sensors.

The results of the usability evaluation of the proposed system have been positive. The evaluation revealed some problems with the environmental lighting, thus, we are going to develop new interfaces suitable for different lighting conditions (i.e. home working rooms or laboratories). At the present, we are investigating how to adopt the proposed technology to define collaborative methodologies to support work group and learning. Future work will also be devoted to make SmartBuilding, developed only for Android platform, portable on different mobile devices.

REFERENCES

[1] ARtoolkit, http://www.hitl.washington.edu/artoolkit/. Retrieved on May 2010.

[2] P. Bahl and V.N. Padmanabhan, "RADAR: An In-Building RF-Based User Location and Tracking System", *In Proc. of IEEE INFOCOM 2002*, vol. 2, pp. 775–784, 2002.

[3] T. Bemers-Lee, "Berners-Lee on the readlwrlte web", *BBC*, August 9, 2005.

[4] J. Burrell and G. Gay, "E-Graffiti: Evaluating real-world use of a context-aware system", *Interacting with Computers*, vol. 14, n. 4, pp. 301–312, 2002.

[5] G. Fitzmaurice, "Situated Information Spaces and Spatially Aware Palmtop Computers", *Communications of the ACM*, Vol. 36, no. 7, pp.39-49, July 1993.

[6] D. Haniff and C. Baber, "User Evaluation of Augmented Reality Systems", *Proc. of the Seventh International Conference on Information Visualization (IV'03)*, 2003

[7] Gartner, "Gartner Identifies the Top 10 Consumer Mobile Applications for 2012", http://www.gartner.com/it/page.jsp?id=1230413. Retrieved on May 2010.

[8] A. Henrysson, M. Billinghurst, M. Olilla, "Face to Face Collaborative AR on Mobile Phones", 2005. *Proc. of Fourth IEEE and ACM International Symposium on Symposium on Mixed and Augmented Reality*, pp. 80 - 89 , 2005.

[9] H. Hile and G. Borriello, "Positioning and Orientation in Indoor Environments Using Camera Phones", *Computer Graphics and Applications*, IEEE Volume: 28, Issue: 4, pp: 32-39, 2008.

[10] Q. Jones and S.A. Grandhi, "P3 Systems: Putting the Place Back into Social Networks", *IEEE Internet Computing*, 2005, pp. 38-46, 2005.

[11] J. Lewis, "IBM Computer Usability Satisfaction Questionnaires: Psychometric Evaluation and Instructions for Use", *International Journal of Human-Computer Interaction, vol.*7 no. 1, pp. 57-78, 1995.

[12] "Labirynth Light", retrieved on May 2010 from http://www.android.com/market/free.html#app=labyrinth. Retrieved on May 2010.

[13] "Layar", http://layar.eu/. Retrieved on May 2010.

[14] "Notescape", Microsoft Research, http://research.microsoft.com/en-us/projects/ncci/default.aspx. Retrieved on May 2010.

[15] S. Oh and W. Woo, "CAMAR: Context-aware Mobile Augmented Reality in Smart Space,"" *In Proc. of IWUVR 2009*, pp. 48-51, 2009.

[16] P. Persson, P. Fagerberg, "GeoNotes: A real-use study of a public location-aware community system", *Technical Report SICS-T–2002/27-SE, SICS, University of Goteborg*, Sweden, 2002.

[17] C. Savarese, J. Rabaey, K. Langendoen, "Robust Positioning Algorithms for Distributed Ad-Hoc Wireless Sensor Networks", *In Proc. of the General Track: 2002 USENIX Annual Technical Conference*, pp. 317–327, 2002.

[18] A. Savidis, M. Zidianakis, N. Kazepis, S. Dubulakis, D. Gramenos, and C. Stephanidis, "An Integrated Platform for the Management of Mobile Location-aware Information Systems", *In Proc. of Pervasive 2008*. Sydney, Australia, pp. 128–145, 2008.

[19] S. Singh, A. David Cheok, G. Loong Ng, F. Farbiz, "Augmented Reality Post-It System", *Proc. in Advances in Computer Entertainment Technology*, p. 359, 2004.

[20] N. A. Streitz, P. Tandler, C. Müller-Tomfelde, S. Konomi, "Roomware: Toward the Next Generation of Human-Computer Interaction Based on an Integrated Design of Real and Virtual Worlds", In J. A. Carroll (Ed.): *Human-Computer Interaction in the New Millennium*, Addison Wesley, pp. 553–578, 2001.

[21] Wikipedia QR-code, retrieved on March 2010 from http://en.wikipedia.org/wiki/QR_Code, Retrieved on May 2010.

[22] F. Zhou, H. Been-Lirn Duh, M. Billinghurst, "Trends in augmented reality tracking, interaction and display: A review of ten years of ISMAR", Proc. of the 7th IEEE/ACM International Symposium on Mixed and Augmented Reality,pp. 193-202, 2008.

 ISBN: 978-1-61208-100-7

Network Architectures for Ubiquitous Home Services

Warodom WERAPUN, Julien FASSON, Beatrice PAILLASSA

University of Toulouse, IRIT Laboratory, INP – ENSEEIHT

email: {warodom.werapun, beatrice.paillassa, julien.fasson}@enseeiht.fr

***Abstract*— The on-going growth of connectivity has brought new opportunities for Home Network; Home network will soon be a place of a large amount of services, from the gadget to the home control. In order to provide and render these services, operator should propose a framework for supporting the deployment of new services. This paper focuses on home services, proposing an overview of potential service architectures. Then a photo sharing services validates through implementation the concepts of Home Services and an analysis of architecture complexity is proposed to conclude this work.**

***Keywords*— home network, network architecture, home services, P2P, IMS, SIP.**

I. INTRODUCTION

With the rapid growth of the Internet, more and more users have an Internet access at home. This connectivity is rendered by a set top box proposing a set of services. However, most of these services are simple and static (mainly triple play) and are only managed by the network operator.

In the same time, web evolution has lead to miscellaneous services like photo sharing (picasaweb), video sharing services (youtube), social networks (facebook), etc. Such services offer more interactivity and can be directly managed by users. Nevertheless, in some cases service providers become owner of personnel data, inducing an issue of privacy for user and an issue of content right for providers. Also, the management by operator of home service enabling a local management and storage of service content solves the issue of privacy that user encounters with web services. Indeed service providers become owner of any piece of information you share through the service. Home service may also simplify the issue of copyright that provider encounters with illegal piece of information. However the responsibility of service provider may be engaged depending on way of referencing the content, as for the P2P trackers for bittorrent.

The challenge is to merge the dynamic web services at the set top box so as to propose a direct management of their services to users through their home network. As we aim at integrating service at home, we need a suitable network architecture to support service deployment and data flow.

This paper introduces home network concept, their services and their needs. Then convenient network architectures for home services are proposed. A simple service is implemented to illustrate our deployment of home services. Eventually an analysis of network architectures concludes this paper.

II. HOME NETWORK AND SERVICES

A. Context

Home Network (HN) is a small network which connects all home terminal devices together (Figure 1). Deploying services between on the HN will bring a lot of possibilities and new service uses (e.g., view photos from mobile phone on a large television screen; remotely control an air condition from any ubiquitous terminal devices, etc.). Since through the HN a user can access private resources and command all connected terminal devices remotely, the network must be secure.

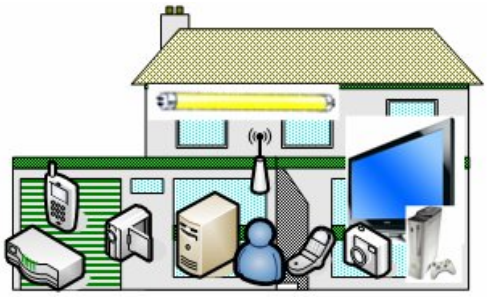


Figure 1: Home network example

Home Services (HS) are a set of miscellaneous ubiquitous services that operate at home. A same user may access to his services through miscellaneous types of access networks from a lot of devices, especially with the growth of portable technologies. In addition, these ubiquitous home services can be deployed from home by a network operator or directly by user. They can be controlled both locally and remotely. (e.g., change radio channel from personnel computer at home, check refrigerator information from mobile home remotely).

There are different types of HS. A HS can be static or dynamic. A static service is managed by the operator (e.g., TV service) and home user can just use it since they are installed by the network operator. Dynamic services are controlled services that the service owner is able to manage or change without interrupting the system (e.g., multimedia

 ISBN: 978-1-61208-100-7

sharing service). Finally HS may be intra-home service, home to home services or community ones.

B. Needs

Supporting Home Services requires allowing users to remotely connect to their HN. HS are dealing with many home equipment devices which suppose to be acquiring, viewing and managing digital content with any ubiquitous devices from any location. Personal home content and control need privacy and rightful access. These conditions raise the need of security. Moreover in order to offer home to home services like content based services, HN must be able to connect to other HN though their home gateway. As a result, HS and HN must have effective authentication and authorization mechanisms.

Since lots of HSs are expected to be deployed, they should be developer-friendly while their exchanges, their installations and their uses should be easy between HNs.

Finally, as resources and services are numerous, they induce a lot of content to share. To locate the data and/or resources it needs a service of indexation. The service of indexation directly impacts the service architectures. In this work two kinds of management are considered:

- The centralized service index: located on one single server. User can search the service or data by requesting the central index server and then go directly to service or data owner at his home.
- The distributed service index: divided in several parts. Each part is located in one or several servers. Users can search the service or data by asking index servers that depends on searching algorithm and then go directly to service or data owner at his home.

III. PROPOSITION OF NETWORK ARCHITECTURES FOR HOME SERVICES

A. Main Architectures

The object of network architecture is to delivery home services to users. As the users may be local, remote or visitor, the architecture has to manage (as transparently as possible) user accesses to their home services. Furthermore, by considering the community framework, the architecture must in addition manage the user data localization.

Basically of achieve service architectures, there are IMS (IP Multimedia Subsystem) [1], P2P (Peer to Peer) [2], VPN (Virtual Private Network) [3] and Web architecture.

IMS is defined as an architectural framework created for the purpose of delivering IP multimedia services to end-users. It supports IP Multimedia sessions, quality of service (QoS) requirement, interworking with the Internet and the circuit switched network, roaming as well as the ability for operators to have a strong control on the services of users. IMS uses SIP (Session Initiation Protocol) [4] as signaling. SIP is the text based signalling protocol to sets up multimedia sessions between endpoints. These sessions may be text, game, voice, video or a combination of these. It is a centralized scheme as the services and users are managed by a central functional entity.

P2P is a relation of connected devices which have equivalent privileges and which share their resources and services together. It is a distributed architecture. In addition, there are several types of P2P such as pure P2P, hybrid P2P and DHT P2P [5] etc.

VPN provides the secure tunneling to establish sessions. There are a lot of interconnecting scenarios in case of using the VPN technology. Users can directly make a VPN tunnel to a server with their home gateways. This aims indeed to create a connection with a VPN server that is able to access to the gateway for achieving services or resources as previously described since some services have policies that clients have to stay in the same network with the server.

Concerning web architecture, it is a client-server based. All data and indexes are stored at one central server. All users achieve them from a central web server.

The choice of an architecture is depended on many factors, especially the performance, the security and the scalability. For a global point of view, considering the performance aspect and scalability aspects, centralized architecture, as IMS, seems to be adapted to home to home service, while distributed P2P technique would be convenient for community services. On the contrary, when tacking account the security aspect, centralized architecture seems preferable whatever the type of service.

Next paragraph details more precisely IMS and P2P architecture on a service example.

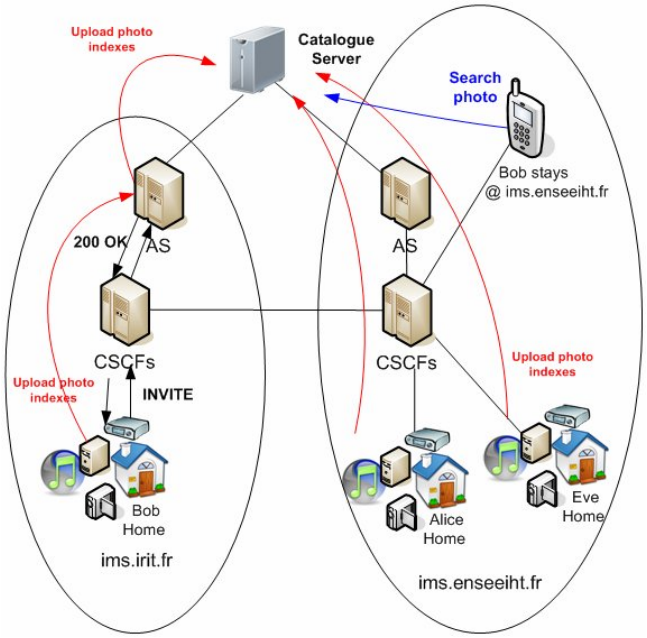


Figure 2 : Photo sharing with IMS Architecture

B. IMS Architecture

The architecture is illustrated on a service scenario. The service is a photo sharing service with a central catalogue server on IMS architecture. The catalogue server is defined to manage resource addresses which are published by clients.

IMS scenario: In Figure 2, Bob, with his home network, registers himself at ims.irit.fr domain. Alice and Eve stay at ims.enseeiht.fr domain. They can upload their photo lists to the catalogue server. The catalogue server is attached with the Application Server (AS). AS is connected with Call Session Control Functions (CSCFs) and it will be trigged by matching the IMS signalling which is defined in Home Subscribe Server (HSS). When Bob stays outside his home, if he would like to search some photos, he just looks in the catalogue server to acquire preferred photo addresses via CSCFs. Then, he can directly download from his friends following retrieved addresses. CSCFs will responsible for session establishment which includes authentication and authorization to the catalogue server.

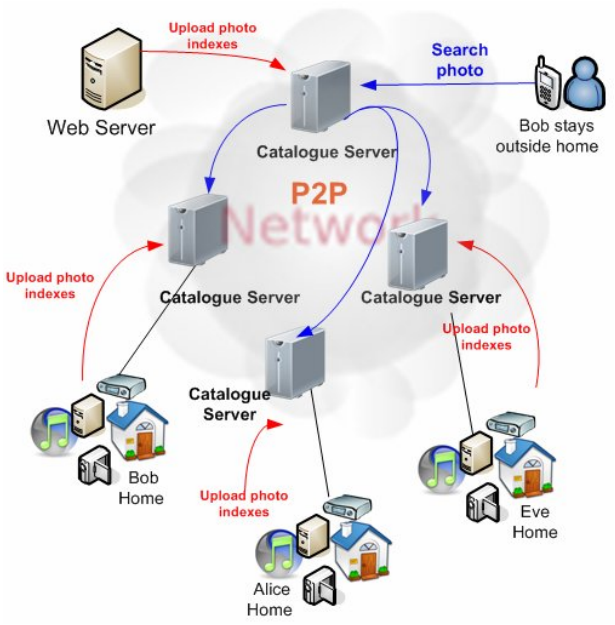


Figure 3: Photo sharing with P2P Architecture

IMS security aspect: IMS standard proposes a security architecture that uses several security protocols such as IMS Authentication Key Agreement (AKA) [6] between User Equipment (UE) and IMS network via P-CSCF, IP Security (IP Sec) [7] between UE and P-CSCF, Diameter [8] between HSS and I/S-CSCF. IMS uses AKA for authentication and IPSec for confidential and integrity. IMS provides strong security mechanisms that suppose to be efficiently secured the platform. Unfortunately, most of the real deployment is not rigorously clung all IMS security standards, e.g., some ubiquitous devices do not support IPv6 and/or IPSec as mandatory in IPv6 [9]. Moreover, only a few IP phones supported AKA. Due to lack of an IMS Subscriber Identity Module (ISIM) in laptops, they use MD5 digestion authentication instead. In addition, because of IMS security architecture implementation is truly complex. As a result, it had led to simplify security mechanism and they also lead to vulnerabilities.

C. Centralized P2P Architecture

The P2P network considered is a centralized P2P. As in IMS, signaling may also be SIP [9].

P2P scenario: The photos sharing service scenario is quite similar to the IMS architecture. There are publishing, searching and retrieving. IMS takes all indexes into a single central catalogue server; instead P2P divides indexes in to several parts and leaves them to several catalogue servers as described in Figure 3. Connected user in the network can share by upload the photo indexes to a catalogue server. Then, another user (e.g., Bob) can search and directly download the photo that he wants from the photo owner.

Centralized P2P security aspect: Many kinds of security architectures that are depended on the efficiency level of protection can be integrated. For minimum level, it can be assumed that all peers are trusted. To increase security, mechanisms can be added as a centralized AAA server, Kerberos [11], a server for generating session tickets to clients, proxy server authentication, peer signatures in the centralized P2P and public/private key cryptography. In addition, it could be used with the challenge/response protocol to authenticate each others. However, lots of certify mechanisms lead to decrease system performances. This should be considered before to decide to apply the security mechanism.

IV. EXPERIMENTAL

Java and JAIN-SIP [12] library have been used to implement the photo sharing service on P2P network architecture with SIP signalling. It consists of 3 phases: publishing, searching and retrieving.

The main components of the service (Figure 4) are:

- Global/local manager: manage contact list for server/client
- Contact List: friend address list, Data & Index: resource addresses
- SIP UA (User Agent). SIP UA is used to be SIP interfaces to other users. It is used to establish the session for specified applications.

Peer stores its data and indexes. Peer publishes its sharing indexes to the catalogue manager. Catalogue manager maintains sharing indexes. Contact list stores a list of peer's friends which is managed by local manager.

 ISBN: 978-1-61208-100-7

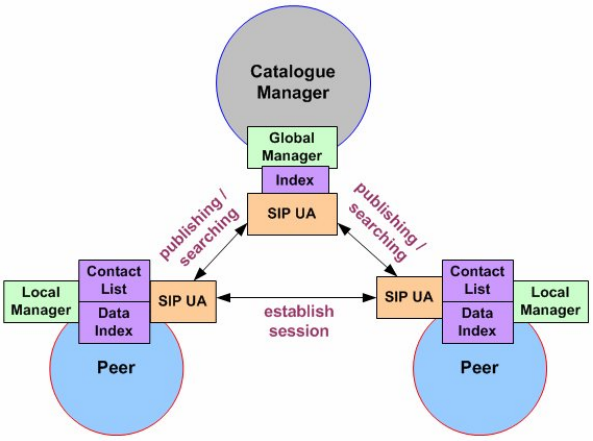


Figure 4: Photo sharing design component

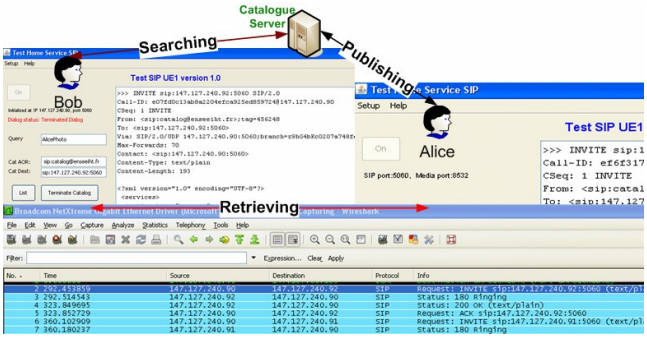


Figure 5: Photo sharing screenshot

Figure 5 shows the implementation screenshot. This application is based on a SIP signalling to communicate with the connected users for session establishment following SIP standard. The specific service communication protocol (e.g., photo sharing service) is defined by attaching xml elements with the SIP body message (MIME type) [13] in text/plain format. Moreover, service communication protocol is mapped with the POJO (Plain Old Java Object) for easier managing. We had tested that our application can established SIP session with the catalogue server and can directly share photo between friends successfully.

V. ANALYSIS AND COMPARISON

The study focuses on the signaling induced by registration, publishing and retrieving, for different architecture and security mechanisms.

A. Client signalling

A first point is to analyze client signalling: counting the procedure weight by number of messages (e.g., sending and receiving by the UE) and the size of these messages. Because the kind of messages is the same, counting them is sufficient.

Let analyze in the client side of the IMS and the SIP central architectures as indicated in Figure 6. There are IMS Client, SIP Client1, SIP Client2 and SIP Client3. IMS Client is an IMS UE that works following IMS standard. The SIP clients differ from their authentication mechanisms. More precisely, when IMS client is connected with different network operators, IPSec session is used to secure the communication (it could be both transport or tunnel mode), however, SIP clients are free to define security protocol, and then it can be applied with lighter security (SIP Client1) until stronger security (SIP Client3) .

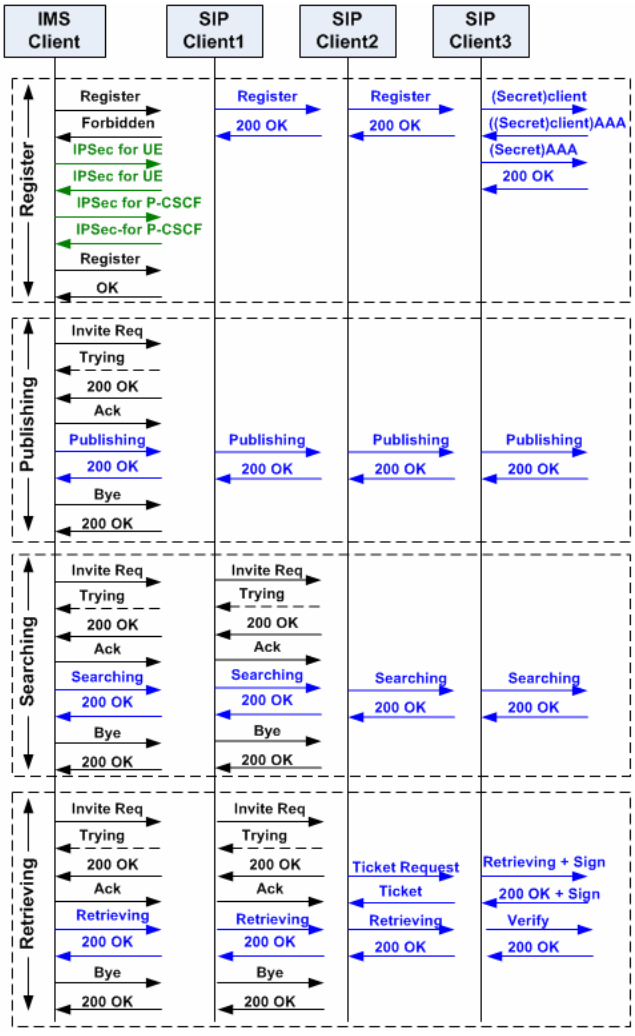


Figure 6: Client signalling

About the signalling is presented in the Figure 6. It is connected to the way to manage content sharing with SIP. A first solution is to have sessions with the catalogue for publishing and searching, and a session with the sharing client for retrieving (as illustrated in Figure 6 with SIP client 1). In this case, the client signalling is similar to the IMS case except the authentication part. However, we can

define specific SIP signalling at least for publishing and searching that does not require a session (as indicated in Figure 6 with SIP client 2 and 3).

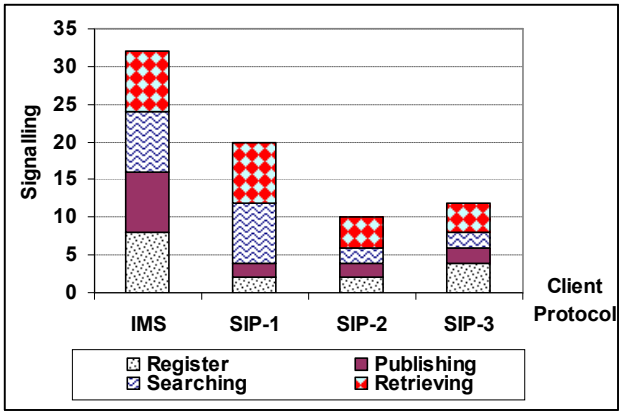


Figure 7: All client signalling summary

Figure 7 shows summary of client exchange signalling for each phases and theirs total. We can see that IMS client handles more procedures than SIP clients. However, IMS solution might be optimized because its standard includes a lot of SIP signalling. In addition, it also proposes a native security mechanism. For the application signalling, photo sharing service needs 2 signalling (request/response), SIP-2 and SIP-3 embed an application query in SIP signalling. This analysis can be extended to any kind of services that have request/response flow like the photo sharing one.

B. Signalling summary

The first evaluation gives a global point of view on the client side but it does not reveal the complexity of the whole architecture. The second point is to evaluate the overhead of signalling over the whole architectures. Thus, counting exchanged messages between all nodes allows us to have an overview of global behaviour.

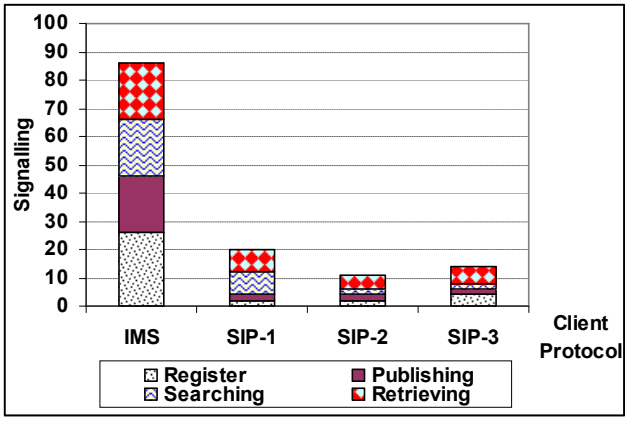


Figure 8: Whole signalling in all network architectures for one searching

Figure 8 shows whole signalling in all considered network architectures for one searching. It focuses on all signalling that is related to a downloader who tries to get the content from a sharer.

IMS has lots of signalling because its standard requires several components (e.g., CSCFs, HSS and ASs) to create and establish the session. In contrast, centralized SIP has defined only a centralized index component. Then, we can build the system with our preferred security mechanism which depends on the required security level and the system performance.

In SIP cases, we propose: 1) Password authentication (SIP-1): login one time to the network, 2) Kerberos ticket (SIP-2): use Kerberos server to issue ticket for communicating between peers, and 3) signature authentication (SIP-3): peers have to sign all signalling and verify with the authentication server. In the fact, there are more security mechanisms which are possible to use. However, these solutions induce different security levels and network performance.

IMS is greedier signalling mainly in whole architecture. It needs an infrastructure, processes in each entity and very complex in the register phase. Register phase is occurred when users firstly connect to the system or move to another location. However, searching and retrieving are considered to be frequently occurred than register phase. These are more dynamic and significant signalling to consider.

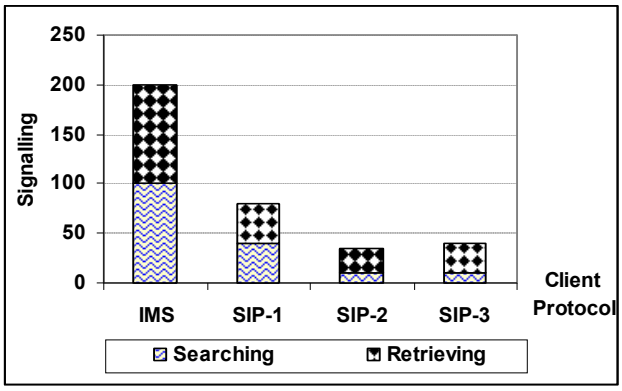


Figure 9: Signalling summary for five searching and retrieving

Figure 9 shows signalling summary for five searching and retrieving. We can see that SIP-3 still has less signalling than standard IMS. Moreover, when we try to increase searching times, the downloader searches the content in the network, signalling in IMS and SIP plus security are increasing undoubtedly. On the other hand, the ratio is decreasing since the major difference is occurred at registration phase. Moreover, when we added security mechanism to SIP central architecture, signalling is increasing. This will lead to decrease network performance. However, it has to consider by the system manager that how much for the system security is required with this tread-off.

 ISBN: 978-1-61208-100-7

VI. CONCLUSION AND FUTURE WORKS

In this work, we presented home services and network architectures for the future delivery of dynamic services to home users. Next, we propose service classification, and network architectures for home services. Based on existing SIP protocol and IMS architecture, this paper exposes a new SIP based framework with security that is compared with IMS solution. We show exchanging message comparison between IMS and centralized SIP architectures with interested security mechanisms and we can see that IMS architecture is more complicated than centralized SIP.

As we previously described for the SIP matter, we have to add security protocol, authenticate and authorize users each time with all peers that they are connected, client application needs to aware security with the SIP application server. In addition, to integrate security protocol in P2P is also interesting since P2P gives more benefit e.g., scalability and availability but it also creates more overhead and complexity especially in security management.

We will attempt to focus more precisely in community networks to do service sharing and build the system by using several P2P architectures (e.g., hybrid P2P, DHT P2P). This study is our future step towards numerous works. A first element might be a real implementation of sharing services. This implementation could be deployed on different types of P2P architectures so as to compare their performances, their relevance and their security aspects. Other services could also be deployed on the more relevant architecture. Finally, the IMS architecture could also provide a support to P2P services. A coupling of these 2 architectures might be an interesting study too.

ACKNOWLEDGMENT

The research is supported by European Celtic Project under Feel@Home project.

REFERENCES

[1] 3GPP, “IP Multimedia Subsystem (IMS),” TS 23.228, Release 8, Version 8.7.0, Dec 08

[2] G. Camarillo, Peer-to-Peer (P2P) Architecture: Definition, Taxonomies, Examples, and Applicability, RFC 5694, IETF Network Working Group, Nov 2009

[3] E. Rosen and Y. Rekhter, BGP/MPLS VPNs, RFC 2547, IETF Network Working Group, July 1999

[4] J. Rosenberg, H. Schulzrinne, G. Camarillo, A. Johnston, J. Peterson, R. Sparks, M. Handley and E. Schooler, SIP: Session Initiation Protocol, RFC 3261, IETF Network Working Group, 2002

[5] I. Stoica, R. Morris, D. Liben-Nowell, D. R. Karger, M. F. Kaashoek, F. Dabek, and H. Balakrishnan, Chord: A Scalable Peer-to-Peer Lookup Protocol for Internet Applications, IEEE/ACM Transactions on Networking, Vol 11, Feb 2003

[6] A. Niemi, J. Arkko and V. Torvine, Hypertext Transfer Protocol (HTTP) Digest Authentication Using Authentication and Key Agreement (AKA), RFC 3310, Sep 2002

[7] S. Kent and R. Atkinson, Security Architecture for the Internet Protocol, RFC 2401, IETF Network Working Group, Nov 1998

[8] P. Calhoun, J. Loughney, G. Zorn and J. Arkko, Diameter Base Protocol, RFC 3588, IETF Network Working Group, Sep 2003

[9] Frank S. Park, Devdutt Patnaik, Chaitrali Amrutkar, Michael T. Hunter, A Security Evaluation of IMS Deployments, IMSAA 08, Dec 2008

[10] C. Jennings, B. Lowekamp, E. Rescorla, S. Baset and H. Schulzrinne, REsource LOcation And Discovery (RELOAD) Base Protocol, draft-ietf-p2psip-base-08, Mar 2010

[11] C. Neuman, T. Yu, S. Hartman and K. Raeburn, The Kerberos Network Authentication Service (V5), RFC 4120, IETF Network Working Group, July 2005

[12] JAIN-SIP Developer Tool, https://jain-sip.dev.java.net/ (Accessed: 4 June 2010)

[13] N. Freed and N. Borenstein, Multipurpose Internet Mail Extensions, RFC 2045, IETF Network Working Group, Nov 1996

 ISBN: 978-1-61208-100-7

The QoE-oriented Heterogeneous Network Selection Based on Fuzzy AHP Methodology

Dong-ming Shen
Beijing University of Posts and Telecommunications
Beijing, China
sdmjordan@gmail.com

Hui Tian
Beijing University of Posts and Telecommunications
Beijing, China
Tianhui@bupt.edu.cn

Lei Sun
Beijing University of Posts and Telecommunications
Beijing, China
buptsunlei@gmail.com

***Abstract*—The next generation of wireless communication will be characterized by heterogeneity. One of the challenges arisen is access selection among various radio access technologies. Meanwhile, the quality of experience (QoE) of user is becoming one of the most concerned topics. Analytic Hierarchy Process (AHP) has been popularly used in network selection, while this method is usually imprecise because consistency index in AHP does not accurately indicate users' perception. In order to deal with this problem as well as optimize the system performance, an effective fuzzy Analytic Hierarchy Process scheme for network selection is proposed in this paper, which takes the multiple criterions of quality of experience (QoE) into consideration. By introducing the fuzzy consistency, the performance of the proposed scheme is consistent with user preferences and experiences. The fuzzy AHP derives relative weights from consistent fuzzy comparison matrices, which eliminates both additional consistency test and modification for the comparison matrix. Simulation results are analyzed in aspects of *session quality*, *availability* and *instantaneity*, and it is indicated that the proposed scheme outperforms the traditional AHP method and load balancing oriented method.**

Keywords- QoE; heterogeneous; network selection; fuzzy AHP; consistency

I. INTRODUCTION

Motivated by the ever-increasing demand for wireless communications, the past few years has witnessed rapid development of varies of wireless networks. It is widely accepted that heterogeneity will be a prominent feature of the next-generation wireless system.

While heterogeneous networks bring multi-access benefit, new challenges also emerge as how to achieve orderly and efficient cooperation across heterogeneous radio networks. Furthermore, the mobile multimedia services are expected as the most promising killer-applications for the next generation wireless systems. As the prime criterion for quality evaluation of multimedia applications, quality of experience (QoE) becomes important to network (service) providers in order to reduce user churn and maintain, and it has been well studied in both the academia and industrial community [1-3]. For QoE, it comprises all elements of an end user's perception of using a service or product. QoE not only covers end-to-end Quality of Service (QoS) parameters such as coverage, throughput, delay, jitter, bit error rate (BER) and so on, but also contains user preference criterions such as cost, mean of score (MOS), mobility, etc. Therefore, an essential issue in heterogeneous radio environments is how to select the most appropriate network according to QoE evaluations including user preference, network capability and service characteristics.

Analytic hierarchy process (AHP) has been applied in many fields, such as network selection and satisfaction evaluation [4]. The relative importance of factors and sub-factors with respect to their parents are estimated through pair-wise comparison based on human's knowledge and experiences. In spite of its popularity, the method is often criticized for its inability to precisely represent human perception, the main reason for this imprecision lies in the *inconsistent* comparison matrices. As a result, AHP requires additional test of comparison matrix's consistency to avoid the violation of common sense that "A is more important than B, B is more important than C, however C is more important than A". However, problems still exist due to the fact that consistency index in AHP is not accurately consistent with user preference.

This paper proposes a fuzzy AHP (FAHP) scheme for network selection based on QoE evaluation in heterogeneous scenarios. In order to deal with the problem mentioned above and effectively capture the ambiguity in user requirements, fuzzy complementary matrix and fuzzy consistent matrix are introduced to relax the consistency requirement in conventional AHP. Then relative weights are deduced based on FAHP theory without consistency test and modification to the judgment matrix. According to [5], several key quality indicators (KQIs) are chosen in this paper to reflect the degree of QoE, including *availability*, *session quality*, and *instantaneity*. Meanwhile, a number of key performance indicators (KPIs) are account for each KQI as its subcategories. Therefore, the FAHP procedure will be implemented in double-layer assessments, and then gives performance ranking of all the networks.

The rest of the paper is organized as follows. The system model and framework for the proposed network selection is studied in Section II. The detail process of FAHP is presented in Section III. A scenario in heterogeneous networks and simulation results are shown in Section IV. Finally, the conclusion is given in Section V.

II. SYSTEM MODEL AND FRAMEWORK

In the typical scenario of heterogeneous radio system, several radio access technologies (RATs) are deployed and different RATs may overlap with each other. These networks are diverse in capabilities of data rates, mobility, coverage, charging mechanisms, etc. For the mobile user, all

 ISBN: 978-1-61208-100-7

available networks are denoted as the set $\Omega = \{AN_1, AN_2, \cdots AN_K\}$. A network is available means that the pilot or beacon of the network can be detected and recognized by the user.

As shown in Figure 1, the resource assessment is decomposed into hierarchical levels. Several typical KQIs are considered to comprehensively to reveal the QoE of service in AN_k, $k = 1, 2, \cdots, K$. One KQI provides a specific aspect of the service performance. Meanwhile different KPIs are listed as sub-factors with respect to the upper layer KQIs.

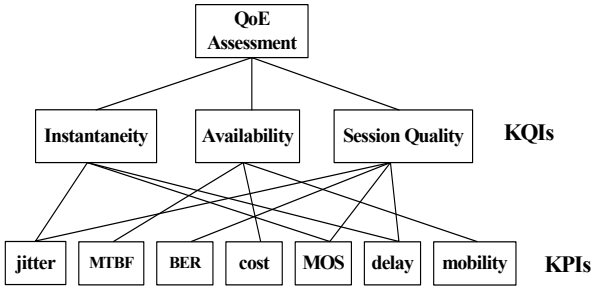


Figure 1. The hierarchy of resource assessment

A. KQIs and KPIs

Session Quality, *Availability* and *Instantaneity* are considered as the three typical KQIs in this paper. Each KQI is described as follows [5].

- **Session Quality—**It represents the collective effect of performances that mainly concern with the definition of the service. A common metric of session quality is the Mean Opinion Score (MOS). Besides, objective performance indicators are also included: delay, BER and jitter. Thus the degree of session quality is achieved by a composite of objective and subjective parameters.
- **Availability—**The service availability is expressed as a percentage of time during which the service is available and the customer has the ability to use the service. It relates to the maintainability performance during the service and the charge cost of the service. Mean time between failure (MTBF) is used as a important metric for multimedia service quality. Besides, cost-effectiveness and mobility are also considered.
- **Instantaneity**—It refers to the punctuality performance of the service. The more prompt service delivery is, the high grade of instantaneity is, especially for real time services. The instantaneity assessment should include MOS, delay and jitter.

B. Framework of network selection

Figure 2 shows the block diagrams of network selection based on QoE evaluation. The whole FAHP system is divided into three parts: sub-category estimator, weights estimator and overall-category estimator, all of which coordinate with each other to give the ranking of network alternatives.

Suppose that QoE assessment is decomposed into N aspects. Then for, $\mathbf{G}_i^k = (g_{i1}^k, g_{i2}^k, \cdots g_{iN_i}^k)$, $i = 1, 2 \cdots N$ denotes the KPI vector that the candidate AN_k is to be judged upon, where N_i KPIs are taken into account for the ith KQI. All the three parts of network selection are introduced based on AN_k as follows.

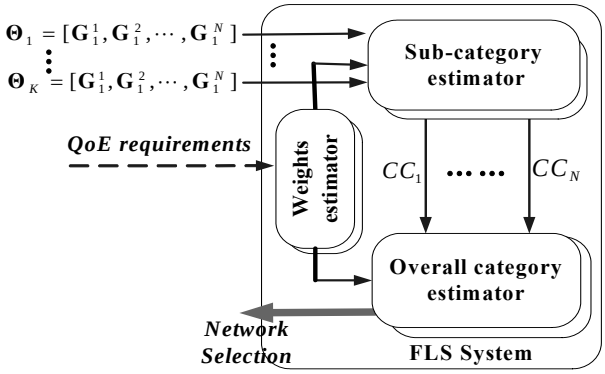


Figure 2. The block diagrams of network selection

Weights estimator is used to deduce the relative weights of KPIs and KQIs based on FAHP theory. According to user preferences described in QoE requirements, the relative weights of KPIs for ith KQI are deduced and denoted by a weight vector $W_i = (w_1, w_2, \cdots, w_{N_i})$. While $\boldsymbol{\alpha}_i = (\alpha_1, \alpha_2, \cdots, \alpha_N)$ stands for relative weights of different KQIs. The KQI weights and KPI weights are then distributed to overall-category estimator and sub-category estimator, respectively.

All KPI measures of each KQI are collected in the **sub-category estimator**. Meanwhile the desirable KPI measures and weights are obtained from QoE requirements. Fuzzy rating of each aspect is represented by a *closeness coefficient* based on the **TOPSIS** (Technique for Order Preference by Similarity to Ideal Solution). According to the concept of **TOPSIS**, the positive ideal solution (PIS) and the negative ideal solution (NIS) will be defined as $\mathbf{G}_i^+$ and $\mathbf{G}_i^-$, respectively **[**6**]**. Then the sub-category estimator calculates the distance of each alternative from PIS and NIS, which are denoted as d^+ and d^- respectively. In this paper, the distances are defined as (1) based on vector norms, since it can be proved that the vector norms satisfy the monotonic property required by TOPSIS method.

$$\begin{cases} d_{ik}^+ = \left\| \mathbf{G}_i^k - \mathbf{G}_i^+ \right\| = \left[\sum_{m=1}^{N_i} w_m{}^2 (g_{im}^k - g_{im}^+)^2 \right]^{1/2} \\ d_{ik}^- = \left\| \mathbf{G}_i^k - \mathbf{G}_i^- \right\| = \left[\sum_{m=1}^{N_i} w_m{}^2 (g_{im}^k - g_{im}^-)^2 \right]^{1/2} \end{cases} \quad (1)$$

 ISBN: 978-1-61208-100-7

Then the closeness coefficient for the *i*th aspect is defined as (2):

$$CC_i^k = \frac{d_{ik}^-}{d_{ik}^- + d_{ik}^+} \quad (2)$$

Overall-category estimator is designed to aggregate the estimation of all aspects to get the whole rank of available networks. The aggregated coefficient is defined as a *Grade of Service Index* (*GSI*)

$$GSI_k = \sum_{i=1}^{N} \alpha_i^k \cdot CC_i^k \quad (3)$$

where CC_i^k is the output of sub-category estimator for the *i*th KQI evaluation. It is worth pointing out that both KPI weights and KQI weights derived from Fuzzy AHP satisfy the consistency requirement, which will be discussed in the Section III. Therefore, according to the *GSI*, the candidate networks will be ranked and the best one can be selected.

III. DESIGN OF FUZZY AHP

To relax the consistency requirement in conventional AHP, fuzzy consistent matrix is introduced, which is well consistent with user's perception and meanwhile can capture the ambiguity in user requirements.

A. Basic concepts

Definition 2.1. [7] For the fuzzy matrix $\mathbf{R} = (r_{mn})_{N\times N}$, if $r_{mn} + r_{nm} = 1$ for any integer *m and n*, then **R** is a fuzzy complementary matrix.

Definition 2.2. [7] For the fuzzy complementary matrix $\mathbf{R} = (r_{mn})_{N\times N}$, if $r_{mn} = r_{mk} - r_{nk} + 0.5$ for any integer *m*, *n*, *k* given at random, then **R** is a fuzzy consistent matrix.

For a given user, the element r_{mn} is a fuzzy membership in that criterion *m* is more important than criterion *n* according to their contribution to the upper layer criterions. When one criterion compares to itself, it is expressed as $r_{mm} = 0.5$. According to definition 2.2, the inherent consistency of fuzzy consistent matrices can be proved by Theorem 1.

***Theorem 1*:** For any fuzzy consistent matrix **R** described in definition 2.2, **R** satisfies the consistency requirement by user perception. Suppose *m* is more important than *n*, and *n* is more important than *k*, then *m* is more important than *k*.

***Proof*:** For different criterions *m*, *n*, *k*, if *m* is more important than *n*, and *n* is more important than *k*, then we have $r_{mn} > 0.5$ and $r_{nk} > 0.5$. According to definition 2.2, $r_{mn} = r_{mk} - r_{nk} + 0.5$, then $r_{mk} = r_{mn} + r_{nk} - 0.5 > 0.5$. Hence, it is true that *m* is more important than *k*, which is consistent with user's common sense of consistency.

B. Fuzzy AHP process

Step 1: Construct pair-wise comparison matrices

It is required that the comparison matrices constructed is fuzzy consistent. As shown in Table I, comparison matrices can be obtained from pair-wise comparison, which is conducted similarly as the nine-point scale used in AHP.

TABLE I. FUZZY PAIR-WISE COMPARISON

Quantitative value	Fuzzy language
0.5	*A* is equally important as *B*.
0.6	*A* is a little important than *B*.
0.7	*A* is more important than *B*.
0.8	*A* is strongly important over *B*.
0.9	*A* is absolutely important over *B*.
0.1~0.4	If the quantitative value is *x* when *B* compares to *A*, then it is 1-*x* when *A* compares to *B*.

For KPI level, the consistent matrix is denoted as (4).

$$\mathbf{P} = (p_{mn})_{N\times N} \quad (4)$$

Step 2: Calculate relative weights

Relative importance of each criterion is derived from consistent matrices. Since element p_{mn} in **P** is the result of importance comparison between sub-factors *m* and *n*, p_{mn} is supposed to be a function of w_m and w_n. The following theorem gives the detailed function expression.

***Theorem 2*:** $\mathbf{P} = (p_{mn})_{N\times N}$ is a fuzzy complementary matrix. Then **P** is a fuzzy consistent matrix if and only if $\exists W = (w_1, w_2, \cdots w_N)^{\mathrm{T}} \in R_+^N$, $\sum_{m=1}^{N} w_m = 1$.

$$st: \quad p_{mn} = \frac{1}{\theta} \ln \frac{w_m}{w_n} + 0.5 \quad (5)$$

where $1 \le m, n \le N$, and $\theta > 0$ is an adjustable parameter.

***Proof*:** On one hand, if **P** is a fuzzy consistent matrix, then the KPI weight can be defined as

$$w_m = \frac{\exp(\frac{\theta}{N}\sum_{n=1}^{N} p_{mn})}{\sum_{m=1}^{N} \exp(\frac{\theta}{N}\sum_{n=1}^{N} p_{mn})} \quad (6)$$

Since $\sum_{m=1}^{M} w_m = 1$, by the definition of fuzzy consistent matrix

$$\begin{aligned}\frac{1}{\theta}\ln\frac{w_m}{w_n} &= \frac{1}{N}\sum_{l=1}^{N} p_{ml} - \frac{1}{N}\sum_{l=1}^{N} p_{nl} \\ &= \frac{1}{N}N(p_{mn} - 0.5) = p_{mn} - 0.5\end{aligned} \tag{7}$$

Therefore, the equation (4) is satisfied.

On the other hand, if the equation (4) is true, that is, for $\forall m, n, k \in 1, 2, \cdots N$,

$$\begin{aligned}p_{mn} &= \frac{1}{\theta}\ln\frac{w_m}{w_n} + 0.5 = (\frac{1}{\theta}\ln\frac{w_m}{w_k} + 0.5) \\ &-(\frac{1}{\theta}\ln\frac{w_n}{w_k} + 0.5) + 0.5 = p_{mk} - p_{nk} + 0.5\end{aligned} \tag{8}$$

Obviously, $\mathbf{P} = (p_{mn})_{N\times N}$ is a fuzzy consistent matrix. Hereby, the poof is completed, and the weights can be constructed as (6). Moreover, this method to construct the weights is not occasional, instead it is reasonable according to Theorem 3.

***Theorem 3*:** For any fuzzy complementary matrix $\mathbf{P} = (p_{mn})_{N\times N}$, assume that the weight vector can be calculated by solving the following constraint programming problem

$$\begin{cases}\min\ f = \sum_{m=1}^{N}\sum_{n=1}^{N}(\frac{1}{\theta}\ln\frac{w_m}{w_n} + 0.5 - p_{mn})^2 \\ s.t.\sum_{m=1}^{N} w_m = 1,\ w_m > 0,\ m = 1, 2, \cdots, N\end{cases} \tag{9}$$

where $\theta > 1$ .Then the solution is

$$w_m(\theta, \mathbf{P}) = \exp(\frac{\theta}{N}\sum_{n=1}^{N} p_{mn}) \Big/ \sum_{m=1}^{N}\exp(\frac{\theta}{N}\sum_{n=1}^{N} p_{mn}) \tag{10}$$

***Proof*:** Firstly, according to the MSE (Mean Squared Error) principle, it is reasonable to form the constraint programming problem based on Theorem 2. Considering the *Lagrange* multiplier method, problem (9) can be transformed into (10), where $\lambda \geq 0$.

$$\min\ L(w, \lambda) = \sum_{m=1}^{N}\sum_{n=1}^{N}(\frac{1}{\theta}\ln\frac{w_m}{w_n} + 0.5 - p_{mn})^2 + \lambda(\sum_{m=1}^{N} w_m - 1) \tag{11}$$

In order to get the optimal solution, assume that $\frac{\partial L(w, \lambda)}{\partial w_m} = 0$, then we have

$$2\sum_{m=1}^{N}\left[\sum_{n=1}^{N}(\frac{1}{\theta}\ln\frac{w_m}{w_n} + 0.5 - p_{mn}) + \lambda w_m\right] = 0 \tag{12}$$

It is proved that for any fuzzy consistent matrix, the sum of all its element equals to $N^2/2$, and $\sum_{m=1}^{N} w_m = 1$. Therefore,

$$\sum_{m=1}^{N}\sum_{n=1}^{N}(\frac{1}{\theta}\ln\frac{w_m}{w_n} + 0.5 - p_{mn}) = 0 \tag{13}$$

Obviously, $\lambda = 0$. Hence, the solution can be calculated from (14).

$$\begin{cases}\sum_{n=1}^{N}(\frac{1}{\theta}\ln\frac{w_m}{w_n} + 0.5 - p_{mn}) = 0 \\ \sum_{m=1}^{N} w_m = 1\end{cases} \tag{14}$$

Then (10) has been satisfied.

Step 3: Closeness coefficient calculation

When all *N*-aspect measures are collected in the sub-category estimator, the fuzzy rating of each aspect is represented by a closeness coefficient based on the TOPSIS. Before calculating the closeness coefficient, the input measures need to be normalized in two situations, larger-the-better and smaller-the-better. For $\mathbf{G}_i^k = (g_{i1}^k, g_{i2}^k, \cdots g_{iN_i}^k)$, it is normalized as (15).

$$\begin{cases}\tilde{g}_{i1}^k = \min\left\{1, \frac{g_{i1}^k}{g_{i1}^*}\right\}, \text{for larger-the-better} \\ \tilde{g}_{i1}^k = \min\left\{1, \frac{g_{i1}^*}{g_{i1}^k}\right\}, \text{for smaller-the-better}\end{cases} \tag{15}$$

where $\mathbf{G}_i^* = (g_{i1}^*, g_{i2}^*, \cdots g_{iN_i}^*)$ is the desirable KPI vector. Then, for all *i, j,* $0 \leq g_{ij}^k \leq 1$. Meanwhile, PIS and NIS can be defined as $\mathbf{G}_i^+ = (1, 1, \cdots 1)$ and $\mathbf{G}_i^- = (0, 0, \cdots 0)$.

According to the concept of TOPSIS, the closeness coefficient for each aspect is calculated as (1-2) based on relative weights.

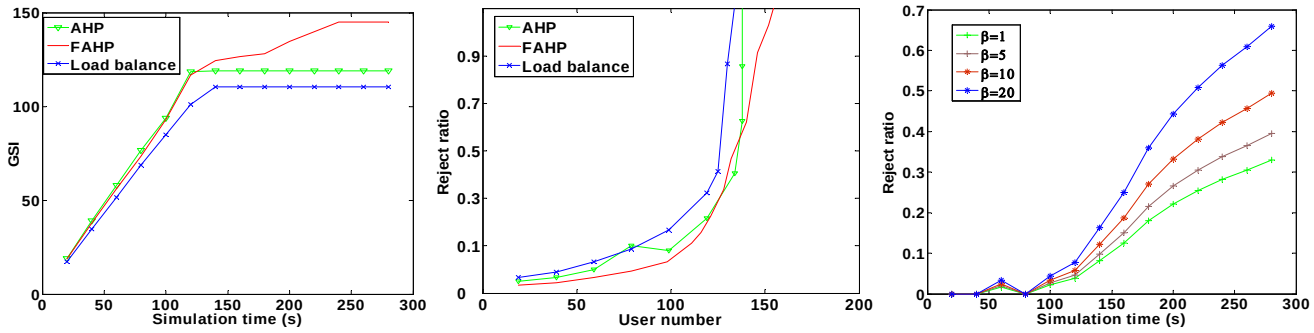
Step 4: Overall estimation and decision making

Similar to the sub-category estimation, both pair-wise comparison and relative weights construction are necessary in KQI-level estimation. With respect to different user, KQI preference may totally different. As to the KQI *comparison matrices*, it is denoted as $\mathbf{Q} = (q_{mn})_{N\times N}$, where q_{mn} describes the fuzzy membership in that the user is more care about criterion *m* than *n*. Based on comparison matrices,

 ISBN: 978-1-61208-100-7

weights of different KQIs are obtained as $\alpha_m(\beta, \mathbf{Q})$ according to (10). The GSI value of each network is obtained through (3). Referring to the performance ranking of all the networks, the best network is selected.

It is worth mentioning that both θ and β are adjustable, which reveal the user's degree of attention on the relative importance. Assuming that $\alpha_m(\beta) > \alpha_n(\beta)$, then it is



(a) System revenue (b) Reject ratio with total user number increases (c) Reject ratio vs. adjustable parameter β

Figure 3. Simulation results

obvious that $\frac{\alpha_m(\beta)}{\alpha_n(\beta)}$ is a monotonically increasing function of β. For $\lim_{\beta \to 0}\left[\frac{\alpha_m(\beta)}{\alpha_n(\beta)}\right] = 1$, it means that when β becomes smaller, it blurs the difference in relative importance. For $\lim_{\beta \to 0}\left[\frac{\alpha_m(\beta)}{\alpha_n(\beta)}\right] = \infty$, it extremely emphasizes the difference in relative importance.

IV. Simulation And Results Analysis

A. Simulation scenario

Four wireless access networks are considered: WiMAX, WCDMA, $WLAN_1$, and $WLAN_2$. The pilots or the beacons of all reachable networks are transmitted periodically. Therefore, measurements criterions listed in Figure 1 can be always obtained by FAHP system for decision making. The subjective MOS value is calculated based on the *R*–factor model defined in [8]. MTBF can be deduced from system reliability model using basic reliability equations [9]. In the implementation scenario, three types of services are considered for simulation: VoIP, FTP and video, with Poisson arrival rates $\lambda = 0.6, 0.5, 0.6$ respectively. In order to check the scheme's performance under heavy load, it is assumed that the serving rate (or the departure rate) of the system is smaller than arrival rate, which are $\mu = 0.4, 0.3, 0.4$ for VoIP, FTP and video, respectively.

Besides, two other schemes are given as comparisons to evaluate the performance of the proposed FAHP scheme, which are named as *AHP* scheme and *load balance* scheme, respectively. Load balance is a technique to distribute workload evenly across different RATs. In the simulation, load balance is realized by utilizing the network with the lightest load. Therefore this scheme improves global resource utilization by reducing regional congestion.

B. Simulation results and further study

Figure 3(a) shows the total revenue of the heterogeneous system in terms of GSI. The revenue upper bound achieved by load balance scheme is about 112. The AHP has a better performance with GSI upper bound at about 120, another considerable gain is achieved by the new proposed scheme, which has an upper bound at about 145, and the improvement becomes more obvious with increasing of arrival users. Compared to the previous two algorithms, FAHP expands network assessment from currently single layer to double layers. More importantly, fuzzy consistency concept is included in FAHP, which helps overcome the weakness of AHP in consistency and therefore makes reasonable decisions. Hence, the new proposed mechanism can find out the actual most appropriate network according to user-specific QoE requirements. Although AHP scheme takes multi-criterions into consideration, it is imprecise because consistency index in AHP does not accurately indicate user's perception. Consequently, system revenue degrades due to the mismatch between network capability and user services.

As indicated in Figure 3(b), the proposed scheme has the lowest reject ratio. Especially when the system is in heavy load situation, FAHP proves robustness due to its well "understanding" of the network situation. Therefore, the proposed scheme has a considerable gain of user number at the same level of reject ratio, especially when the system bears a heavy load.

In Figure 3(c), it depicts the comparison of reject ratio when using different parameter β. As analysis given in Section III, when $\beta = 1$, the difference in relative importance is blurred. Therefore, network selection is executed passively and user service will not be rejected

 ISBN: 978-1-61208-100-7

easily. The larger β is chosen, the user is more sensitive to the relative importance of all factors.

Therefore, different strategies are implemented by choosing different θ and β. Under different circumstances they are adjust to upgrade the overall revenue. When there exists a network connection for the user, then a passive access strategy is adopted to avoid unnecessary vertical handoff. That means small θ and β are chosen to make the user less sensitive to relative importance of factors. Then, vertical handoff is not worth, since the potential revenue improvement may not be able to compensate the cost caused by traffic handoff. The detail of this point will be studied in near future.

V. CONCLUTION

In this paper, an effective network selection scheme considering multiple QoE criterions is proposed to meet the challenges of multimedia applied in heterogeneous radio environments. For the sake of QoE evaluation, this scheme decomposes heterogeneous resource evaluation into multidimensional aspects, which are represented by KQIs. Meanwhile several KPIs are taken into account for each KQI, including both objective and subjective criterions. Then a fuzzy AHP system is designed for QoE reasoning and then gives performance ranking of all the network alternatives. Numerical results show that the proposed scheme outperforms the conventional AHP scheme and the load balance scheme.

REFERENCES

[1] Peter and B. Hestnes., “User measures of quality of experience: why being objective and quantitative is important,” Network, IEEE (Journal), vol. 24, Mar. 2010, pp. 8-13.

[2] C. Kuan-Ta, T. Cheng-Chun, and X. Wei-Cheng, “OneClick: A Franmework for Measuring Network Quality of Experience,” INFOCOM. IEEE, April. 2009, pp. 702-710, doi: 10.1109/INFCOM.2009.5061978.

[3] T. Srisakul, K. Shoaib, S. Eckehard, and K. Wolfgang. “QoE-Driven Cross-Layer Optimization for High Speed Downlink Packet Access,” Journal of Communication, IEEE, vol. 4, no. 9, Oct. 2009, doi: 10.43.4/jcm.4.9.669-680.

[4] G.Chen, S. Mei, Z. Yong, W. Li, and S. Jun-de. “Novel Network Selection Mechanism AHP and Enhanced GA,” Communication Networks and Services Research Conference, 2009(CNSR’09), pp. 397-401, doi: 10.1109/CNSR.2009.68.

[5] The Open Group, “SLA Management Handbook-Volume 4: Enterprise Perspective,” Oct. 2004.

[6] C.Chen-Tung, “Extensions of the TOPSIS for group decision-making under fuzzy environment,” Fuzzy Sets and Systems, vol. 114, Aug. 2000, pp. 1-9, doi: 10.1016/S0165-0114(97)00377-1.

[7] Z. Ji-jun “Fuzzy analytical hierarchy process,” Fuzzy Systems and Mathematics, vol. 14, no. 2, 2000, pp. 80–88.

[8] “The E-model, a computational model for user in transmission planning”, G.107, ITU-T, Series G: Transmission Systems and Media, Digital Systems and Networks, May. 2000.

[9] M. J. Mondro, “Approximation of Mean Time Between Failure When a System has Periodic Maintenance,” Reliablity, IEEE Transactions, vol. 51, Aug. 2002, pp. 166-167, doi: 10.1109/TR.2002.1011521.

 ISBN: 978-1-61208-100-7

EAP-Kerberos: Leveraging the Kerberos Credential Caching Mechanism for Faster Re-authentications in Wireless Access Networks

Saber Zrelli
Nobuo Okabe
Corporate R&D Headquarters
Yokogawa Electric Corporation
Tokyo, Japan
saber.zrelli,nobuo.okabe@jp.yokogawa.com

Yoichi Shinoda
Center for Information Science
Japan Advanced Institute of Science and Technology
Ishikawa, Japan
shinoda@jaist.ac.jp

***Abstract*—Although the wireless technology nowadays provides satisfying bandwidth and higher speeds, it still lacks improvements with regard to handoff performance. Existing solutions for reducing handoff delays are specific to a particular network technology or require expensive upgrades of the whole infrastructure. In this paper, we investigate performance benefits of leveraging the Kerberos ticket cashing mechanism for achieving faster re-authentications in IEEE 802.11 wireless access networks. For this purpose, we designed a new EAP authentication method, EAP-Kerberos, and evaluated re-authentication performance in different scenarios.**

***Keywords*-Wireless; Authentication; Handoff; Performance**

I. INTRODUCTION

Mobile wireless stations perform handoffs in order to change their point of attachment to the network. A handoff from an old access point to a new access point involves several steps that each may introduce delays.

A. What causes handoff delays

Handoff latency has long been an acknowledged issue in wireless networks. Some of the experimental studies [1] [2] have attributed the handoff delay in wireless local area networks (IEEE 802.11) to the scanning phase during which a wireless station discovers neighboring access points. These studies however, did not take authentication delays into consideration. In a previous work [3], we have shown that authentication using the Extensible Authentication Protocol [4] can take substantial delays especially when authentication servers are located in remote locations far from the access point.

B. How security impacts handoff delays

The Extensible Authentication protocol (EAP), is a core component in standard AAA (Authentication Authorization and Accounting) frameworks for access control in various network technologies such as 802.3, 802.11 and 802.16. In these frameworks, EAP authentication delays may become an issue especially in roaming situations; AAA frameworks support cross-domain authentication that enables an access network to authenticate a roaming client that belongs to a remote domain. The cross-domain authentication requires message exchange between the AAA server of the visited network and the AAA server of the roaming station's home network. Because these inter-domain exchanges occur over the Internet, they are subject to degradations such as packet loss and network delays which increases the overall authentication time. When a roaming station changes of access point, the same authentication procedure takes place again, disrupting the user traffic at each handoff.

C. Contributions

In this paper, we investigate performance benefits from using the Kerberos authentication protocol within wireless authentication frameworks that rely on the Extensible Authentication Protocol (EAP).

By relying on the legacy Kerberos authentication protocol as defined in [5], our scheme provides the same security properties as Kerberos and inherits its highly prized performance and simplicity. There are several aspects in the design of the Kerberos protocol that makes it suitable for use as the underlying authentication mechanism in wireless networks where handoff performance is a desired property. First, the Kerberos protocol uses symmetric key cryptography which consumes much less computing resources and hence introduces less delays compared to common methods based on public key cryptography. Second, the use of *Tickets* in Kerberos allows the client to perform fast re-authentication through a two round-trips exchange with the local authentication server, without the need for contacting any remote entity even if the client is in a roaming situation (i.e. The client belongs to a domain different from the domain that owns the local access network).

II. IEEE 802.1X EAP AUTHENTICATION

In order to gain access to the infrastructure, a wireless station (STA) needs to authenticate and share a key with the

 ISBN: 978-1-61208-100-7

Access Point (AP) using the *Extensible Authentication Protocol (EAP)* [6] and IEEE 802.1X. During EAP authentication, the AP acts as a pass-through between the STA and a back-end authentication server. As shown in Figure 1, EAP packets are transported over IEEE 802.1X between the AP and the STA in the front-end side, and using a AAA (Authentication Authorization and Accounting) protocol such as RADIUS [7] [8] or Diameter [9] [10] between the AP and the authentication server in the back-end side. After a successful authentication, the STA and the authentication server derive a shared key called *Master Session Key (MSK)*. Finally, the the back-end authentication server sends the MSK to the AP along with a notification of successful authentication.

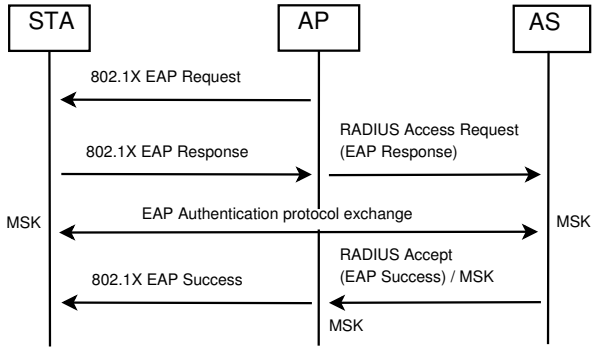


Fig. 1: IEEE 802.1X/ EAP authentication

III. The Kerberos Authentication Protocol

Kerberos [5] is a widely deployed authentication system. The authentication process in Kerberos involves *principals* and a *Key Distribution Center (KDC)*. Principals represent users and services registered in the Kerberos domain or realm. The KDC maintains a database of principals and shares a secret key with each one of them. In order to access an actual service, the client must submit valid Kerberos credentials to the service.

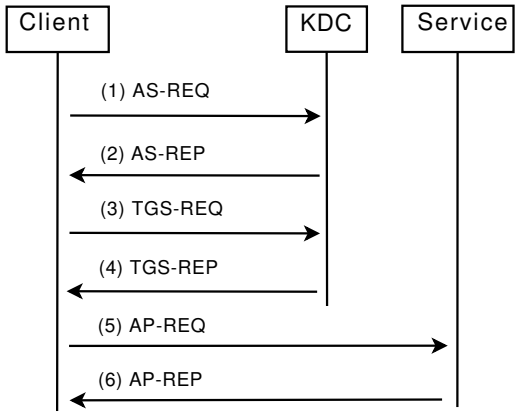


Fig. 2: The Kerberos authentication exchange

The Kerberos protocol specifies three exchanges, the *Authentication Server (AS) exchange*, the *Ticket Granting Service (TGS) exchange* and the *Client Server (AP) exchange*. The three exchanges are depicted in Figure 2. The AS exchange allows the client to obtain credentials that it can use to prove its identity to the KDC. These credentials consist of a Ticket referred to as *Ticket Granting Ticket (TGT)*, and a session key referred to as *TGS session key*. A Ticket is a message created by the Kerberos key distribution center and encrypted using the secret key of the target service.

The TGS exchange, on the other hand, allows the client to authenticate to the KDC using the TGT and to obtain a Ticket for a certain service. After validating the client request (TGS-REQ), the KDC issues a Ticket for the client and sends it along with the associated session key in a TGS Reply message (TGS-REP).

The AP exchange is performed between the client and the service to authenticate the client before granting it access to the resources. The client initiates the authentication by issuing an AP Request message (AP-REQ) that contains a Ticket for the service. After validating client's credentials the service authorizes the client and optionally sends an AP Reply message (AP-REP) to achieve mutual authentication.

IV. The EAP Kerberos Authentication Method

The EAP-Kerberos authentication method allows network clients to use Kerberos credentials to achieve mutual authentication with back-end authentication servers in wireless access networks.

The EAP-Kerberos method requires that each network access providers deploys a Kerberos realm with one or more Key distribution centers and that network clients are registered in the Kerberos principals database. In order to gain network access, a wireless station must possess a Kerberos login and password pair that can be used to authenticate the station to the network access provider's Kerberos KDC.

In the following sections, we present the design and operations of the EAP-Kerberos authentication method.

A. Overview

Our approach for using Kerberos in network access control is based on the notion of *Network Access Zones* that we define as a collection of lightweight access points managed by a single back-end authentication server. A set of network access zones that belong to the same provider constitutes an Access Network. Although an average sized access network may consist of a single network access zone, the partitioning of the access network into different zones is important for larger access networks such as those of wireless Internet providers. Generally, the use of multiple zones in large access networks makes management easier and ensures a scalable infrastructure.

To each network access zone corresponds a Kerberos service registered in a Kerberos key distribution center managed by an access network provider. Furthermore, the authentication server managing a certain zone has the secret Kerberos key of the corresponding zone. This secret key shared with the KDC

 ISBN: 978-1-61208-100-7

allows the authentication server to validate Kerberos AP-REQ messages it receives over EAP from wireless clients that are requesting access in the zone.

In order to gain network access within a zone, a wireless station must obtain a service Ticket for the local network access zone and present the Ticket to the zone's authentication server. The EAP-Kerberos method described hereafter specifies how the station obtains Kerberos credentials and how it uses them to authenticate and gain network access.

B. Station behavior

The STA and the authentication server negotiate the use of the EAP-Kerberos method as they would do for legacy EAP methods [6]. After a station have initiated the EAP-Kerberos method, the first message issued by the authentication server includes the Kerberos realm name as well as information identifying the local network access zone (see Section IV-A for the definition of network access zones). This information represented by REALM and ZONE in Figure 3 constitutes the local zone's Kerberos principal name that uniquely identifies it within the global Kerberos name space.

Upon reception of this first message, the STA checks its Kerberos credential cache for service Tickets and Kerberos Ticket Granting Tickets. Depending on what credentials are available, the station's behavior varies as follows.

1) Service Ticket for the local zone available: If the station has a Kerberos service Ticket for the local zone in its credential cache, then the STA initiates a Kerberos AP exchange over EAP with the authentication server managing the local zone.

2) Ticket Granting Ticket for local realm available: If the STA does not have a Ticket for the zone, but has a Ticket Granting Ticket for the local zone's Kerberos realm, then the STA must acquire a Ticket by performing a Kerberos TGS exchange with the Key Distribution Center where the zone is registered. The TGS exchange is tunneled in EAP between the STA and the local zone's authentication server. From there, the local zone's authentication server proxies the TGS exchange between the STA and the Kerberos KDC. For this purpose, the authentication server extracts the TGS-REQ message from the EAP-Kerberos message issued by the STA and sends it to the Kerberos KDC. The reply message from the KDC is sent back to the STA in an EAP-Kerberos message. After obtaining the service Ticket, the STA can perform an AP exchange with the authentication server.

3) No tickets available: If the STA does not have a service Ticket for the zone nor a TGT for the local realm, then it first needs to obtain a TGT for the local realm. The process of obtaining a TGT for the local realm depends on whether the Kerberos realm where the zone is registered is the same as the STA's home Kerberos realm or not. In the former case, the STA uses an AS exchange with the Kerberos KDC of the local realm. In the latter case, the STA first gets a TGT for its

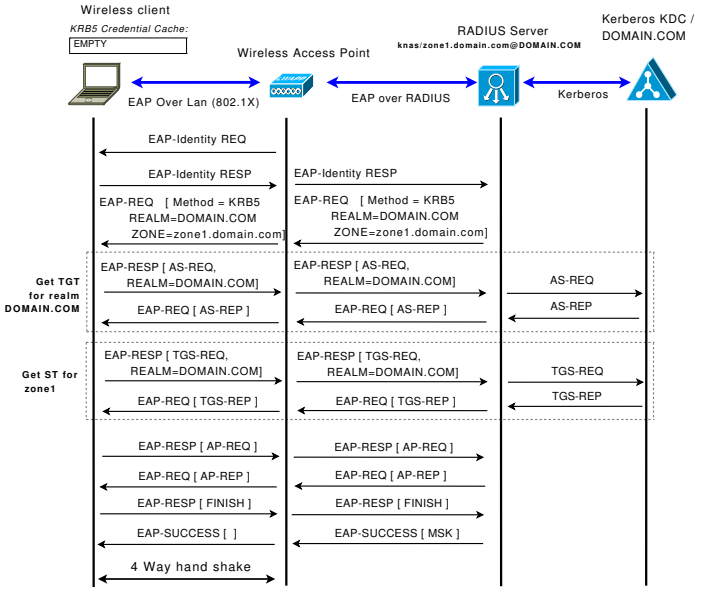


Fig. 3: Initial EAP-Kerberos authentication in the home access network

home Kerberos realm using an AS exchange with its home KDC, then, the STA performs Kerberos cross-realm TGS exchanges as specified in [5]. After the STA has obtained a TGT for the local realm, it performs a TGS exchange with the local realm's KDC to obtain a service Ticket for the local zone, then, it initiates an AP exchange with the local zone's authentication server to achieve mutual authentication and gain network access.

C. Use cases

When a STA is booted or is performing a handoff from an access point to another, it follows the same behavior described in Section IV-B. In the following, we provide more details and illustrate how the EAP-Kerberos method works in different use cases.

1) Initial authentication in the home network: The first use case we consider is the case where the STA needs to gain network connectivity through a network access zone that belongs to the station's home domain (e.g, when a subscriber is using her ISP's infrastructure). If the STA does not possess any cached Kerberos credentials for network access, then it needs to carry out three Kerberos exchanges; AS and TGS exchanges with the Kerberos KDC and an AP exchange with the authentication server managing the local network access zone.

As shown in Figure 3, the STA receives the Kerberos realm name (REALM) as well as the current zone's principal name (ZONE) in the initial EAP-Kerberos message issued by the zone's RADIUS authentication server. Since the STA does not possess any credentials yet, it first obtains a TGT using an AS exchange relayed by the access network infrastructure to

 ISBN: 978-1-61208-100-7

the STA's home Kerberos KDC. The EAP-Kerberos message carrying the AS-REQ message also contains the Kerberos realm name of the STA's home KDC. This information, will be used by the RADIUS server to locate the IP address of the Kerberos KDC to which the Kerberos message must be forwarded. In practice, IP addresses of Kerberos KDCs are resolved from Kerberos realm names using DNS SRV records[11] or mappings using static configuration files.

After obtaining the TGT, the STA requests a service ticket for the service "knas/zone1.domain.com@DOMAIN.COM". For this, the STA performs a TGS exchange with the Kerberos KDC of the current zone. As with the AS exchange, the TGS exchange is relayed by the RADIUS authentication server. The STA then initiates an AP exchange to authenticate with the RADIUS server managing the network access zone. After the AP exchange is completed, the STA issues a FINISH message to indicate to the RADIUS server that mutual authentication has been established. The authentication server then sends the EAP Master Session Key (MSK) to the AP. Finally, the STA and the AP use the shared MSK to establish a security association. In the case of IEEE 802.11, they perform the four-way handshake to derive Transient Session Keys from the MSK.

2) Intra-zone Handoff: The first handoff scenario we consider consists of the mobile STA's handoff within the same network access zone. In this case, the STA does not need to acquire new credentials since it already has a Ticket for the local zone (unless when the Ticket has expired). The authentication with the access network only requires an AP exchange with the local authentication server.

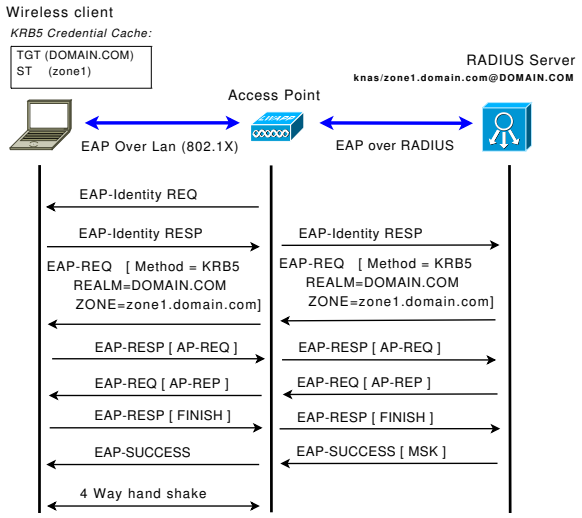


Fig. 4: EAP-Kerberos re-authentication in an Intra-zone handoff: The STA re-uses the Kerberos ticket for the current access network zone to re-authenticate with the RADIUS server.

As shown in Figure 4, the authentication, including all EAP messages, uses 3.5 round trips. All messages are exchanged within the zone and no entities in remote locations are involved in the intra-zone handoff process.

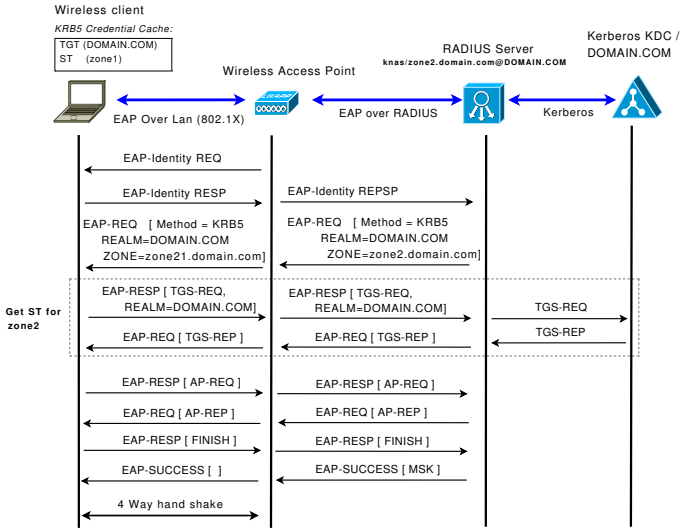


Fig. 5: EAP-Kerberos authentication in an Inter-zone handoff: The STA re-uses the TGT for the local realm to obtain a ticket for the new zone and authenticate with the RADIUS server.

3) Inter-zone Handoff: When the STA moves to a new access network zone, it may need to acquire a new Ticket for the new zone. If the new zone belongs to the same access network as the previous zone, the STA can re-use the TGT for the local realm to obtain a service Ticket for the new zone then authenticate with the new zone's authentication server. The inter-zone handoff scenario, as shown in Figure 5, requires an additional round-trip (for a total of 4.5) in comparison to the intra-zone handoff scenario.

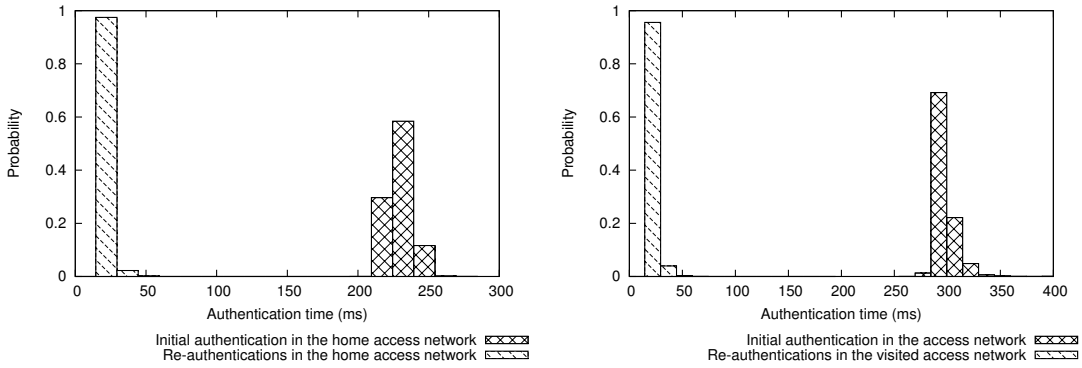
V. Performance Evaluation

We implemented the EAP-Kerberos method by extending the open-source *hostapd* [12] RADIUS server and the *wpa_supplicant* [13] EAP supplicant. For comparison, we performed performance evaluation of the EAP-PEAPv0 authentication method using Microsoft Windows 2003 server's Internet Authentication Server (IAS) on the same test-bed.

The test-bed consisted of two access networks, each composed of one network access zone. The two network access zones belong to different Kerberos realms and each has its own RADIUS authentication server and Kerberos Key Distribution Center. In order to emulate network delays, we used the Linux *netem* [14] utility. The resulting test-bed was equivalent to the reference architecture depicted in Figure 6.

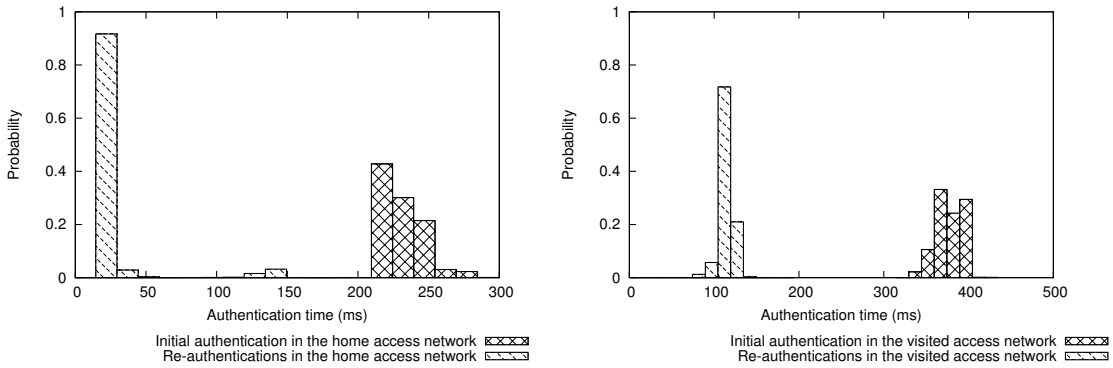
Figure 7 shows authentication delays using the EAP-Kerberos method in different scenarios. The re-authentication delay with EAP-Kerberos (30ms) is the same whether the wireless station is in its home network or in a visited network. When comparing intra-zone re-authentication delays in the home access network for the EAP-PEAPv0 method (Figure 8a)

 ISBN: 978-1-61208-100-7



(a) EAP-Kerberos performance in the home access network (b) EAP-Kerberos performance in the visited access network

Fig. 7: Probability Density Functions of re-authentication delays for the EAP-Kerberos authentication method



(a) EAP-PEAPv0 performance in the home access network (b) EAP-PEAPv0 performance in the visited access network

Fig. 8: Probability Density Functions of re-authentication delays for the EAP-PEAPv0 authentication method

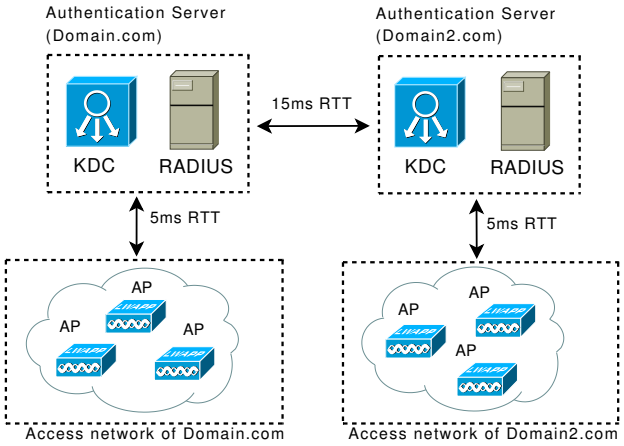


Fig. 6: Reference architecture

and the EAP-Kerberos method (Figure 7a), we can notice that both methods offer similar performance. However, in roaming scenarios when the station is performing handoffs in a visited access network, the EAP-Kerberos method shows superior performance.

As shown in Figure 7b, intra-zone re-authentication delays remained acceptable in the roaming case for EAP-Kerberos (30ms) while re-authentication latency increased about four folds from 30ms to around 120ms for EAP-PEAPv0 (Figure 8b). This is due to the fact that EAP-PEAPv0 (as it is the case for all legacy EAP authentication methods) require message exchange with the roaming station's home RADIUS server for performing re-authentications in foreign access networks while the EAP-Kerberos method involves only entities in the the visited access network.

VI. Conclusion

In order to achieve true ubiquitous applications, the handoff delays in wireless networks must be kept to the minimum. Several steps in the handoff process may be subject to enhancements. In this paper, we consider authentication delays during handoffs. The problem with handoff latency arises when a roaming wireless station performs handoffs in a foreign access network. The inter-domain exchanges necessary for

 ISBN: 978-1-61208-100-7

authenticating the roaming station may introduce large delays that would affect quality of service in real-time applications.

We have designed, implemented and evaluated a Kerberos-based EAP authentication method that achieves strong authentication with reduced latency during handoffs. Experimental results from our test-bed show that EAP-Kerberos re-authentications in roaming scenarios took around 30 milliseconds, more than 3 times faster than EAP-PEAPv0 that took around 120 milliseconds.

When compared to existing solutions for reducing EAP re-authentication delays such as IEEE 802.11r [15] and ERP [16], the approach presented in this paper has three main advantages; (1) The proposed method extends the EAP layer by specifying a new EAP method which ensures that the proposed approach is link layer independent. (2) The proposed approach does not require changes in the access point, and therefore it has an advantage from deployment cost point of view, and (3) The EAP method proposed in this paper supports fast inter access point and inter access network handoffs by relying on Kerberos cross-realm authentication capabilities. Other existing approaches enable fast re-authentication only within the same access network.

REFERENCES

[1] A. Mishra, M. Shin, and W. Arbaugh, "An empirical analysis of the IEEE 802.11 mac layer handoff process," SIGCOMM Comput. Commun. Rev., vol. 33, no. 2, pp. 93–102, 2003.

[2] H. Velayos and G. Karlsson, "Techniques to reduce the IEEE 802.11b handoff time," Tech. Rep., 20-24 June 2004.

[3] S. Zrelli and Y. Shinoda, "Experimental evaluation of EAP performance in roaming scenarios," in Sustainable Internet, Third Asian Internet Engineering Conference, ser. Lecture Notes in Computer Science, S. Fdida and K. Sugiura, Eds., vol. 4866. Springer, 2007, pp. 86–98. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-76809-8_8

[4] B. Aboba, L. Blunk, J. Vollbrecht, J. Carlson, and H. Levkowetz, "Extensible Authentication Protocol (EAP)," RFC 3748 (Proposed Standard), Jun. 2004. [Online]. Available: http://www.ietf.org/rfc/rfc3748.txt

[5] C. Neuman, T. Yu, S. Hartman, and K. Raeburn, "The Kerberos Network Authentication Service (V5)," RFC 4120 (Proposed Standard), Internet Engineering Task Force, Jul. 2005. [Online]. Available: http://www.ietf.org/rfc/rfc4119.txt

[6] B. Aboba, L. Blunk, J. Vollbrecht, J. Carlson, and H. Levkowetz, "Extensible Authentication Protocol (EAP)," RFC 3748 (Proposed Standard), Jun. 2004. [Online]. Available: http://www.ietf.org/rfc/rfc3748.txt

[7] C. Rigney, S. Willens, A. Rubens, and W. Simpson, "Remote Authentication Dial In User Service (RADIUS)," RFC 2865 (Draft Standard), Internet Engineering Task Force, Jun. 2000. [Online]. Available: http://www.ietf.org/rfc/rfc2865.txt

[8] B. Aboba and P. Calhoun, "RADIUS (Remote Authentication Dial In User Service) Support For Extensible Authentication Protocol (EAP)," RFC 3579 (Informational), Sep. 2003. [Online]. Available: http://www.ietf.org/rfc/rfc3579.txt

[9] P. Calhoun, J. Loughney, E. Guttman, G. Zorn, and J. Arkko, "Diameter Base Protocol," RFC 3588 (Proposed Standard), Internet Engineering Task Force, Sep. 2003. [Online]. Available: http://www.ietf.org/rfc/rfc3588.txt

[10] P. Eronen, T. Hiller, and G. Zorn, "Diameter Extensible Authentication Protocol (EAP) Application," RFC 4072 (Proposed Standard), Internet Engineering Task Force, Aug. 2005. [Online]. Available: http://www.ietf.org/rfc/rfc4072.txt

[11] A. Gulbrandsen, P. Vixie, and L. Esibov, "A DNS RR for specifying the location of services (DNS SRV)," RFC 2782 (Proposed Standard), Feb. 2000. [Online]. Available: http://www.ietf.org/rfc/rfc2782.txt

[12] "hostapd: IEEE 802.11 AP, IEEE 802.1X/WPA/WPA2/EAP/RADIUS Authenticator," Web page, As of July 2010. [Online]. Available: http://hostap.epitest.fi/hostapd/

[13] "Linux WPA/WPA2/IEEE 802.1X Supplicant," Web page, As of July 2010. [Online]. Available: http://hostap.epitest.fi/wpa_supplicant/

[14] "Netem: The Linux Network Emulator," Web page, As of July 2010. [Online]. Available: http://www.linuxfoundation.org/collaborate/workgroups/networking/netem

[15] 802.11r, "IEEE Standard for Information technology, Telecommunications and information exchange between systems, Local and metropolitan area networks – Specific requirements Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications Amendment 2: Fast Basic Service Set (BSS) Transition," IEEE Standards, 2008. [Online]. Available: http://dx.doi.org/10.1109%2FIEEESTD.2008.4573292

[16] V. Narayanan and L. Dondeti, "EAP Extensions for EAP Re-authentication Protocol (ERP)," RFC 5296 (Proposed Standard), Aug. 2008. [Online]. Available: http://www.ietf.org/rfc/rfc5296.txt

 ISBN: 978-1-61208-100-7

A One-Shot Dynamic Optimization Methodology for Wireless Sensor Networks

Arslan Munir and Ann Gordon-Ross
Department of Electrical and Computer Engineering
University of Florida, Gainesville, Florida 32611
Email: amunir@ufl.edu, ann@chrec.org

Susan Lysecky and Roman Lysecky
Department of Electrical and Computer Engineering
University of Arizona, Tucson, Arizona 85721
Email: {slysecky, rlysecky}@ece.arizona.edu

***Abstract*—Wireless sensor networks (WSNs), consisting of autonomous sensor nodes, have emerged as ubiquitous networks which span diverse application domains (e.g., health care, logistics, defense) each with varying application requirements (e.g., lifetime, throughput). Sensor nodes possess tunable parameters (e.g., processor voltage, sensing frequency), which enable platform specialization for particular application requirements. WSN application design can be daunting for application developers, which are oftentimes not trained engineers (e.g., biologists, agriculturists) who wish to utilize the sensor-based systems within their given domain. Dynamic optimizations enable sensor-based platforms to tune parameters in-situ to automatically determine an operating state. However, rapidly changing application behavior and environmental stimuli necessitate a lightweight and highly responsive dynamic optimization methodology. In this paper, we propose One-Shot – a lightweight dynamic optimization methodology that determines initial tunable parameter settings to give a high-quality operating state in *One-Shot* for time-critical and highly constrained applications. Results reveal that One-Shot solution is within 5.92% of the optimal solution on average. To assist dynamic optimizations in determining an operating state, we propose an application metric estimation model to establish a relationship between application metrics (e.g., lifetime) and sensor-based platform parameters.**

***Keywords*-Wireless sensor networks; dynamic optimization; application metric estimation**

I. Introduction and Motivation

Wireless sensor networks (WSNs) consist of spatially distributed autonomous sensor nodes that observe a phenomenon (environment, target, etc.). WSNs are becoming ubiquitous due to their proliferation in diverse application domains (e.g., defense, health care, logistics) each with varying application requirements. For example, a security/defense system may have a high throughput requirement whereas an ambient conditions monitoring application may be more sensitive to lifetime. This diversity makes WSN design challenging using commercial-off-the-shelf (COTS) sensor nodes.

COTS sensor nodes are mass-produced to optimize for cost and are not specialized for any particular application. Furthermore, WSN application developers oftentimes are not trained engineers, but rather biologists, teachers, or agriculturists who wish to utilize the sensor-based systems within their given domain. Fortunately, many COTS sensor nodes possess tunable parameters (e.g., processor voltage and frequency, sensing frequency) whose values may be *tuned* for a specific application. Faced with an overwhelming number of tunable parameter choices, WSN design may be a daunting task for non-experts and necessitates an automated parameter tuning process for assistance.

Parameter optimization is the process of assigning appropriate (optimal or near-optimal) values to tunable parameters either statically or dynamically to meet application requirements. *Static optimizations* assign parameter values at deployment and these values remain fixed during a sensor node's lifetime. Accurate prediction/simulation of environmental stimuli is challenging and applications with changing environmental stimuli do not benefit from static optimizations. *Dynamic optimizations* assign parameter values during runtime and reassign/change these values in accordance with changing environmental stimuli, thus enabling closer adherence to application requirements.

There exists much research in the area of dynamic optimizations (e.g., [1][2][3][4]), but most previous work targets the memory (cache) or processor in computer systems. Little work exists on WSN dynamic optimization, which presents additional challenges due to a WSN's unique design space, energy constraints, and operating environment. Shenoy et al. [5] presented profiling methods for dynamically monitoring sensor-based platforms and analyzed the associated network traffic and energy, but did not explore dynamic optimizations. In prior work, Munir et al. [6] proposed a Markov Decision Process (MDP)-based methodology as a first step towards WSN dynamic optimization, but this methodology required excessive computational resources for larger design spaces. Wang et al. [7] proposed an energy efficient optimization method for target tracking applications that consisted of dynamic awakening and an optimal sensing scheme. Khanna et al. [8] proposed a genetic algorithm for secure and dynamic deployment of resource-constrained multi-hop WSNs. Some previous works [9][10] explored WSN dynamic voltage and frequency scaling (DVFS) for dynamic optimization, but DVFS only considered two sensor node tunable parameters (processor voltage and frequency).

In this paper, we explore a fine-grained design space for sensor-based platforms with many tunable parameters

 ISBN: 978-1-61208-100-7

to more closely meet application requirements (Gordon-Ross et al. [11] showed that finer-grained design spaces provide interesting design alternatives and result in increased benefits in the cache subsystem). The exploration of a fine-grained design space coupled with limited battery reserves and rapidly changing application requirements and environmental stimuli necessitates a lightweight and highly responsive dynamic optimization methodology.

We propose One-Shot – a lightweight dynamic optimization methodology that determines appropriate initial tunable parameter values to give a good quality operating state (tunable parameter value settings) in *One-Shot* with minimal design exploration for highly constrained applications. Results reveal that the One-Shot operating state is within 5.92% of the optimal solution (obtained from exhaustive search) averaged over several different application domains and design spaces. To assist dynamic optimizations in determining an operating state, we for the first time, to the best of our knowledge, propose an *application metric estimation model*, which estimates high-level metrics (lifetime, throughput, and reliability) from sensor-based platform parameters (e.g., processor voltage and frequency, sensing frequency, transceiver transmission power, etc.). The dynamic optimization methodology leverages this estimation model while comparing different operating states for optimization purposes.

II. Dynamic Optimization Methodology

In this section, we give an overview of One-Shot and associated algorithm. We also formulate the state space and objective function for One-Shot.

A. Overview

Fig. 1 depicts our One-Shot dynamic optimization methodology for WSNs. WSN designers evaluate application requirements and capture these requirements as high-level *application metrics* (e.g., lifetime, throughput, reliability) and associated *weight factors*. The weight factors signify the weightage/importance of each application metric with respect to each other. One-Shot leverages an application metric estimation model to determine application metric values offered by an operating state.

Fig. 1 shows the per-node One-Shot process (encompassed by the dashed circle), which is orchestrated by the *dynamic optimization controller*. The dynamic optimization controller invokes *One-Shot* wherein the sensor node operating state is directly determined using an intelligent tunable parameter value setting selection methodology (i.e., in *One-Shot*). One-Shot also determines an exploration order (ascending or descending) for the tunable parameters. This exploration order can be leveraged by an *online optimization algorithm* to provide improvements over the One-Shot solution by further design space exploration and is the focus of our future work. This

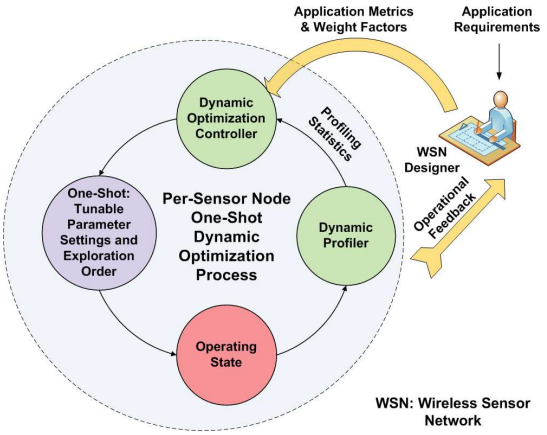


Figure 1. One-Shot dynamic optimization methodology for wireless sensor networks.

exploration order is critical in reducing the number of states explored by the online optimization algorithm. The sensor node moves directly to the operating state specified by One-Shot. A *dynamic profiler* records profiling statistics (e.g., remaining battery energy, wireless channel condition) given the current operating state and environmental stimuli and passes these profiling statistics to the dynamic optimization controller.

The dynamic optimization controller processes the profiling statistics to determine if the current operating state meets the application requirements. If the application requirements are not met, the dynamic optimization controller reinvokes One-Shot to determine the new operating state. This feedback process continues to ensure the selection of an appropriate operating state to better meet application requirements.

B. State Space

The state space S for One-Shot given N tunable parameters is defined as:

$$S = P_1 \times P_2 \times \cdots \times P_N \tag{1}$$

where P_i denotes the state space for tunable parameter i, $\forall\, i \in \{1, 2, \ldots, N\}$ and $\times$ denotes the Cartesian product. Each tunable parameter P_i consists of n values:

$$P_i = \{p_{i_1}, p_{i_2}, p_{i_3}, \ldots, p_{i_n}\} \;\; : \;\; |P_i| = n \tag{2}$$

where $|P_i|$ denotes the tunable parameter P_i's state space cardinality (the number of tunable values in P_i). S is a set of n-tuples (each n-tuple represents a sensor node state) formed by taking one tunable parameter value from each tunable parameter. A single n-tuple $s \in S$ is given as:

$$\begin{aligned} s \;\; = \;\; & (p_{1_y}, p_{2_y}, \ldots, p_{N_y}) : p_{i_y} \in P_i, \\ & \forall\, i \in \{1, 2, \ldots, N\},\, y \in \{1, 2, \ldots, n\} \end{aligned} \tag{3}$$

We point out that some n-tuples in S may not be feasible (such as invalid combinations of processor voltage and frequency) and can be treated as *do not care* tuples.

C. Optimization Objection Function

The sensor node dynamic optimization problem can be formulated as an unconstrained optimization problem:

$$\begin{aligned} \max f(s) \;&=\; \sum_{k=1}^{m} \omega_k f_k(s) \\ s.t. \quad & s \in S \\ & \omega_k \geq 0, \quad k = 1, 2, \ldots, m. \\ & \omega_k \leq 1, \quad k = 1, 2, \ldots, m. \\ & \sum_{k=1}^{m} \omega_k = 1, \end{aligned} \tag{4}$$

where $f(s)$ denotes the objective function characterizing application metrics and weight factors. $f_k(s)$ and ω_k in (4) denote the objective function and weight factor for the k^{th} application metric, respectively, given that there are m application metrics. Each state $s \in S$ has an associated objective function value and the optimization goal is to determine a state that gives the maximum (optimal) objective function value $f^{opt}(s)$ where $f^{opt}(s)$ indicates the best possible adherence to the application requirements given the design space S.

For our dynamic optimization methodology, we consider three application metrics (m = 3), which are lifetime, throughput, and reliability, each with piecewise linear objective functions. A piecewise linear objective function captures the desirable and acceptable ranges of a particular application metric. For example, for a particular application, a lifetime metric may have an acceptable minimum value of 40 days and reliability may be a more important metric than the lifetime. The objective function delineates this inter-metric relative importance and attainable application metric values. Even though we consider piecewise linear objective functions, our methodology works well for any other objective function characterization (e.g., linear, non-linear).

D. One-Shot Dynamic Optimization Algorithm

In this subsection, we describe associated algorithm for One-Shot. The algorithm determines initial tunable parameter value settings and exploration order (ascending or descending).

Algorithm 1 describes One-Shot's algorithm to determine initial tunable parameter value settings and exploration order. The algorithm takes as input the objective function $f(s)$, the number of tunable parameters N, the number of values for each tunable parameter n, the number of application metrics m, and $\boldsymbol{P}$ where $\boldsymbol{P}$ represents a vector containing the tunable parameters, $\boldsymbol{P} = \{P_1, P_2, \ldots, P_N\}$. For each application metric k, the algorithm calculates vectors $\boldsymbol{P_0^k}$ and $\boldsymbol{P_d^k}$ (where d denotes the exploration direction (ascending or descending)), which store the initial value settings and exploration order, respectively, for the

Input: *f(s), N, n, m,* $\boldsymbol{P}$
Output: Initial tunable parameter value settings and exploration order

```
1  for k ← 1 to m do
2      for P_i ← P_1 to P_N do
3          f^k_{p_{i_1}} ← k-metric objective function value when
           parameter setting is {P_i = p_{i_1}, P_j = P_{j_0}, ∀ i ≠ j} ;
4          f^k_{p_{i_n}} ← k-metric objective function value when
           parameter setting is {P_i = p_{i_n}, P_j = P_{j_0}, ∀ i ≠ j} ;
5          δf^k_{P_i} ← f^k_{p_{i_n}} − f^k_{p_{i_1}} ;
6          if δf^k_{P_i} ≥ 0 then
7              explore P_i in descending order ;
8              P^k_d[i] ← descending ;
9              P^k_0[i] ← p^k_{i_n} ;
10         else
11             explore P_i in ascending order ;
12             P^k_d[i] ← ascending ;
13             P^k_0[i] ← p^k_{i_1} ;
14         end
15     end
16 end
```

return $\boldsymbol{P_d^k}$, $\boldsymbol{P_0^k}, \forall\, k \in \{1, \ldots, m\}$

Algorithm 1: One-shot dynamic optimization algorithm.

tunable parameters. The algorithm determines $f^k_{p_{i_1}}$ and $f^k_{p_{i_n}}$ (the k^{th} application metric objective function values) where the parameter being explored P_i is assigned its first p_{i_1} and last p_{i_n} tunable values, respectively, and the rest of the tunable parameters $P_j, \forall\, j \neq i$ are assigned initial values (lines 3 - 4). $\delta f^k_{P_i}$ stores the difference between $f^k_{p_{i_n}}$ and $f^k_{p_{i_1}}$. $\delta f^k_{P_i} \geq 0$ means that p_{i_n} results in a greater (or equal when $\delta f^k_{P_i} = 0$) objective function value as compared to p_{i_1} for parameter P_i (i.e., the objective function value decreases as the parameter value decreases). To reduce the number of states explored while considering that an online optimization algorithm (e.g., greedy-based algorithm) will typically stop exploring a tunable parameter if a tunable parameter's value yields a comparatively lower (or equal) objective function value, P_i's exploration order must be descending (lines 6 - 8). The algorithm assigns p_{i_n} as the initial value of P_i for the k^{th} application metric (line 9). If $\delta f^k_{P_i} < 0$, the algorithm assigns the exploration order as ascending for P_i and p_{i_1} as the initial value setting of P_i (lines 11 - 13). This $\delta f^k_{P_i}$ calculation procedure is repeated for all m application metrics and all N tunable parameters (lines 1 - 16).

III. Application Metric Estimation Model

In this section, we propose an application metric estimation model leveraged by One-Shot. This estimation model estimates high-level application metrics (lifetime, throughput, reliability) from sensor node parameters (e.g., processor voltage and frequency, transceiver voltage, etc.). For brevity, we describe only the estimation model's key elements.

 ISBN: 978-1-61208-100-7

A. Lifetime Estimation

Lifetime of a sensor node is defined as the time duration between the deployment time and the time before which the sensor node fails to perform the assigned task due to sensor node failure. The sensor failure due to battery energy depletion is normally taken into account for lifetime estimation. The sensor node typically contains AA alkaline batteries whose energy depletes gradually as the sensor node consumes energy during operation. The critical factors in determining sensor node lifetime are battery energy and energy consumption during operation.

The sensor node lifetime in days $\mathcal{L}_s$ can be estimated as:

$$\mathcal{L}_s = \frac{E_b}{E_c \times 24} \tag{5}$$

where E_b denotes the sensor node's battery energy (Joules) and E_c denotes the sensor node's energy consumption per hour.

We model E_c as the sum of processing energy, communication energy, and sensing energy:

$$E_c = E_{proc} + E_{com} + E_{sen} \ \ (J) \tag{6}$$

where E_{proc}, E_{com}, and E_{sen} denote processing energy per hour, communication energy per hour, and sensing energy per hour, respectively.

The *processing energy* accounts for the energy consumed in processing the sensed data by the sensor node's processor. We assume that the sensor node's processor operates in two modes: active mode and idle mode [12]. We point out that although we only consider active and idle modes, a processor operating in other sleep modes apart from idle mode (e.g., power-down, power-save, standby, etc.) can also be incorporated in our model. E_{proc} is given by:

$$E_{proc} = E^a_{proc} + E^i_{proc} \tag{7}$$

where E^a_{proc} and E^i_{proc} denote the processor's energy consumption per hour in active mode and idle mode, respectively.

The sensor nodes communicate with each other (e.g., send packets containing the sensed data), which consumes *communication energy*. The communication energy is the sum of transmission, receive, and idle energy for a sensor node's transceiver:

$$E_{com} = E^{tx}_{trans} + E^{rx}_{trans} + E^i_{trans} \tag{8}$$

where E^{tx}_{trans}, E^{rx}_{trans}, and E^i_{trans} denote the transceiver's transmission energy per hour, receive energy per hour, and idle energy per hour, respectively.

The energy consumption due to sensing the observed phenomenon is termed as *sensing energy*. The sensing energy mainly depends upon the sensing (sampling) frequency and the number of sensors attached to the sensor board (e.g., the MTS400 sensor board [13] has Sensirion SHT1x temperature and humidity sensors [14]). The sensors consume energy while taking sensing measurements and switch to the idle mode for energy conservation while not sensing. E_{sen} is given by:

$$E_{sen} = E^m_{sen} + E^i_{sen} \tag{9}$$

where E^m_{sen} denotes the sensing measurement energy per hour and E^i_{sen} denotes the sensing idle energy per hour.

B. Throughput Estimation

In the context of dynamic optimizations, *throughput* can be interpreted relative to the state (tunable parameter value settings) that deliver the maximum quality (rate) sensing process, processing, and transmission to observe a phenomenon while minimizing the cost (energy consumption). Three processes contribute to the throughput for sensor nodes: sensing, processing, and communication. The throughput interpretation may vary depending upon the WSN application design as these throughputs can have different relative importance for different applications. The aggregate throughput R (typically measured in bits/second) can be considered as a weighted sum of constituent throughputs:

$$R = \omega_s R_{sen} + \omega_p R_{proc} + \omega_c R_{com} \ : \ \omega_s + \omega_p + \omega_c = 1 \tag{10}$$

where R_{sen}, R_{proc}, and R_{com} denote the sensing throughput, processing throughput, and communication throughput, respectively. ω_s, ω_p, and ω_c denote the weight factors for sensing, processing, and communication throughput, respectively.

The sensing throughput is the throughput due to sensing activity and measures the sensing bits sampled per second. R_{sen} is given by:

$$R_{sen} = F_s \cdot R^b_{sen} \tag{11}$$

where F_s and R^b_{sen} denote sensing frequency and sensing resolution bits, respectively.

The processing throughput is the throughput due to the processor's processing of sensed measurements and measures the bits processed per second. R_{proc} is given by:

$$R_{proc} = F_p / N^b \tag{12}$$

where F_p and N^b denote processor frequency and the number of processor instructions to process one bit, respectively.

The communication throughput R_{com} results from the transfer of data packets over the wireless channel and is given by:

$$R_{com} = P^{eff}_s \times 8 / t^{pkt}_{tx} \tag{13}$$

where t^{pkt}_{tx} denotes the time to transmit one packet and P^{eff}_s denotes the effective packet size excluding the packet header overhead.

 ISBN: 978-1-61208-100-7

C. Reliability Estimation

The reliability metric measures the number of packets transferred reliably (i.e., error free packet transmission) over the wireless channel. Accurate reliability estimation is challenging because the various factors involved change dynamically, such as network topology, number of neighboring sensor nodes, wireless channel fading, sensor network traffic, packet size, etc. The two main factors that affect reliability are transceiver transmission power P_{tx} and receiver sensitivity. For example, the AT86RF230 transceiver [15] has a receiver sensitivity of -101 dBm with corresponding packet error rate (PER) $\leq$ 1% for additive white Gaussian noise (AWGN) channel with physical service data unit (PSDU) equal to 20 bytes. Reliability can be estimated using Friis free space transmission equation [16] for different P_{tx} values, distance between transmitting and receiving sensor nodes, and fading models (e.g., shadowing fading model). Reliability values can be assigned corresponding to P_{tx} values such that the higher P_{tx} values give higher reliability. However, more accurate reliability estimation requires profiling statistics for the number of packets transmitted and the number of packets received.

IV. Experimental Results

A. Experimental Setup

We base our experimental setup on the Crossbow IRIS mote platform [17], which has a battery capacity of 2000 mA-h with two AA alkaline batteries. The IRIS mote platform integrates an Atmel ATmega1281 microcontroller [12], an Atmel AT-86RF230 low power 2.4 GHz transceiver [15], an MTS400 sensor board [13] with Sensirion SHT1x temperature and humidity sensors [14].

We analyze six tunable parameters: processor voltage V_p, processor frequency F_p, sensing frequency F_s, packet size P_s, packet transmission interval P_{ti}, and transceiver transmission power P_{tx}. In order to evaluate our methodology across small and large design spaces, we consider two design space cardinalities (number of states in the design space): $|S| = 729$ and $|S| = 31,104$ ($|S| = 729$ is a subset of $|S| = 31,104$). The tunable parameters for $|S| = 31,104$ are V_p = {1.8, 2.7, 3.3, 4, 4.5, 5} (volts), F_p = {2, 4, 6, 8, 12, 16} (MHz) [12], F_s = {0.2, 0.5, 1, 2, 3, 4} (samples per second) [14], P_s = {32, 41, 56, 64, 100, 127} (bytes), P_{ti} = {10, 30, 60, 300, 600, 1200} (seconds), and P_{tx} = {-17, -3, 1, 3} (dBm) [15]. All state space tuples are feasible for $|S| = 729$, whereas $|S| = 31,104$ contains 7,779 infeasible state space tuples (e.g., all V_p and F_p pairs are not feasible).

We model three application domains (a security/defense system (S/D), a health care application (HC), and an ambient conditions monitoring application (AC)) to evaluate the robustness of One-Shot across different applications. We assign objective function parameter values such as minimum and maximum values of application metrics and their associated weight factors considering typical application requirements [18].

In order to evaluate One-Shot's solution quality, we compare the solution from One-Shot's initial parameter settings $\mathcal{I}$ with the solutions obtained from the following four potential initial parameter value settings (although any feasible n-tuple $s \in S$ can be taken as the initial parameter settings): $\mathcal{I}_1$ assigns the first parameter value for each tunable parameter (i.e., $\mathcal{I}_1 = p_{i_1},\ \forall\ i \in \{1, 2, \ldots, N\}$); $\mathcal{I}_2$ assigns the last parameter value for each tunable parameter (i.e., $\mathcal{I}_2 = p_{i_n},\ \forall\ i \in \{1, 2, \ldots, N\}$); $\mathcal{I}_3$ assigns the middle parameter value for each tunable parameter (i.e., $\mathcal{I}_3 = \lfloor p_{i_n}/2 \rfloor,\ \forall\ i \in \{1, 2, \ldots, N\}$); $\mathcal{I}_4$ assigns a random value for each tunable parameter (i.e., $\mathcal{I}_4 = p_{i_q} : q = \texttt{rand()} \,\%\, n,\ \forall\ i \in \{1, 2, \ldots, N\}$).

B. Results

We implemented One-Shot in C++. We compare our results with four different initial parameter arrangements (Section IV-A) and normalize the objective function value corresponding to the operating state attained by One-Shot with respect to the optimal solution obtained using an exhaustive search. We compare the relative complexity of One-Shot with two other dynamic optimization methodologies.

1) Percentage Improvements over other Initial Parameter Settings: Table I depicts the percentage improvements attained by One-Shot parameter settings $\mathcal{I}$ over other parameter settings for different application domains. We observe that some arbitrary settings may give a comparable solution for a particular application domain, application metric weight factors, and design space cardinality, but that arbitrary setting would not scale to other application domains, application metric weight factors, and design space cardinalities. For example, $\mathcal{I}_1$ achieves the same solution quality as of $\mathcal{I}$ for AC, but yields 73.27% and 147.87% lower quality solutions than $\mathcal{I}$ for HC and S/D, respectively, for $|S| = 31,104$. Furthermore, $\mathcal{I}_1$ yields a 51.85% lower quality solution than $\mathcal{I}$ for AC when $|S| = 729$. The average percentage improvement attained by $\mathcal{I}$ over all application domains and design spaces is 44.79%. In summary, results reveal that on average $\mathcal{I}$ gives a solution within 5.92% of the optimal solution obtained from exhaustive search.

2) Comparison with Greedy- and SA-based Dynamic Optimization Methodologies: In order to investigate the effectiveness of One-Shot, we compare the One-Shot solution's quality (indicated by the attained objective function value) with two other dynamic optimization methodologies, which leverage SA-based and greedy-based (denoted by GD^{asc} where asc stands for ascending order of parameter exploration) design space exploration. We assign initial parameter value settings for greedy and SA-based methodologies as $\mathcal{I}_1$ and $\mathcal{I}_4$, respectively. For brevity we

 ISBN: 978-1-61208-100-7

Table I
PERCENTAGE IMPROVEMENTS ATTAINED BY $\mathcal{I}$ OVER OTHER INITIAL PARAMETER SETTINGS FOR $|S| = 729$ AND $|S| = 31,104$.

–	$\|S\| = 729$				$\|S\| = 31,104$			
Application Domain	$\mathcal{I}_1$	$\mathcal{I}_2$	$\mathcal{I}_3$	$\mathcal{I}_4$	$\mathcal{I}_1$	$\mathcal{I}_2$	$\mathcal{I}_3$	$\mathcal{I}_4$
Security/Defense System (S/D)	154.9%	10.25%	56.62%	29.18%	147.87%	0.318%	9.72%	91.88%
Health Care (HC)	77.6%	6.66%	30.73%	10.86%	73.27%	0.267%	9.62%	45.17%
Ambient Conditions Monitoring (AC)	51.85%	6.17%	20.39%	6.85%	0%	75.5%	50.97%	108.31%

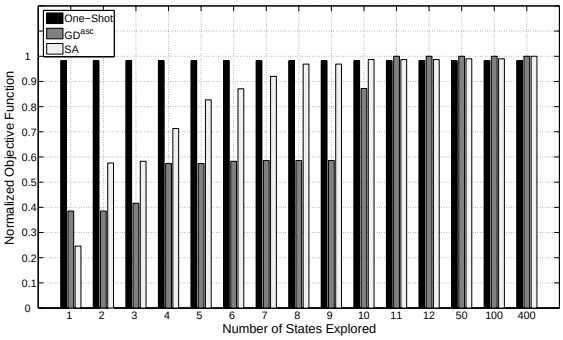


Figure 2. Objective function value normalized to the optimal solution for S/D where $\omega_l = 0.25$, $\omega_t = 0.35$, $\omega_r = 0.4$, $|S| = 729$.

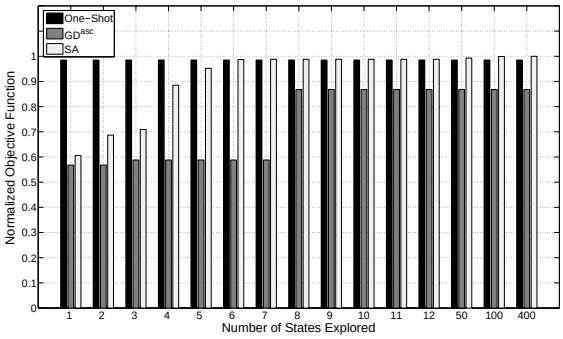


Figure 3. Objective function value normalized to the optimal solution for HC where $\omega_l = 0.25$, $\omega_t = 0.35$, $\omega_r = 0.4$, $|S| = 31,104$.

present results for $\mathcal{I}_1$ and $\mathcal{I}_4$, but results for $\mathcal{I}_2$ and $\mathcal{I}_3$ revealed similar trends.

Fig. 2 shows the objective function value normalized to the optimal solution versus the number of states explored for One-Shot, GD^{asc}, and SA algorithms for S/D for $|S| = 729$. One-Shot's solution is within 1.8% of the optimal solution. The figure shows that GD^{asc} and SA explore 11 states (1.51% of the design space) and 10 states (1.37% of the design space), respectively, to attain an equivalent or better quality solution than the One-Shot solution. Although greedy and SA explore few states to reach a comparable solution as that of One-Shot, One-Shot is suitable when design space exploration is not an option due to an extremely large design space and/or extremely stringent computational, memory, and timing constraints. These results reveal that other arbitrary initial value settings do not provide a good quality operating state and necessitate additional design space exploration to provide a good quality operating state.

Fig. 3 shows the objective function value normalized to the optimal solution versus number of states explored for HC for $|S| = 31,104$. One-shot's solution is within 1.5% of the optimal solution. The figure shows that GD^{asc} converges to a lower quality solution than One-Shot's solution after exploring 8 states (0.026% of the design space) and SA explores 6 states (0.019% of the design space) to yield a better quality solution than One-shot's solution. These results reveal that the greedy exploration of parameters may not necessarily attain a better quality solution than One-Shot.

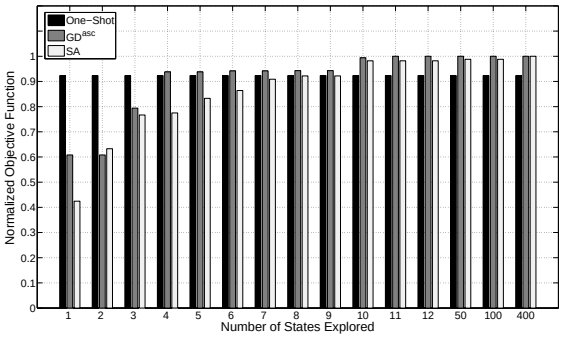


Figure 4. Objective function value normalized to the optimal solution for AC where $\omega_l = 0.4$, $\omega_t = 0.5$, $\omega_r = 0.1$, $|S| = 729$.

Fig. 4 shows the objective function value normalized to the optimal solution versus number of states explored for AC for $|S| = 729$. One-Shot solution is within 7.7% of the optimal solution. The figure shows that GD^{asc} and SA converge to an equivalent or better quality solution than One-Shot solution after exploring 4 states (0.549% of the design space) and 10 states (1.37% of the design space), respectively. These results show that greedy and SA can provide improved results over One-Shot, but require additional state exploration.

3) Computational Complexity: To verify that One-Shot (Section II) is lightweight, we compared the data memory requirements and execution time of One-Shot with greedy- and SA-based dynamic optimization methodologies.

The data memory analysis revealed that One-Shot requires only 150, 188, 248, and 416 bytes for (number of tunable parameters N, number of application metrics m) equal to (3, 2), (3, 3), (6, 3), and (6, 6), respectively. Greedy requires 458, 528, 574, 870, and 886 bytes, whereas SA requires 514, 582, 624, 920, and 936 bytes of storage for $|S| = 8, 81, 729, 31104, 46656$, respectively. The data memory analysis shows that SA has comparatively larger memory requirements than greedy. Our analysis reveals that the data memory requirements for One-Shot increases

 ISBN: 978-1-61208-100-7

linearly as the number of tunable parameters and the number of application metrics increases. The data memory requirements for greedy and SA increases linearly as the number of tunable parameters and tunable values (and thus the design space) increases. The data memory analysis verifies that although One-Shot, greedy, and SA have low data memory requirements (on the order of hundreds of bytes), One-Shot requires 203.94% and 457.94% less memory on average as compared to greedy and SA, respectively.

We measured the execution time for One-Shot, greedy, and SA averaged over 10,000 runs (to smooth any discrepancies in execution time due to operating system overheads) on an Intel Xeon CPU running at 2.66 GHz [19] using the Linux/Unix `time` command [20]. We scaled the execution time to the Atmel ATmega1281 microcontroller [12] running at 8 MHz. Although microcontrollers have different instruction set architectures and scaling does not provide 100% accuracy, scaling enables relative comparisons and provides reasonable runtime estimates. Results showed that One-Shot required 1.66 ms both for $|S| = 729$ and $|S| = 31,104$. Greedy explored 10 states and required 0.887 ms and 1.33 ms on average to converge to the solution for $|S| = 729$ and $|S| = 31,104$, respectively. SA took 2.76 ms and 2.88 ms to explore the first 10 states (to provide a fair comparison with greedy) for $|S| = 729$ and $|S| = 31,104$, respectively. The execution time analysis revealed that our dynamic optimization methodologies required execution times on the order of milliseconds, and One-Shot required 18.325% less execution time on average as compared to greedy and SA. One-Shot required 66.26% and 73.49% less execution time than SA when $|S| = 729$ and $|S| = 31,104$, respectively. These results indicate that the design space cardinality affects the execution time linearly for greedy and SA whereas One-Shot's execution time is affected negligibly by the design space cardinality and hence One-Shot's advantage increases as the design space cardinality increases.

V. Conclusions and Future Work

In this paper, we proposed One-Shot – a dynamic optimization methodology for highly-constrained WSNs that provides a high-quality operation state using intelligent initial tunable parameter value settings. We also proposed an application metric estimation model to estimate high-level metrics from sensor node parameters. This estimation model was leveraged by One-Shot and provided a prototype model for application metric estimation. To evaluate the effectiveness of initial parameter settings, we compared One-Shot's solution quality with four other typical initial parameter settings. Results revealed that the percentage improvement attained by One-Shot over other initial parameter settings was as high as 154.9% and within 5.92% of the optimal solution. Computational complexity analysis revealed that One-Shot used 203.94% and 457.94% less memory and required 18.325% less execution time on average as compared to greedy- and SA-based methodologies. Execution time and data memory analysis confirmed that One-Shot is lightweight and suitable for time-critical or highly constrained applications.

Future work includes incorporating profiling statistics into One-Shot to provide feedback with respect to changing environmental stimuli.

Acknowledgments

This work was supported by the National Science Foundation (CNS-0834080). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

[1] K. Hazelwood and M. Smith, "Managing Bounded Code Caches in Dynamic Binary Optimization Systems," *ACM Trans. on Architecture and Code Optimization*, vol. 3, no. 3, pp. 263–294, Sep. 2006.

[2] S. Hu, M. Valluri, and L. John, "Effective Management of Multiple Configurable Units using Dynamic Optimization," *ACM Trans. on Architecture and Code Optimization*, vol. 3, no. 4, pp. 477–501, Dec. 2006.

[3] S. Patel and S. Lumetta, "rePLay: A Hardware Framework for Dynamic Optimization," *IEEE Trans. on Computers*, vol. 50, no. 6, pp. 590–608, June 2001.

[4] C. Zhang, F. Vahid, and R. Lysecky, "A Self-Tuning Cache Architecture for Embedded Systems," *ACM Trans. on Embedded Computing Systems*, vol. 3, no. 2, pp. 407–425, May 2004.

[5] A. Shenoy, J. Hiner, S. Lysecky, R. Lysecky, and A. Gordon-Ross, "Evaluation of Dynamic Profiling Methodologies for Optimization of Sensor Networks," *IEEE Embedded Systems Letters*, vol. 2, no. 1, pp. 10–13, Mar. 2010.

[6] A. Munir and A. Gordon-Ross, "An MDP-based Application Oriented Optimal Policy for Wireless Sensor Networks," in *Proc. ACM CODES+ISSS'09*, October 2009.

[7] X. Wang and et al., "Distributed Energy Optimization for Target Tracking in Wireless Sensor Networks," *IEEE Trans. on Mobile Computing*, vol. 9, no. 1, pp. 73–86, Jan. 2009.

[8] R. Khanna, H. Liu, and H.-H. Chen, "Dynamic Optimization of Secure Mobile Sensor Networks: A Genetic Algorithm," in *Proc. IEEE ICC'07*, June 2007.

[9] R. Min, T. Furrer, and A. Chandrakasan, "Dynamic Voltage Scaling Techniques for Distributed Microsensor Networks," in *Proc. IEEE WVLSI'00*, April 2000.

[10] L. Yuan and G. Qu, "Design Space Exploration for Energy-Efficient Secure Sensor Network," in *Proc. IEEE ASAP'02*, July 2002.

[11] A. Gordon-Ross, F. Vahid, and N. Dutt, "Fast Configurable-Cache Tuning With a Unified Second-Level Cache," *IEEE Trans. on Very Large Scale Integration (VLSI) Systems*, vol. 17, no. 1, pp. 80–91, Jan. 2009.

[12] Atmel, "ATMEL ATmega1281 Microcontroller with 256K Bytes In-System Programmable Flash," 2010. [Online]. Available: http://www.atmel.com/dyn/resources/prod_documents/2549S.pdf

[13] Crossbow, "MTS/MDA Sensor Board Users Manual," July 2010. [Online]. Available: http://www.xbow.com/

[14] Sensirion, "Datasheet SHT1x (SHT10, SHT11, SHT15) Humidity and Temperature Sensor," July 2010. [Online]. Available: http://www.sensirion.com/

[15] Atmel, "ATMEL AT86RF230 Low Power 2.4 GHz Transceiver for ZigBee, IEEE 802.15.4, 6LoWPAN, RF4CE and ISM Applications," July 2010. [Online]. Available: http://www.atmel.com/dyn/resources/prod_documents/doc5131.pdf

[16] H. Friis, "A Note on a Simple Transmission Formula," *Proc. IRE*, vol. 34, p. 254, 1946.

[17] Crossbow, "Crossbow IRIS Datasheet," July 2010. [Online]. Available: http://www.xbow.com/

[18] I. Akyildiz and et al., "Wireless Sensor Networks: A Survey," *Elsevier Computer Networks*, vol. 38, no. 4, pp. 393–422, Mar. 2002.

[19] "Intel Xeon Processor E5430," July 2010. [Online]. Available: http://processorfinder.intel.com/details.aspx?sSpec=SLANU

[20] "Linux Man Pages," July 2010. [Online]. Available: http://linux.die.net/man/

 ISBN: 978-1-61208-100-7

BeAware: A Framework for Residential Services on Energy Awareness

Christoffer Björkskog[12], Giulio Jacucci[12], Topi Mikkola[3], Massimo Bertoncini[4], Luciano Gamberini[5], Carin Torstensson[6], Tatu Nieminen[7], Luigi Briguglio[4], Pasquale Andriani[4], Giampaolo Fiorentino[4]

[1]Helsinki Institute for Information Technology HIIT, Aalto University, Finland, name.surname@hiit.fi, [2]Department of Computer Science, University of Helsinki, Helsinki, Finland, name.surname@cs.helsinki.fi, [3]BaseN, Helsinki, Finland, name.surname@basen.net, [4]Engineering Ingegneria Informatica, Rome, Italy, name.surname@eng.it, [5]HTLab, Department of General Psychology, University of Padova, Padova, Italy, name.surname@unipd.it, [6]The Interactive Institute, Sweden, name.surname@tii, [7]Department of Electrical Engineering, Aalto University School of Science and Technology, Helsinki, Finland, name.surname@tkk.fi

Abstract - **Lately a wide variety of products and services are emerging to address energy efficiency in households by providing timely information. This area is still relatively young in terms of adoption of advanced ICT still exploring what functionality and features should be part of such services. Here we propose a framework for residential energy awareness services that divides technologies in three layers: sensing, services and applications. First we review expected features of emerging products and services also presenting a service and user application we implemented to exemplify the functionality. Each layer of the BeAware framework is presented by discussing the objective, composition and current challenges.**

Keywords- Ubiquitous services; Energy Awareness; Monitoring sensing.

I. INTRODUCTION

There are generally three methods available today for a homeowner to access his or her energy consumption information. First, consumption data can be obtained from their energy bill. A second method is through reading his or her own energy meter installed in the house, and lastly by purchasing an external energy meter device. A report conducted from the Centre for Sustainable Energy stated that most homeowners are only looking at the amount of money they owe and ignore the rest of the data presented on their energy bill [1]. Although the consumption figures and tariffs are stated, they are often ignored so consumers are still unaware of their consumption habits.

The available information to energy consumers in household is soon to be expanded by a wide variety of products and services. These prove timely consumption information that allows consumers to track the energy consumption in their household. The services range from web-services and sites to hardware and software products.

The area is still relatively young and is exploring what functionalities and features should be part of such services. A wide diversity of energy conservation products and services exist that in different ways include hardware, methods for analysis and interface components. We maintain that it is needed to discuss to what extent different functionalities are needed and to propose frameworks that aid the implementation of services. We propose a framework for residential services providing energy awareness than is divided in three layers: sensing platform, web services, and user applications. The framework is based on a concrete implementation is already running in 15 sites and with over 50 users across three countries.

In the following section, we review expected features of emerging products and services. In Section III, we present an example user application "Energylife". In Section IV, we present each layer of the BeAware framework discussing objectives, composition and current challenges. We end the paper with discussion and conclusions.

II. RELATED WORK

In this section, we review expected features of emerging products and services. First, we examine Web-based solutions and home displays, then we take a look at real-time plug meters and lastly we discuss ambient interfaces.

A. Web-based solutions and home displays

A method to provide consumption feedback to users today is via personal websites where users can log in that electricity companies provide. The key feature of the web-based solutions is that they provide aggregated overall data on the consumption of the home and this data is visualized with a histogram to show when consumption is low or high. The most advanced concepts of such systems provide real-time information on the consumption, user configurable alarms, billing information, payment services, details by floor, room, appliance, circuit, or utility and they can even provide automatic calculation of carbon footprint and have trend analysis and historical comparatives. However, these advanced visualizations are mostly at conceptual or trial stage. An example of a web-based user interface to monitor electricity consumption is from Agilewaves or Greenbox [2]: a web-based solution that enables households to track, understand and manage their home energy usage and environmental footprint. Greenbox automatically categorizes electricity usage and allows users to compare their own usage anonymously with other homes. Other web-based solutions are TREE – Tendril Residential Energy Ecosystem [3]. TREE is a solution that connects "smart" consumer devices (like thermostats and outlets) to the existing utility back office and establishes a dialog between consumers and their energy providers. The in-home wireless network allows

 ISBN: 978-1-61208-100-7

TABLE I: ANALYSIS OF FEATURES OFFERED BY SOME AVAILABLE ENERGY AWARENESS PRODUCTS COMPARED TO OUR RESEARCH PROTOTYPE ENERGYLIFE IMPLEMENTED IN THE BEAWARE FRAMEWORK

See references [2],[3],[6]-[8] and [15]-[24]		Manodo	Greenbox	Tendril TREE	Google Pow.	Innohome	Homemanag.	PowerCost M.	Ewgeco	BaseN Beat	CC Envy	efergy meter	The OWL	Eco-eye	Wattson	Energy hub	Home Joule	**EnergyLife**
Consumption	Real time	x	x	x	x	x		x	x	x	x	x	x	x	x	x	x	**x**
	Over time	x	x	x	x	x		x	x	x				x	x	x		**x**
Control							x					x				x		**x**
Information can be actual or exemplary	Actual	x	x	x	x	x		x	x	x	x	x	x	x	x	x		**x**
	Typical		x															
The information can be about a device or whole household, or broken down to all devices	Device	x									x	x	x		x	x		**x**
	Household		x	x	x	x		x	x	x	x		x	x				**x**
	Breakdown																	**x**
Medium or interface can range between, web, dedicated display mobile or ambient	Web/PC	x	x	x	x							x			x			**x**
	Mobile	x				x												**x**
	Display	x						x	x		x	x	x	x	x	x		
	Ambient														x			**x**
Format of the information in power measure, monetary or environmental impact	W or Wh		x	x		x		x	x	x	x	x	x	x		x	x	**x**
	Monetary €		x	x	x	x		x	x	x	x	x	x	x				
	CO2		x	x					x				x					
Advanced User Applications such as awareness features beyond consumption data	Tips, quiz																	**x**
	Community																	**x**
	Games																	**x**

appliances and electrical outlets communicate to a home energy monitor the energy consumption.

B. Real-time plug meters

To measure energy used by individual appliances, homeowners can purchase plug-in electricity meters. These devices are placed between the power inlet and a home appliance and provide both real-time readout of the electricity consumption (kW) and measures consumption over time (kW/h). They can aid the user by identifying major energy consuming thieves as well as devices that consume much standby electricity. Several studies show that measures such as better billing, smart metering and feedback on energy consumption can encourage households to save energy, about an average of 5-10% [4]. In addition, a recent public survey [5] conducted by Future Foundation in the UK found that 82% of the respondents would consider changing their energy consumption habits if they had a "screen telling the homeowner how much energy they are using at any one moment". Therefore, the problem may not be the technology but the poorly presented information that limits consumers to understand the impact of their behavior and sustain awareness of their energy consumption. Some of these devices focus on safety since they can prevent fires and malfunctioning of appliances besides monitoring their consumption. An example is the Innohome Guard [6], which shows data also on a mobile phone rather then a dedicated screen/device. Innohome is a device that can be attached to the socket of an individual appliance. Separate devices can send data through the power line to the Innohome Base Station device, which can send all the collected data from the house through the Internet and show it on the end user's mobile phone. Features also include shut down control action for individual appliances as a mean for safety against fires originating from faulty usage of electric appliances.

C. Ambient Interfaces

Ambient interfaces address the periphery of user attention by embedding information into the objects that surround us. These interfaces provide the user with information in the form of sound, air pressure, motion, light, smell, and other media that complement the full range of our human sensory modalities. They are designed to work with our peripheral senses, where they provide continuous information without being distracting or obtrusive. Notable examples are Watson and the Ambient Orb. Watson is an energy-monitoring device that has been praised for its style and simplicity [7]. By clamping an external wireless sensor to the home energy supply (mains fuse box), the sleek device shows the running total in real-time of the wattage output. This is represented in a digital readout on the Wattson display or with ambient light. The Ambient Orb [8] is a frosted-glass ball that slowly shift between thousands of colors to show changes in the weather, stock market trends, electricity usage or any other ambient information channel (chosen by the user). It can also flash when important information is received or if a special threshold has been reached.

 ISBN: 978-1-61208-100-7

In our work we have used the Table 1 as a starting point to identify basic features to be addressed.

III. EXAMPLE OF APPLICATION : ENERGYLIFE

In this section we present an example application called Energylife. Energylife is an engaging informational game-like application that is designed to raise awareness on energy conservation and help users to embrace a sustainable energy conservation lifestyle. It makes the users aware of their energy consumption and gives information regarding the electricity consumed by the whole household and certain appliances. It also lets the users understand whether or not they are saving energy compared to their normal consumption and how much above or under it they are.

Inside the households the system consists of a base station, sensors connected to appliances and main fuse box and iPhones that present the application to the household members (See Figure 2). The system also includes servers that process the measured data and delivers services to the phones in order to present and visualize the data.

Users can track their consumption history and get an overview on what appliances are saving electricity and which ones are consuming.

Every day users receive advice on energy saving and every third day they are presented with a quiz where they can choose what would be the best way of conserving energy in different situations. The application behaves as a game where users can reach new levels by gaining enough points. Saving points are gathered by saving electricity and you can get awareness points by answering correct on quiz, reading advice and participating in the community.

IV. A LAYERED APPROACH :SENSING, SERVICE AND APPLICATION

Three layers compose the BeAware Framework. A sensing layer to collect next-to real-time information on detailed consumption of energy; A service layer providing web services as general level functionality in managing data from consumptions, users, and knowledge database; finally a user application layer providing applications and interfaces on selected smart phones and newer web browsers (See Figure 1).

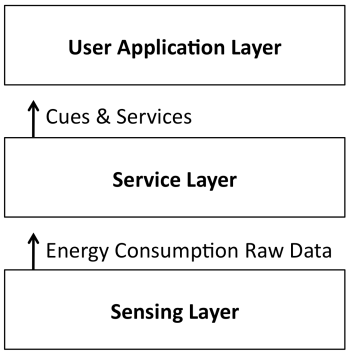


Figure 1. The overall organization of the layers

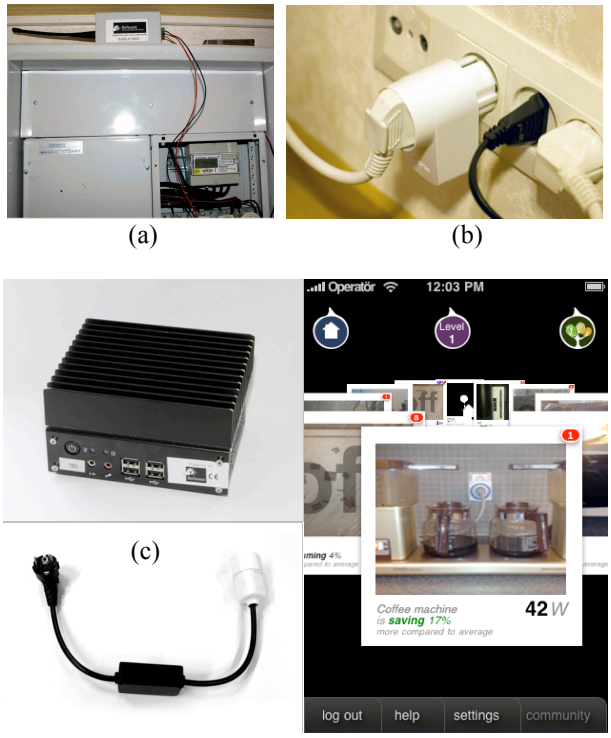


Figure 2. The hardware in the household consist of a sensor measuring the main meter in the fuse box (a), sensors that measure appliances; either off the shelf sensors (a) or custom made sensors (d), a base station (c) that receives the data from the sensors, controls the ambient interface and delivers data to the data storage. Next to real time consumption data is visible in the client user interface (e).

A. *Application Layer*

The Application Layer provides the end user functionality. The web application communicates asynchronously with the service layer and dynamically updates its user interface.

The ambient interface consists of electronics that connect to the existing lighting of a household. It communicates wirelessly with the base station in the household. The ambient interface gets the required data to function as intended from the base station.

The mobile web application it is built up using JavaScript, HTML5 and CSS3 (See Figure 2e). Client side JavaScript running in the web browser builds the user interface and communicates with the Web services provided by the Service Layer.

1) *Interface to Service Layer*

The application retrieves needed data from service layer and utilizes the twitter platform for its community part. Data from is fetched via XMLHttpRequest calls delivering the data in JSON format. Calls fetching the current consumption and state of devices a are made in short intervals in order to provide a next to real time experience.

Interaction with the service layer is done by executing service methods on a JavaScript object provided by the JAX_WS framework. These objects have access to all the public methods provided by its web service in the service

 ISBN: 978-1-61208-100-7

layer. Access to such an object is obtained by including a JavaScript tag in the header of the HTML document that points to the web service.

The application layer needed a way to effectively communicate with the services provided by the Service layer and the users. It was chosen to build the application as an iPhone web app where both the benefits from having a touch screen would be taken advantage of and also that it would be easy to port to different target devices without having to build the application again from scratch.

The Service Layer exposes their services via Java API for XML Web Services (JAX-WS) that you can communicate with using different technologies. At first we thought we needed to build some kind of server side middleware in order to allow easy JavaScript interaction with the services. We noticed however that there are JSON bindings for JAX-WS that allows you to interact with the web services directly via JavaScript. The framework allows you to point a script tag to the exposed web service, which outputs a JavaScript object. This object provides all the methods exposed by the web service.

Example usage:

```
<script src="/consumption-manager-
ws/ConsumptionManager?js&var=ConsumptionManager">
</script>
```

As an example, the previous code is included in the head tag. This script supplies you with a similarly named object (consumptionManager) that has all the methods that the web service supplies. You can call these methods by supplying an object as parameter with attributes named as the variables in the web service method, and a callback function to be executed once the request is done.

Some problems were encountered with this approach. We found out that the project for developing the JSON extension had been discontinued, and some bugs in the framework was found. For instance, if the output is defined to be an array, but the array only contains a single item, the single item will be outputted instead of an array containing one item. To remedy this we needed to have client side checks to see if the returned element is an array or not.

2) Scripts and resources

As the application is web based running in a browser, it requires resources from the server in form of documents, JavaScript and CSS files and images. To support tailored versions for specific devices the file structure on the server for each resource type is divided into general resources and platform specific (for instance iPhone or desktop). JavaScript, CSS and PHP library files are first included from the general repository and then from the platform specific repository. Due to large amounts of files that were needed, each JavaScript and CSS batch are automatically combined and minified to speed up loading and reduce the amount of HTTP calls.

B. Service Layer

The rationale beyond the Service Layer is to build a reference web services based middleware for residential energy awareness and conservation. It provides an open and distributed web service infrastructure (i.e., based on JAX-WS JSON binding) that can be plugged to any sensing back-end. The basic feature of the Service Layer is the acquisition of raw data from the plugged sensing platform and the generation of energy-related information in an understandable format, according to residential household consumers. The service layer aims to represent the natural bridge between a sensing platform and a consumer application built on top of it. For instance, in the BeAware projects Service Layer interacts with the sensing layer through the data store client using Google Protocol Buffer (GPB).

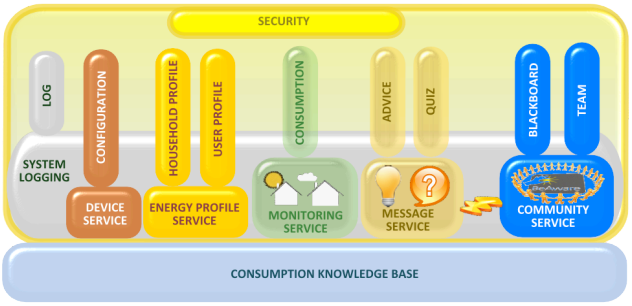


Figure 3. Service Layer

Hereafter is more detailed description of Service Layer components (See Figure 3) where a more comprehensible rational is given for each of them.

A common layer for each offered service has been designed. The Consumption knowledge base is a key part of the BeAware Service Layer and provides to all the services the basic information model and concepts of energy consumption (e.g., historical consumption, over consuming or saving, power consumption, state of the appliance, etc).

Device Service offers information about available sensors deployed in the household and configuring them as household's appliances in a more fashionable way according to user preferences.

Energy Profile Service allows the application layer to set and obtain information regarding the household (e.g., location, rooms, inhabitants, type of building, etc) and the users (accounts, credentials, age, language, etc.). This information is useful for characterizing and classifying the energy consumption data.

Monitoring Service provides a set of information extracted by the raw data of the sensing layer as far as next-to-real time power consumption and historical consumption (both for the whole household and for the specified appliances), state of the appliance, overconsumption or saving for a specified period.

Message Service is a notification system whose objective is to improve awareness on energy consumption. Message Service includes delivering of quiz, advice and real time tips in order to lead the user towards a more understandable and responsible way of using its appliances. Message Service supports multiple languages according to the profile of the household and its users. The Message Service also calculates an Awareness Score that has been modeled in order to offer an index of the users' awareness about their energy behavior.

Community Service receives notifications when the user has posted a message in the community. This is used to update the community part of the Awareness Score. In the

 ISBN: 978-1-61208-100-7

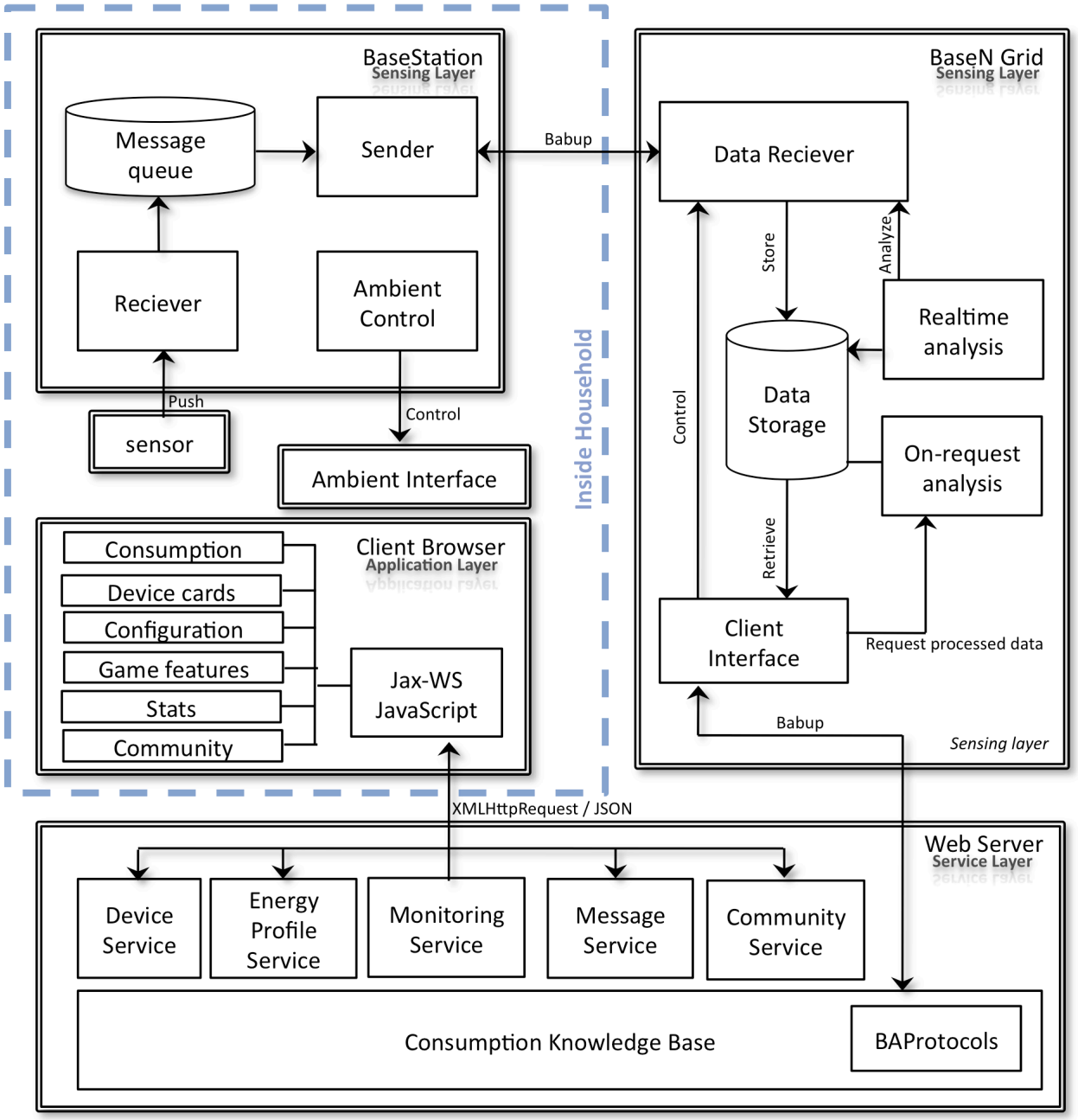


Figure 4. The BeAware Architecture

current version the community itself is built on top of Twitter.

System logging is a background component that is in charge of tracking the whole Service Layer behavior. It receives a wide range of input going from the specific sensor's status to a more detailed feedback on Service Layer usage.

C. *Sensing Layer*

The goal of sensing layer is to provide a deployable measuring system, which is able to break down the consumption of the whole household, monitor it in real time and also effecting the ambient interface connected to particular lights. The architecture should be able to scale.

The sensing layer consists of two separate parts, the physical sensing infrastructure installed in households, and distributed data storage and analysis platform for collecting measurements from all sources (See Figure 4). While one household installation should be able to handle hundreds of discrete measurements per minute (the full BeAware installation is 10 sensors, measuring 9 values once per 2 second, resulting in roughly 400kbit/min uncompressed dataflow), the whole system must be able to handle thousands to millions of such installations, meaning that the whole system must be able to cope with millions of measurements per minute even in moderate sized installation. Inside the household the system must also be as

 ISBN: 978-1-61208-100-7

energy efficient as possible not to impact the household power usage.

The system has been designed as open from the start, allowing for easy interfacing with any 3rd party product, which has a documented logging interface. Currently pulse counters for water, heating and older electricity meters, and Plugwise [9] wireless energy sensors are supported as prototype examples. The system must also be easily updateable to support new sensors - this means that it should support both remote updating, and keep the household sensing system as simple as possible, while keeping the analysis capability in the storage platform. It also allows customer to use as much of old sensor equipment as possible, as quite a few households have at least some energy/environment measurements available.

Inside the household the wireless communication is in a low level byte coded protocol called BASP (BeAware Sensor protocol) while the communication between the base station and storage (plus outwards from the storage) is based on the Google Protocol Buffers transported over HTTP. Compared to normally used XML-based approaches, this gives far better signal/overhead -ratio, allows data marshalling/ unmarshalling to be much lighter and makes 3rd party integration easier. Conceptually, each measurement type from sensors is modeled as a channel containing time series of values. This allows for easier hierarchical grouping and modeling of data.

1) Sensing Infrastructure

The household sensing infrastructure includes various wireless sensors, ambient interfaces and a base station for sensor and actuator control. The design goal is to allow user to use what sensors they have available while also providing a state of the art approach for new installations.

Any Linux supporting computer can be used as a BeAware base station. A dedicated low-power passively cooled system is probably the optimal, but the system has been demonstrated to work with many types of servers, from remote virtual server over Ethernet-USB adapters to local GPRS based microcomputers. The base station will cache measurements in non-volatile memory and transfer them to the data storage in bursts. This allows the system to survive network outages with no data loss and reduces the transfer overheads. The base station supports both BeAware's own 433 MHz wireless network and also allows for interfacing with commercial ZigBee based networks. A wireless connection is provided by radio transceiver over ISM band of 433 MHz. Radio chip handles receiving and transmitting automatically on microcontroller input and provides interrupt signals for message and signal handling. Data is sent Manchester coded over air and the radio chip provides automatically preamble for radio channel synchronization. The chip operates both transmitter and receiver operations. The BeAware sensor is meant to be a low-power, low-cost wireless energy sensor. The sensor takes its' power from the line being measured and needs no external source. It does not have any electrical connection to protective earth, so it does not affect protective devices and can be installed to either normal socket or to main fuse box to read the whole household consumption. Unlike normal commercial solutions, BeAware sensor measures multiple variables[1] in addition to active power and energy allowing sensor to be used to load fingerprinting.

Initially the household measuring system will consist of 9 such sensors, while the long term goal is to see, if fine enough granularity can be achieved with only 3 BeAware sensors monitoring the main phases of household electricity. This would need a system to be able to detect individual appliances via load fingerprinting but would allow us to lower the amount of measurements needed over the whole household significantly.

An Ambient interface combines a normal BeAware sensor with the ability to control lights that are connected to it. It is used to signal to users whether household is saving energy compared to baselines and to tell users if any of the alerting conditions they have set have been triggered. It provides a non-intrusive interface, which provides necessary information to user without need for checking any external system and also provides cues for user to check EnergyLife for more detailed information.

Main challenges encountered in the sensing infrastructure have been to create a system, which is as easily installable as possible – the system should need no configuration to install and should be as easily serviced as possible. Also, finding an off the rack reliable mini server with suitably low power consumption has been harder than initially anticipated. The current generation of pico-computers have either had problems with overheating, general stability or consume over 15W.

2) Data Storage

*T*he data storage is based on the BaseN [10] Platform computation grid, which has been optimized for receiving and analyzing large amounts of real time data measurements. It currently handles millions of measurements per minute, and with BeAware the target is to push the measurement interval from once per minute closer to one per second. The platform consists of data receiver agents, storage units, and real time, on-request analysis services. Data receiver handles communication from households, caching data and feeding it to both long-term storage and real time analysis services. The real time analysis systems allow a user to receive alerts on trigger conditions within one minute of a performed measurement. On-request analysis provides historical data and visualization services to the user. The platform is also able to provide open interfaces to other measurement sources, such as other energy meters, environmental monitors, industrial sensors, and general information systems. The platform also monitors itself and other parts of the BeAware system for service consistency and outages.

The current storage platform has been tested with a steady flow of roughly 5 million measurements per minute

[1] These variables are: Root Mean Square (RMS) value of current, RMS of voltage, Active power, Apparent power, Power factor, Phase shift, Energy,Crest factor, Harmonics and Total harmonic distortion

 ISBN: 978-1-61208-100-7

load in telecom applications. With the current design the data receival can easily be scaled to accept tenfold more, but as with any interactive system, load with peak concurrent interactive users is more problematic, as much more data analysis is needed and most users want to see data from several sensors simultaneously. Either systems must keep considerable resources on stand by for possible load peaks, or accept that performance will momentarily degrade in peak usage situations, or the better load handling algorithms are needed.

V. DISCUSSION AND CONCLUSIONS

Recently research on ubiquitous computing for energy efficiency in the household has received attention. Some of the latest advancements include: 1) visualization of detailed data attained through pervasive sensing [11] 2) aesthetic displays using novel interfaces [12], 3) theoretically informed implementation of feedback that address behavior change [13].

As we have seen from the review of Table 1, products and services exists all addressing partial functionality. Generally solutions tend to track only a single device or the total consumption of the household. Solutions that comprehensively give a breakdown of consumption in the household are unaddressed. The user applications are simply about displaying consumption in different formats. Most of the solutions utilize a dedicated display device.

As an example recent research work on connecting energy meters with mobile phones such as [14] shows that research would benefit from a more comprehensive approach in detailing requirements and solutions in a framework for future services.

We have presented EnergyLife an energy awareness application that addresses more features than current available solutions. We then presented a framework to deploy EnergyLife services that is composed by three layers. This has served to discuss requirements of hardware and software in an approach that can help implement services.

The BeAware framework is being deployed in trials in 15 sites including 12 households and 3 laboratories, including 50 users and 120 sensors. While the data collection is ongoing preliminary findings and experiences indicate that the decomposition is the three layers has been very useful to support interoperability and independent development. As an example, physical sensors can change leaving the rest of the system unchanged. Similarly, new applications can be developed without affecting the lower layers. To be able to fully take advantage of the layers, each layer has a monitoring interface to debug and identify issues.

REFERENCES

[1] S. Roberts, H Humphries, and V. Hyldon “Consumer preferences for improving energy consumption feedback”. Report for Ofgem. Centre for Sustainable Energy, Bristol, UK. 2004.

[2] Greenbox Technologies. Retrieved July 19, 2010, from http://www.getgreenbox.com/public/contact.

[3] Tree. Retrieved July 19, 2010, from http://www.tendrilinc.com/utilities/utility-products/

[4] S. Darby “The Effectiveness of Feedback on Energy Consumption – A Review for Defra of the Literature on Metering, Billing and Direct Displays”. Environmental Change Institute, Oxford University. 2006.

[5] Future Foundation “Energy Efficiency – Public attitude, private action”, LogicaCMG, 2006, p. 21.

[6] Innohome. Retrieved July 19, 2010, from http://www.innohome.com/

[7] Diykyoto. Retrieved July 19, 2010, from www.diykyoto.com/

[8] http://www.ambientdevices.com/products/energyjoule.html

[9] Plugwise. Retrieved July 19, 2010, from http://plugwise.nl

[10] BaseN. Retrieved July 19, 2010, from https://www.basen.net/corporate/#Energy

[11] Y.Kim, T. Schmid, Z. M. Charbiwala, and M. B. Srivastava “ViridiScope: design and implementation of a fine grained power monitoring system for homes”, Ubicomp '09, ACM Press, 2009, pp. 245-254.

[12] A. Gustafsson and M. Gyllenswärd, “The Power Aware cord: energy awareness through ambient information display”. In Ext. Abs. CHI '05, ACM Press, 2009, pp. 1423-1426.

[13] T. Ueno, R. Inada, O. Saeki, and K Tsuji, “Effectiveness of an energy-consumption information system for residential buildings”. Applied Energy, 83, 2006, 868-883.

[14] M. Weiss, F. Mattern, T. Graml, T. Staake, and E. Fleisch, “Handy feedback: Connecting smart meters with mobile phones”. In: MUM2009, Proceedings of the 8th International Conference on Mobile and Ubiquitous Multimedia, ACM Press, 2009, pp. 1-4.

[15] Manodo. Retrieved July 19, 2010, from http://en.manodo.se

[16] Home Manageables. Retrieved July 19, 2010, from http://www.homemanageables.com

[17] Google Power Meter. Retrieved July 19, 2010, from http://www.google.com/powermeter

[18] Energy Monitor. Retrieved July 19, 2010, from http://www.bluelineinnovations.com/Products/

[19] Ewgeco. Retrieved July 19, 2010, from http://www.ewgeco.com/

[20] Current Cost. Retrieved July 19, 2010, from http://www.currentcost.com

[21] Efergy Meter. Retrieved July 19, 2010, from http://www.efergy.com/

[22] Owl. Retrieved July 19, 2010, from http://www.theowl.com/

[23] Eco-eye. Retrieved July 19, 2010, from http://www.eco-eye.com/

[24] EnergyHub. Retrieved July 19, 2010, from http://www.energyhub.com/

 ISBN: 978-1-61208-100-7

Acceptance Models for the Analysis of RFID

Markus Haushahn, Michael Amberg,
Armin Schwingenschlögel, Florian Bernhardt
Chair of Information Systems III
University of Erlangen-Nuremberg
Lange Gasse 20, 90403 Nuremberg, Germany
markus.haushahn@wiso.uni-erlangen.de

Krzysztof Malowaniec
Business Development
DATEV eG
Paumgartnerstr. 6-14, 90329 Nuremberg, Germany
krzysztof.malowaniec@datev.de

***Abstract*—Although there is a high dispersion of RFID in many areas of the economy, it can be said that up until now this technology has been barely implemented and accepted within law firms. Considering the severe problems when tracking documents and although these systems facilitate specific improvement in various sub processes handling legal cases, lawyers are still mostly disapproving of the use of RFID. Therefore, the motivation as well as the acceptance indicators of lawyers, which are responsible for such behavior, need to be observed. Within the scope of this article the currently valid acceptance models will be analyzed regarding their applicability to RFID-Systems and their possible application in law firms. Furthermore, both the user's point of view as well as the involved IT technology shall be considered within this evaluation. This is the only way to ultimately ensure the achievement of pursued objectives such as increasing customer and employee satisfaction, optimizing internal processes as well as continuously improving business results within law firms. The primary result of this analysis shows that the DART acceptance model by Amberg and Wehrmann explains best all eight of the RFID-relevant acceptance levels such as the psychological or the task-related level for instance.**

Keywords: acceptance analysis; RFID; ubiquitous computing

I. INTRODUCTION

In recent years RFID-technologies have not only caused quite a stir in science but also in many areas of the service sector, purchase and outbound logistics, the industry as well as in manufacturing companies. Hundreds of companies that are actively involved in the development and sale of RFID systems indicate that this market is taken very seriously. While global sales of RFID systems reached about 1.2 billion U.S. dollars in 2008, the forecast for 2012 predicts a sales growth of 3.5 billion U.S. $ [1]. The market for RFID is therefore one of the fastest growing sectors in the industry of radio technology. In spite of the obvious progress and the expected efficiency and cost potentials, the diffusion rate and therefore the implementation in many sectors as well as in companies is still to be seen as a niche solution [2].

Enhancement in productivity and efficiency are not only practicable in the field of supply chain management but also in many areas of the service sector in terms of a cross-sectional technology [3]. A particular setting for the application of RFID-Systems is the tracking of documents and the administration of books within the scope of the document management. Especially in the day-to-day handling of documents and books, companies gather a significant amount of data. These vast amounts of documents are often stored in boxes, folders or cabinets and filed in special rooms. In order to find and process stored data additional costs emerge for the company and cause extra time expenses for the employees. Furthermore, this loss of efficiency leads to a waste of human resources and employee productivity decreases [4].

Even today the handling of the so-called "paper files" is still required by law, particularly in law firms and tax attorney offices. According to §50 of the Federal Code for the Legal Profession, it is a lawyer's responsibility to give an orderly insight into his work by creating reference files [5]. However, this legally demanded system is being affected significantly by the in some extent complex procedure of processing a file. If one analyzes the working process of an attorney, this problem can clearly be seen. Depending on the complexity of a lawsuit there are up to eight people sequentially working on one case. Thereby the paper card changes the staff member up to 26 times on average.

If one combines the complexity and diversity of the processing steps in a lawsuit with the number of cases a lawyer has to work on per day, it is clear that a single paper file might get lost easily. However, it is mandatory to have a hardcopy of the document while talking to clients in the office or being in court in order to ensure the possibility of making changes at all times and having a successful legal dispute. Therefore, the loss of a file would be linked to far-reaching consequences for the office and for the client. Particularly affected by this problem are law firms with more than 20 lawyers, which are distributed on different floors and buildings. If at least 5 files a week go missing in a law firm with 20 lawyers and a stock of 700 cases, and the average search time equals 1.5 hours per file, the consequent time spend on searching is at least 7.5 hours a week, which is a serious problem for the efficient work cycle in this office.[1] However, with the help of RFID as a cross-sectional technology, it is possible to improve the workflow of a lawyer and thus the handling of paper files as well as legal texts considerably.

[1] The numbers result from a process analysis carried out for a law firm in Munich.

 ISBN: 978-1-61208-100-7

Although many law firms are aware of these problems and lawyers know that RFID technology could eliminate these deficits, they are still not willing to invest in an innovative technology. The goal of this paper is therefore to analyze the acceptance models for RFID systems in law firms and to identify the factors, which are able to describe the acceptance of RFID technology.

Within the scope of the pilot project "RFID in lawyer's offices", the phenomenon of RFID-technologies in the daily use is ought to be analyzed scientifically. Therefore, its distribution in both the literature and in practice, depending on the costs and benefits, need to be assessed in order to assure a holistic implementation with the help of practitioners in the next step. The evaluation of the acceptance is an essential component next to the creation of a business case. This evaluation of the theoretical acceptance models illustrates the first step of the analysis of RFID in lawyer's offices. Although acceptance models offer diverse possibilities, they only fit partly into the scenario RFID in lawyer's offices, which is the reason why this evaluation is being carried out. Therefore, the following analysis is ought to indicate to what extent the selected models fit the scenario RFID in lawyer's offices and thus which practice seems the most suitable.

At the beginning of this paper it is ought to describe the technology of RFID and the main procedure applied in order to carry out the analysis. Thereupon, individual dimensions and indicators are being described and evaluated by experts. In the end, these results are the basis for the decision to use the Dart Model according to Wehrmann**.**

A. Research Design

The scientific background was provided by a comprehensive literature research being the preliminary stage of the acceptance research area and RFID. **Therefore, 556 articles of the IEEE and 137 German and English books were studied regarding these models. Within the scope of this procedure 10 acceptance models were identified due to their number of mentions.** The main objective, however, is the evaluation of the user acceptance in law firms, which are supported by RFID. The purpose is to create a general understanding for those areas involving RFID and research acceptance. Based on this knowledge it is ought to identify, outline, and monitor existing acceptance models on whether they are suitable for the evaluation of the user acceptance of RFID in law firms. Besides, both the user's point of view and the underlying IT technology are to be considered within this evaluation and review. **Thus, the second step of the explorative study involved an execution of two workshops. A total of fourteen experts participated in these workshops, which were held half-day in March and April 2010. Four of these participants were employees of the law firm and they were working with this technology constantly, while there were three experts out of science, two experts out of user's offices, and three experts out of the hard- and software industry of RFID-systems (see Figure 1). These persons were chosen as experts since they had a long-time experience and thus a wide knowledge of RFID. Besides, they were trained by using this technology directly within lawyer's offices. In order to ensure an equal knowledge of the participants and a successful design of the workshops, the required documents were sent out about a week prior to the meeting. Within the scope of the workshop a set of questions was assessed. One of them dealt with the expert's opinion on how well these models covered the different factors of the individual levels such as the social level, for instance. Therefore, both the interviews as well as the related discourses within the project group made it possible to carry out an evaluation of the individual models.** In doing so, individual models were presented, discussed several times, and indicators were chosen and improved. Afterwards the acceptance models were evaluated according to the previously identified indicators. While developing the results, there were three questions being focused on:

1. Which acceptance models are available in the literature?
2. Which of these models suits an analysis of RFID in law firms the best?
3. Which indicators need to be considered by the acceptance model?

The objective of this approach is to generate an adapted model of acceptance, which possesses those indicators adjusted to RFID systems in order to generate an acceptance analysis in a pilot office.

Research Design	Explorative Study	
Iteration	1	2
Research Method	Draft design for the selection of an adequate acceptance model	2 workshops for the selection of an adequate acceptance model
Duration	January- February 10	March-April 10
Number	556 articles (IEEE) & 137 books	14 experts out of science, industry, and user's area

Figure1. Procedure while analyzing the acceptance of RFID in law firms

B. RFID technology

The abbreviation RFID means Radio Frequency Identification, which could be translated via radio waves for identification. RFID is also commonly described as an automatic identification and data capture system with contactless transmission of data between an RFID tag and RFID reader based on radio frequency technology. If products, pallets, truckloads or documents are being equipped with RFID tags, they can give a feedback signal on their position, motion or texture automatically [6].

RFID has been used for several years and in some areas it is already seen as an important part of the process management. It has established itself, particularly in the area

 ISBN: 978-1-61208-100-7

of production, logistics, theft protection and access security [7]. Recently this technology has gained a foothold in other areas as well. At the moment there are strong efforts in establishing itself in the medical and nutritional area as well as in the field of document management. Since the advantages and disadvantages of this technology have been analyzed more than once in the literature and public, this paper will omit to discuss this matter in detail.

C. Acceptance Models

In recent years, there has been a development towards a new understanding of acceptance research, which is also referred to as the "recent acceptance research" by Kollmann. The following features can illustrate this new perception. [8] p. 149f.

Timeframe: As mentioned previously, the acceptance research is seen traditionally as a study object of various areas of science, which observes the "acceptance" isolated from the rest. This separation is currently not applicable in terms of innovative applications since most technologies (including RFID) are able to establish themselves not only in organizations but also in private households. An integrative perspective counteracts with this isolated observation and summarizes all of the critical factors of the various disciplines [8] p. 149.

Decision criterion: The classic dichotomy of acceptance decision cannot be transferred on innovative applications in general. Especially when it comes to innovative utilization, an acceptance continuum has to be acknowledged. Therefore, business informatics rather prefer gradations between different acceptance levels than a dichotomous notion, since it is being focused on the utilization of the innovation [9].

Utilization motivation: Since products of information technology (such as PCs, notebooks, netbooks, or PDA's), telecommunications (mobile phone, smartphone) and multimedia communications can be used both due to organizational requirements as well as on a voluntary basis, a strict separation of organizational use or voluntary use are not longer appropriate. Therefore, acceptance research and used models must take into account that innovative products have reached both organizations and private households [8] p. 149.

Objective: The traditional point of view considered the acceptance only ex-post. Thus, this approach was used primarily to generate appropriate marketing strategies for already established products and services. As a result no action has been taken in order to detect product deficits and counteract previously to launching. An ex-ante analysis, however, makes it possible to analyze the acceptance at an early stage in order to carry out formative and corrective actions. Therefore, an acceptance analysis, which carries out both, an ex-ante and ex-post analysis, is desirable.

A great amount of acceptance models were developed over the last years in order to analyze the acceptance. Since different research areas have arisen, various assumptions are being made on the main aspects or influential factors, which have a considerable impact on the acceptance [8] p. 150. A short overview of the acceptance models dealt with in this paper is being given in Figure 2. It shows the development of the ten models over time. As a result of the analysis, it could be observed that these models have been specified even more over the periods of time and thus are able to fulfill the demands and application scenarios needed today.

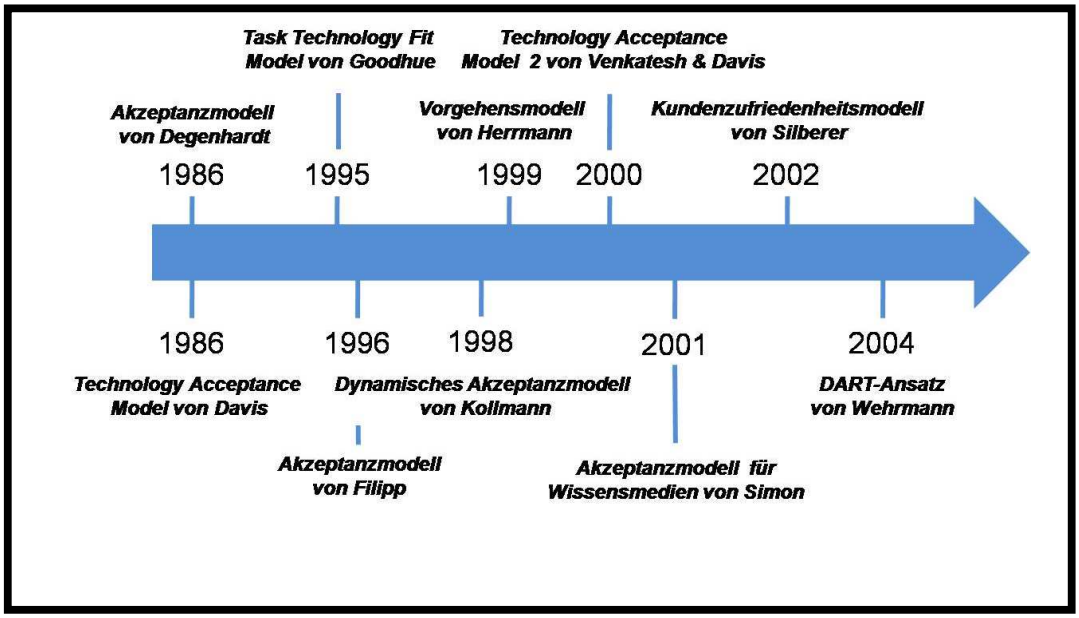


Figure 2. Overview of the relevant acceptance models

 ISBN: 978-1-61208-100-7

Acceptance Models	Influencing Factors/Dimension	Alternative Model
TAM (Davis)	- Perceived benefits - Perceived easy handling	**Input-/Output Model**
Acceptance Model (Degenhardt)	- Task characteristics - System configuration - User characteristics	**Input-/Output Model**
TIFM (Goodhue)	- Tasks - Technology - Individual	**Input-/Output Model**
Acceptance Model (Filipp)	- Organizational environment - User - Technology system (content & user guidance)	**Feedback Model**
Dynamic Acceptance Model (Kollmann)	- Product-related determinants - Consumer-related determinants - Company-related determinants - Environment-related determinants	**Input-/Output Model**
Procedure Model (Herrmann)	- Global criteria checklist	**Feedback Model**
TAM 2 (Venkatesh & Davis)	- Perceived benefits - Perceived easy handling • Subjective standard • Image • Job relevance • Output quality • Traceability of the results	**Input-/Output Model**
Acceptance Model for Knowledge Media (Simon)	- Knowledge media design - User design	**Feedback Model**
Customer Satisfaction Model (Silberer et al.)	- Hardware - Transmission costs - Mobile commerce applications	**Input-/Output Model**
DART Approach (Wehrmann)	- Perceived benefits - Perceived usability - Perceived costs - Perceived amplified benefits	**Input-/Output Model with Feedback Character**

Figure 3. Influencing factors and dimensions of relevant Acceptance Models

D. Aggregation of the acceptance models

While identifying and analyzing all relevant acceptance models, the most suitable models were selected and their characteristics were observed. Since these differ to some extend fundamentally regarding the factors and dimensions, Figure 3 aggregates all the relevant acceptance models, which affect the attitude and behavior of user acceptance. At the same time a corresponding version of each model is being outlined briefly. Taking the model of Kollman as an example for the other models, it can be shown that indicators related to products, consumers, companies, and the environment, are being considered in this feedback model and therefore facilitate the application for specific scenarios. Likewise these indicators were identified in the other models as well and thus are being subsumed in Figure 3. This figure is therefore the basis for the further progress of this paper. Based on these findings the third section indicates appraisals of the models concerning their capability of considering the adoption of RFID technology in law firms.

II. ADEQUACY OF ACCEPTANCE MODELS FOR RFID

Previous results indicated that the acceptance does not only focus on the simple utilization of an application, but also refers to many individual, social, organizational, technical, economic, task-related, psychological as well as cultural indicators. This situation is the same for implementing and using RFID technologies within organizations. The introduction of such technologies does not only change individual habits but also involves organizational adaption. The following section attempts to identify such indicators in various steps. Afterwards it will be checked which model considers them best and therefore seems the most appropriate in order to evaluate the acceptance of an RFID-based document management system in libraries.

 ISBN: 978-1-61208-100-7

Levels	Acceptance Indicators
Individual Level (user)	- Perceived benefits - Perceived usability - Job relevance - Capabilities & Skills
Task-related Level	- Improvement of the work results - Acceleration in accomplishing tasks - Traceability of productivity - Task characteristics
Organizational Level	- Integration and implementation methods in the company - Rationalization measures - Organizational adaption - Restructuring measures - Integration of benefits
Technical Level	- Maturity level of the technology - System configuration - Degree of standardization - Awareness level of the technology - Other companies experiences - General user-optimization - Modularity
Economic Level	- Productivity - Acquisition costs - Maintenance costs - Monetary benefits - Ability to retrofit
Social Level	- Network effects - Synergy effects - Opinion leadership
Cultural Level	- Cultural sensitivity - Mentality
Psychological Level	- Enhancement of motivation - Enhancement of self-esteem - Improvement of the individual output - Safety in the work routine
	- Fear of job loss - Insecurity during the accomplishment of tasks - Individual readjustment burdens

Figure 4. Assignment of the acceptance indicators according to the levels

A. Identification of relevant acceptance levels

First of all representative acceptance levels will be detected, which may have an impact on the introduction and the utilization of this technology. As a result possible acceptance indicators will be classified more precisely. At the same time it was attempted to derive behavioral psychological and work psychological dimensions. The agreement on the acceptance levels is based on discussions within two workshops carried out with a project team. After having evaluated current acceptance models and other researches, eight possible levels have been concluded in Figure 3.

Individual level: All of those factors that may affect the acceptance of an application, both positively and negatively, at the level of each individual can be found here. Two of the key factors within this level are the *perceived benefits* and the *perceived usability* of an application or a system.

Task-related level: All of those aspects that could have a positive or negative impact on the acceptance of an application and are linked to the task, which needs to be accomplished, are being subsumed in this section. After having analyzed the models cited above, the *improvement of the work results* or the *acceleration in accomplishing tasks* could be mentioned as examples in this context.

Organizational level: a further level to be considered is the organization in which a system is being introduced. The establishment of an RFID technology may result in great *human and structural actions*, such as *rationalization* or *department mergers*.

Technical level: in order to accept a technology or a system, the development of this technology is extremely

 ISBN: 978-1-61208-100-7

important. In this context, the acceptance is influenced by the awareness and maturity level of the applied technology.

Social level: every indicator, which explains how the acceptance within a collective such as a group of users or a organization can be influenced, is being summarized here. In this regard, the so-called network effect is a strong influencing factor. It may be assumed that the acceptance of a system is positively affected if a large number of users has already adopted it [10]. The opinion leadership can also have a great impact on acceptance. This indicates the extent to which an individual is able to influence his social environment [11]. The opinion leadership originates from the marketing theory and can be associated with the reference value model [12].

Cultural level: The consideration of cultural aspects in matters of the acceptance is also required. The cultural sensitivity is of great importance. This means considering country-specific characteristics such as adapting oneself to the mentality of the country for instance [13]. Therefore, the reaction to the introduction of a new system could turn out quite differently in Europe and in Asia.

Economic level: In addition to the dimensions mentioned above, economic aspects are also very important, especially for the management. The focus in this context is the profitability of such actions. Therefore, potential costs such as acquisition o maintenance costs need to be contrasted with the benefits, which are expected by introducing a new technology. A positive result of this analysis could contribute to the acceptance within the corporation.

Psychological level: This level mainly includes factors that are usually not directly visible and measurable. They are rather results due to the changes within the other dimensions. An example could be a department merger, which provokes the fear of job loss and therefore affects employees psychologically. These indicators can have a positive or negative impact on the acceptance.

B. Model evaluation with the help of acceptance indicators

In order to determine which model fits best to evaluate the acceptance of RFID supported document management systems, an evaluation matrix with an adequate rating scale was established for every indentified level mentioned above. This approach made it possible to establish a ranking of the applicable models within the framework of this paper.

The following scale has been chosen for the 35 accumulated indicators in order to evaluate the ten examined acceptance models:

0 = No consideration
1 = Poor consideration
2 = Consideration
3 = Strong consideration

The mentioned rating scale was chosen for several reasons. In order to avoid a tendency towards the centre, a four-way specification was chosen on the one hand. On the other hand, however, a two-way specification (yes/no) did not seem adequate due to the fact that various models offer a lot of space for interpretation and adaptation.

C. Evaluation matrix

The evaluation of the model was carried out by every project member individually within the specified levels with the help of matrices. A subsequent group discussion compared the results and revealed contradictions. A final meeting with all participants completed the assessment. The results were satisfactory for all parties. Figure 5 indicates an example for the evaluation of the individual level.

The individual level is included in the DART approach according to Wehrmann and even stronger in the acceptance model according to Degenhardt. Degenhardt's model is focusing very effectively on individual characteristics. The DART model allows a flexible design of these features within the scope of the sub-dimensions and the process model. The poor performance of the customer satisfaction model according to Silberer et al. can be traced back to its origin. The remaining models cover the identified indicators only partially which explains the rather poor results.

 ISBN: 978-1-61208-100-7

Maximum Number of Points
Individual Level: 12

Acceptance Indicators / Acceptance Models	Perceived Benefits	Perceived Usability	Job Relevance	Skills	Obtained Points	Overall Results (%)
TAM (Davis)	2	1	1	1	5	41,6
Acceptance Model (Degenhardt)	3	2	3	2	10	83,3
TTFM (Goodhue)	2	1	1	2	6	50
Acceptance Model (Filipp)	1	2	1	2	6	50
Dynamic Acceptance Model (Kollmann)	2	2	0	1	5	41,6
Procedure Model (Herrmann)	1	1	1	2	5	41,6
TAM 2 (Venkatesh & Davis)	2	1	2	1	6	50
Acceptance Model for Knowledge Media (Simon)	3	0	3	1	7	58,3
Customer Satisfaction Model (Silberer et al.)	2	1	0	0	3	25
DART Approach (Wehrmann)	3	3	2	2 (3)	10 (11)	83,3 (91,6)

Figure 5. Evaluation matrix of the individual level

D. Determination of the overall results

After having evaluated every model by means of the acceptance levels and the included indicators, the obtained results will be visualized once again in order to get a better comparison of each approach. Besides, the overall results will be calculated. One possibility is the evaluation of the results using the arithmetic mean [14]. Figure 6 indicates the obtained results employing an equal weighting for all levels.

Acceptance Levels / Acceptance Models	Individual Level	Task-related Level	Organizational Level	Technical Level	Social Level	Cultural Level	Psychological Level	Economic Level	Overall Results	Ranking
DART Approach (Wehrmann)	83,3 %	83,3 %	80 %	76,2 %	88,8 %	66,6 %	80,9 %	80 %	79,89 %	*1*
TAM 2 (Venkatesh & Davis)	50 %	83,3 %	13,3 %	14,3 %	66,7 %	66,6 %	38,1 %	0	41,54 %	*2*
Acceptance Model (Degenhardt)	83,3 %	50 %	13,3 %	28,6 %	33,3 %	33,3 %	47,6 %	0	36,18 %	*3*
Acceptance Model (Filipp)	50 %	33,3 %	46,6 %	38,1 %	33,3 %	33,3 %	42,6 %	0	34,65 %	*4*
Procedure Model (Herrmann)	41,6 %	50 %	20 %	28,6 %	11,1 %	0	38,1 %	26,7 %	27,01 %	*5*
Acceptance Model for Knowledge Media (Simon)	58,3 %	50 %	26,7 %	23,8 %	22,2 %	0	33,3 %	0	26,79 %	*6*
Dynamic Acceptance Model (Kollmann)	41,6 %	16,6 %	26,6 %	9,5 %	33,3 %	0	33,3 %	40 %	25,11 %	*7*
Customer Satisfaction Model (Silberer at al.)	25 %	0	13,3 %	28,6 %	44,4 %	0	19,1 %	33,3 %	20, 46 %	*8*
TTFM (Goodhue)	50 %	50 %	13,3 %	14,3 %	0	0	0	0	15,95 %	*9*
TAM (Davis)	41,6 %	25 %	6,7 %	4,8 %	0	0	19,1 %	0	12, 15 %	*10*

Figure 6. Ranking of the acceptance models with the arithmetic mean

Figure 6 clearly indicates the superiority of the DART approach according to Wehrmann against the other acceptance models. This dominance is also reflected in the overall result at each level. The reasons for this advantage can be explained with the basic structure of the model. The high flexibility creates particularly high dynamic extensibility and thus a great scope for interpretation at all levels. The TAM 2 model according to Venkatesh and Davis scores surprisingly well in the overall results. In addition to the comprehensive consideration of the individual level, which has been focused on already in the TAM model by Davis, the obtained result is due to a detailed elaboration of

 ISBN: 978-1-61208-100-7

external stimuli. This step was positively perceived on the task-related as well as the social level. According to the project team both approaches, the one according to Wehrmann as well as the TAM 2 model, allow the integration of cultural aspects sufficiently.

Having observed the first overall results it became clear that the equal weighting of all levels could lead to a falsification of the model rankings [15]. According to the project team, the reason might be a different degree of influence a single acceptance level might have on the acceptance of a technology such as RFID [16]. Therefore, a weighting of the levels was introduced based on the assessments, findings and experience of the project team. Those levels, which are highly relevant in this context, are assessed with a **weighting factor of three**. Levels, which have a different effect on the acceptance depending on the situation, will be assessed with a **weighting of two**. Lower levels, which are expected to have a low impact, will be assessed with a **weighting of one**.

The individual level has a decisive influence on the acceptance of an individual. This fact is also postulated in most models as a key factor. Thus, it is necessary to assess this level and its involved indicators with a **weighting of three**. The social level also plays an important role within the scope of this project. Hence, network effects, synergies and opinion leadership could have a great impact on the acceptance or use in institutions such as law firms and libraries. Therefore it makes sense to assess these levels with a **weighting factor of three**. In this context the psychological level should also be assessed with a **weighting factor of three** since it has influence on all levels of acceptance. Due to its connection it plays an important role in the evaluation. The task-related, technical, organizational and economic level can have a completely different effect on the acceptance of an individual or a collective, depending on the institution, design and other conditions. This can be observed in the examined models in which they are not treated equally. In order to accentuate this fact, a weighting factor of two was assessed. The cultural level, a subordinate factor that should not be underestimated, is not being considered in any model. However, a **weighting factor of one** was assigned in order include the level in the evaluation model. Figure 7 illustrated the changes that have taken place in the ranking of the evaluated models. Although each acceptance model of Simon, Degenhardt, and Kollmann improve their values by more than three percent, the DART approach according to Wehrmann is still considered to be the best model in the analysis.

Overall Results after two Evaluations Acceptance Models	Overall Result Evaluation 1	Overall Result Evaluation 2	Alteration (%)
DART Approach (Wehrmann)	79,89 %	81,36%	+ 1,47 %
TAM 2 (Venkatesh & Davis)	41,54 %	41,82%	+ 0,28 %
Acceptance Model (Degenhardt)	36,18 %	39,43 %	+ 3,25 %
Acceptance Model (Filipp)	34,65 %	33,58 %	- 1,07 %
Acceptance Model for Knowledge Media (Simon)	26,79 %	30,13 %	+ 3,34 %
Procedure Model (Herrmann)	27,01 %	29,06 %	+ 2,05 %
Dynamic Acceptance Model (Kollmann)	25,11 %	28,33 %	+ 3,22 %
Customer Satisfaction Model (Silberer et al.)	20, 46 %	23, 11 %	+ 2,65 %
TTFM (Goodhue)	15,95 %	16,96 %	+ 1,01 %
TAM (Davis)	12, 15 %	14, 17 %	+ 2,02 %

Figure 7. Effects on the second evaluation of the model

III. FINAL EVALUATION

After a detailed analysis of the mentioned acceptance models, an examination has taken place in order to identify whether they were suitable for an evaluation of RFID-based documant management system. As a result it became clear that the approach according to Wehrmann was offering the best conditions in almost every research area. This dominance is not only due to the overall results. The model also achieved optimum values in those levels, which according to the project team had a great impact on the acceptance, as well. Furthermore an ideal organization based on the characteristics for evaluating the acceptance of RFID technology in law firms is admitted by the DART approach due to its high flexibility and modifiability. Based on these findings, the project team has opted for the DART approach according to Wehrmann. In the further progress of the project it is ought to adjust the basic structure of the model,

 ISBN: 978-1-61208-100-7

which was outlined in the literature, to the specific research circumstances in order to carry out a promising acceptance analysis. Based on this analysis of acceptance models, the DART model is being used as the basis for the actual analysis of the acceptance of RFID in lawyer's offices. Dimensions such as perceived network effects, perceived costs, and perceived benefits are being examined according to indicators such as investment costs or surface handling. As a result of analyzing the 10 acceptance models it can me stated that the DART model describes those indicators, which are necessary for the scenario of RFID in lawyer's offices, best.

In order to achieve generalization, it was ought to include structural, local and temporal limitations. Structural limitations affect the chosen research design. The analytical structure of this paper involves a study of scientific literature dealing with acceptance models in general. Additionally, it is being specialized by using four workshops until it reaches the complex issue of RFID in law firms. However, this approach is correct due to the very poor literature provided on evaluating the acceptance of RFID in law offices. Based on the lack of knowledge, the results may vary when applying different research designs.

Besides, the German legislation as well as the composition of the workshops need to be considered as an important reason of local limitations. Law firms and attorneys operating in different European countries or in other parts of the world are facing distinct legal standards and working methods. This is why the identified acceptance levels and indicators can be transferred only partially. The second limitation relates to the time circumstances during the investigation. Since there are no scientific studies related to RFID in law firms and the use of RFID technology in this environment, the declaration given by the experts only reflects their current opinion. However, the identified results are representative for these issues. It can be assumed that it is possible to transfer acceptance indicators on defined processes in law firms due to the legal actions in Germany and the rigid operations in this profession.

REFERENCES

[1] Gartner, “Market Trends: Radio Frequency Identification. Worldwide, 2007-2012,” Stamford, 2005.

[2] S. Weigert, “Radio Frequency Identification (RFID) in der Automobilindustrie - Chancen, Risiken, Nutzenpotentiale,” Gabler, Wiesbaden, 2007.

[3] VDEB, “RFID Anwendertag 2009 des Verband IT-Mittelstand – Anwender treffen Experten,“ http://www.pressebox.de/pressemeldungen/vdeb-verband-it-mittelstand-ev/boxid-300867.html, accessed 11 June 2010.

[4] V. Sohmer, “Management – Denn Sie wissen, was Sie tun,” in Handelszeitung no. 21, Zürich, 2008, p. 25.

[5] German Federal Ministry of Justice, “Bundesrechtsanwaltsordnung,“ http://bundesrecht.juris.de/brao/__50.html, accessed 11 May 2010.

[6] J. Gabius, “RFID als erfolgsversprechendes Instrument einer Marktforschung,“ diploma thesis, 1 edition, GRIN Verlag, Norderstedt, 2007, p. 4.

[7] G. Schoblick and R. Schoblick, “RFID- Radio Frequenzy Identification: Grundlagen, eingeführte Systeme, Einsatzbereiche, Datenschutz, praktische Anwendungsbeispiele,” Franzis Verlag, Poing, 2005, p. 153ff.

[8] J. Wehrmann, “Situationsabhängige mobile Dienste: Konzepte und Modelle zu ihrer effizienten Entwicklung unter besonderer Berücksichtigung der Benutzerakzeptanz,” WiKu Verlag, Berlin, 2004.

[9] T. Kollmann, “Akzeptanz innovativer Nutzungsgüter und -systeme: Konsequenzen für die Einführung von Telekommunikations- und Multimediasystemen,” Gabler Verlag, Wiesbaden, 1998, p. 61.

[10] G. Tamm and O. Günther, “Webbasierte Dienste: Technologien, Märkte und Geschäftsmodelle,” Institut für Wirtschaftsinformatik, Universität St. Gallen, Psycho Verlag, Heidelberg, 2005, p. 67.

[11] N. C. Schneider, “Kundenwertbasierte Effizienzmessung – Der Beitrag von Marketingmaßnahmen zur Unternehmenswerterhöhung in der Automobilindustrie, ” Deutscher Universitäts - Verlag (DUV) Gabler, 2006, p. 151.

[12] J. Cornelsen, “Kundenwertanalysen im Beziehungsmarketing - Theoretische Grundlegung und Ergebnisse einer empirischen Studie im Automobilbereich,” GIM-Verlag, 2000, p. 199.

[13] M. Müller, “Die Identifikation kultureller Erfolgsfaktoren bei grenzüberschreitenden Fusionen: Eine Analyse am Beispiel der DaimlerChrysler AG,” 1st edition, Deutscher Universitäts – Verlag, Wiesbaden, June 2007, p. 151ff.

[14] P. Shannon, “Valuing Business: The Analysis and Appraisal of closely Held Companies”, 5th edition, Mac-Graw Hill Books, 2008, p. 209ff

[15] P. Niven, „Balanced Scorecard: for Government an non profit agencies”, 2nd edition, Wiley Publishing, 2008, p. 95ff

[16] M. G.Brown, „Beyond the Balanced Scorecard: Improving Business Intelligence with Analytics”, 1st edition, Productivity Press, 2007, p. 67ff

 ISBN: 978-1-61208-100-7

System Architecture for Mobile-phone-readable RF Memory Tags

Iiro Jantunen, Jyri Hämäläinen, Timo Korhonen
Department of Communications and Networking
Aalto University
Espoo, Finland
{iiro.jantunen, jyri.hamalainen, timo.korhonen}@tkk.fi

Harald Kaaja, Joni Jantunen, Sergey Boldyrev
Nokia Research Center
Helsinki, Finland
{harald.kaaja, joni.jantunen,
sergey.boldyrev}@nokia.com

***Abstract* — We have developed an open architecture platform for implementing passive radio-frequency identification (RFID) tags with a mass memory. Purposes for such mass memory tags are, e.g., multimedia files embedded in advertisements or logged sensor data on a low-power sensor node. In the proposed architecture, a mobile phone acts as the reader that can read or write the memory of these RFID tags. The architecture is designed so that development path to a full Network on Terminal Architecture (NoTA) is feasible. The wireless reading speed of the mass memory tags, demonstrated to be 112 Mbit/s, is in range that a 3-minute VGA size video can be loaded from the tag to the phone in less than 10 s.**

Keywords — Memory architecture; Multimedia systems; RFID; telephone sets; RF memory tags

I. INTRODUCTION

RFID tags are increasingly a part of our life; transport, traceability, and secure access are some of the main uses of this technology today. Conventional machine-readable wireless tags, e.g., Near Field Communication (NFC) tags, normally have a very small memory in range of hundreds of bytes or kilobytes. Some RFID standards include an option to have a flexible-use memory, but the capacity is low compared to factory-set fixed-content memory. Tag selection is based on reading the content in a selected tag memory address (e.g., tag or manufacturer ID). As the memory capacity of these tags is small, the amount of data to be transferred is also small and power consumption of RF communication is, thus, not a critical issue.

To overcome the storage capacity limitation of passive tags, Wu et al. increase effective tag storage sizes with proposed distributed RFID tag storage infrastructure (D-RFID stores) [1]. Ahmed et al. focus on RFID system unreliability and improvements in middleware for object tracking and object location with moving readers or tags [2][3]. Ying described a verification platform for RFID reader that utilized Ultra High Frequency (UHF) frequency [4]. Pillin et al. have developed a passive far-field RFID tag using the 2.45 GHz Industrial, Scientific and Medical (ISM) band, with a data rate of 4 Mbit/s on the range of 5.5 cm [5]. As an example of a proprietary solution, HP's *Memory Spot* RFID tag also works on the 2.45 GHz band and has demonstrated 4 MB memory and 10 Mbit/s data rate [6].

Today's mobile phones provide music and video players, which make it possible for consumers to enjoy entertainment while on the move. Acquiring new multimedia content by downloading or streaming, however, is hampered by the high cost and slow speed of Internet connections, as well as by the fact that commonly used physical multimedia formats, such as optical disks, cannot be read with a mobile phone. Thus, to make acquiring new content easier, cheaper and less power-consuming, we propose a new technology based on RF memory tags readable and writable by mobile phones.

The problems of low data reading rate and small memory size provided by contemporary RFID tags become emphasized if one considers mobile users reading multimedia files from tags embedded on, e.g., paper media. The attention span of a mobile user is about 10 seconds [7]. Within this period, the user could get a single multimedia content file from a memory tag. Considering a movie trailer, the file size for a 2-minute 640×320-pixel 30-fps (3 Mbit/s), encoded with H.264, would be in range of 50 MB [8]. The required minimum data transfer rate from the user point-of-view is thus 50 Mbit/s. This exceeds the maximum data rate available by 13.56 MHz NFC technology, 848 kbit/s, by a factor of 60. Even the maximum data rate for NFC demonstrated on a laboratory set-up, 6.78 Mbit/s [9], is not enough. Thus, there is a need for a new high-speed touch-range RFID radio interface.

The aim of our research has been to develop a high-capacity memory tag, which is wirelessly readable with a mobile phone and suitable for consumer markets in ubimedia applications [10][11]. The mobile phone acts as the user interface for reading and writing passive RF memory tags that contain a high-capacity memory (0.1–1 GB). The reasoning for the proposed technology was justified by modern trends in non-volatile memory technologies, according to which the power consumption, physical size, and price of memory are continuously decreasing.

In this paper we describe and specify a network architecture, which enables mobile phones to read and write passive RF memory tags. The architecture has been developed and demonstrated in the EU's 6th Framework Programme (FP) "Micro-Nano integrated platform for transverse Ambient Intelligence applications" (MINAmI) project [12], and thus the architecture is referred to as MINAmI Architecture. Important architecture requirements include openness, modularity, scalability, and energy efficiency. Openness and modularity are needed to support creation of novel applications and services by 3rd parties. Scalability of data rate is needed to enable evolution of the technology along with evolution of multimedia services. Energy efficiency is essential to enable passive operation of the tags as well as to save the phone's battery.

The paper is organized as follows. In Section II, we introduce the system architecture, along with a key component of the architecture, RF memory tags. In Section

 ISBN: 978-1-61208-100-7

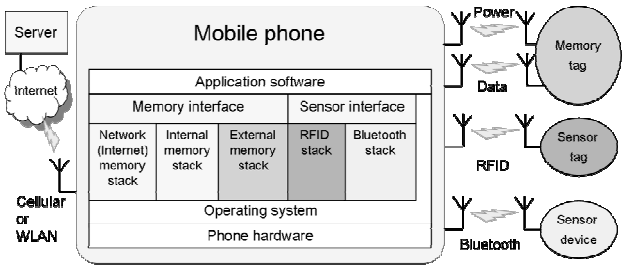


Figure 1. MINAmI Architecture

III we introduce a novel dual-band radio subsystem and its hardware and software implementation. In Section IV, we present the current status of implementation of the architecture and discuss possibilities for future development. Section V concludes the paper.

II. MINAMI ARCHITECTURE

The proposed MINAmI architecture makes use of the mobile phone's capability of running software and providing several radio interfaces (Figure 1 [11]). The architecture is modular, enabling simpler and faster development of new technical extensions (e.g., RF memory tags). Our architecture focuses on utilization of modularity on component level (e.g., where to plug memory tag functionality) and on communication level (e.g., how the available memory tags are utilized). At short proximity domain (range < 1 m), different tags are communicating locally with a mobile phone. In the present work we have concentrated on the RF memory tags. The sensor parts of the architecture (RFID sensor tags and Bluetooth sensor devices) have been studied in an earlier project [13][14].

The main RF memory tag architectural design challenges include target platform performance obstacles, such as available bus operations (R/W) and power requirements, especially when drawing the line for autonomous operations in described MINAmI architecture. The other challenge is minimizing changes to the existing system communication layering, only to the external memory stack block. The choices in the MINAmI system architecture were able to support both existing standard radios for low-rate sensors, and the high-rate high-capacity memory tags.

A. Network-on-Terminal Architecture (NoTA)

NoTA is modular service-based system architecture for mobile and embedded devices offering services and applications to each other. The concept is being defined in an open initiative, in NoTA World [15]. NoTA is also known as an open device distributed architecture, which allows direct connections between different nodes, within subsystem or between subsystems. This architecture supports both messaging and streaming services. The beauty in the architecture resides in modularity and transport independency. Direct connection between subsystems improves the efficiency as they do not necessarily require any processor involvement, when subsystems have all the needed functionalities available for their independent operations. Transport-specific portion is hidden underneath NoTA communication layering.

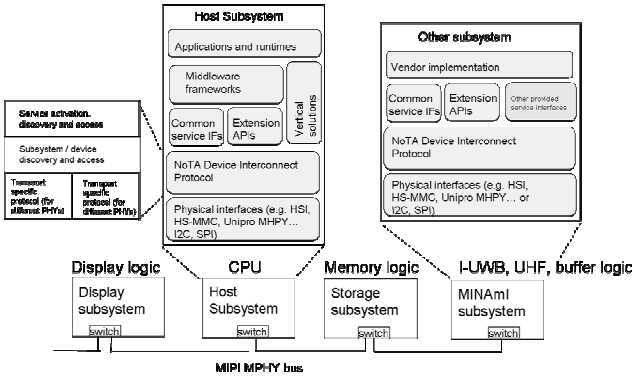


Figure 2. NoTA extension architecture for MINAmI subsystem, where HSI - High Speed Serial Interface; HS-MMC - High Speed MultiMediaCard: SPI - Serial Peripheral Interface bus.

NoTA communication layering is built around transport-independent parts and interfacing towards transport-specific parts (Figure 2). Device Interconnect Protocol (DIP) provides logical links between a requesting subsystem and other subsystems or within a subsystem.[15]. DIP is a device-level communication protocol that can be implemented for various physical interfaces ranging from MIPI (Mobile Industry Processor Interface) high speed serial interfaces and Universal Serial Bus (USB) to wireless interfaces, such as Bluetooth [16][17].

NoTA host subsystem and neighboring subsystems are connected via the high speed physical interface. DIP adapts physical interfaces to the upper layers. It is the lowest layer that is common for all subsystems (i.e., also for MINAmI subsystem) and hides the physical dependencies underneath. Above DIP there is a common service interface used for resource management, file systems, and system boot-ups. Middleware frameworks, e.g., for multimedia, USB, and other applications, use a common service interface or extension Application Programming Interface (API). The architecture also takes into account vertical solutions, which may require an optimized protocol design for certain requirements that are tied to HW-specific applications.

NoTA subsystem structure takes into account possibility to add different types of independent (service/application) subsystems to the architecture, and MINAmI architecture forms one high data rate high capacity subsystem.

MINAmI subsystem offers memory tag read/write, storage and local connectivity services to other subsystems within mobile device, and its architecture is compatible with NoTA communication layering. MINAmI subsystem includes both the mobile phone (Mobile Reader/Writer) and the tag and all the relating hardware and software resources. Mobile Reader/Writer sees the contents of the memory of a passive RF memory tag only when there is an established connection, i.e., power field and data connection exists.

B. RF Memory Tags

The focus of our research has been on mobile-phone-operable memory tags suitable for consumer markets and ubimedia applications. The tag is developed as a part of our mobile-phone-centric architecture. Our memory tag

 ISBN: 978-1-61208-100-7

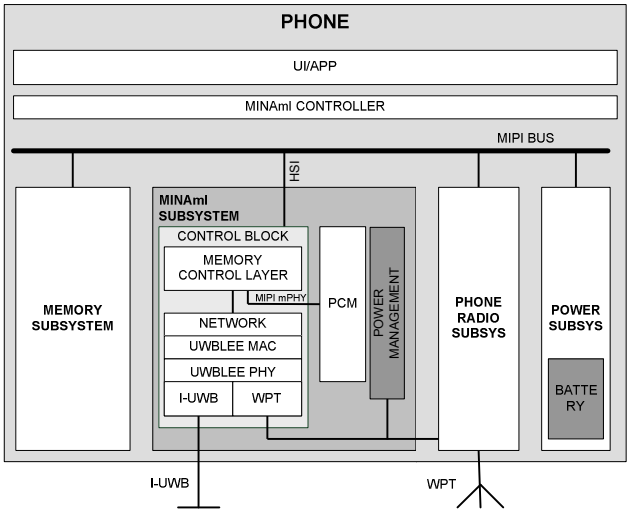


Figure 3. MINAmI Architecture on phone

development targets improving both transfer speed and storage capacity. These improvements give direct benefit for ubimedia users.

The target memory capacity of our memory tag has been in the range of gigabits and mobile reader/writer transfer speed to and from memory tag in excess of 10 Mbit/s. The same design platform is usable for both ends, for mobile phone platform reader/writer and for tag implementation. When designing the platform, various important design parameters, such as the selection of the used radio technology, were considered to provide an efficient and low-power solution for mobile reader/writer and tags.

It was important to make sure that connectivity technology is simple enough for the user, e.g., it should facilitate easy content selection (see Section III.D). Memory tag content selections should be based on metadata (e.g., filenames, file content types, file content keywords). Due to the large memory size, power consumption for memory access is a critical design issue, both for reading and writing the memory tag. To be successful on the market, RF memory tags for ubimedia must be passive to make them as small (size) and cheap as possible, and to achieve autonomous usage with minimum maintenance (e.g., usage without charging of battery). This severely limits the power budget. On the other hand, a short communication range (even touch) is sometimes preferable to make it easier for the user to physically select the tag. An RF memory tag will be read many times by different users, but written more rarely – in some cases, only once. The memory unit must work reliably even with several consecutive read cycles. A limited write throughput due to power constraint is not an issue, as data is rarely written by the users.

III. UWB LOW END EXTENSION

As memory tags have high data storage capacity, a high-speed radio is needed for communication to enable reading even all the contents of the tag in an acceptable time. Currently available mobile phones contain several radio transceivers, such as cellular, Bluetooth, and Wi-Fi, along

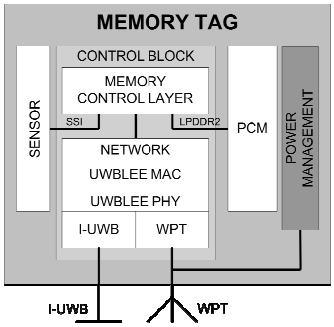


Figure 4. MINAmI Architecture on a RF memory tag

with NFC. Most of the technologies are made for well-established communication between active devices, consuming a relatively large amount of power. These technologies are also not inherently designed for ad-hoc, possibly one-time, connections between devices that have not communicated with each other before, resulting in long latency in establishing the communications. For example, in an environment with many unknown Bluetooth devices, the Bluetooth connection setup latency can be over 10 seconds [18]. NFC enables communications between an active and a passive battery-less device and is physically more selective; its communication range is almost in touch. However, it has severe limitations in data transfer speed.

To provide higher data rates, a wider frequency band available on higher frequencies needs to be used. On the other hand, the efficiency of wireless power transfer (WPT) decreases as a function of center frequency. To solve the problem of providing high-speed communication (high frequency needed) while simultaneously providing power wirelessly to the tag, a dual-band radio interface has been proposed [19]. One narrowband signal on RFID frequencies (e.g., RFID frequency bands globally available between 860–960 MHz) is used to power the tag and to provide a mutual clock reference for both ends of the communication link, whereas the communication link itself is based on impulse UWB technique to provide a high communication bandwidth and scalability for even higher data rates.

As the selected RFID frequencies are approximately in the same frequency range as GSM/WCDMA 900 MHz, in the reader there is a possibility of integrating the WPT function to the existing Phone Radio Subsystem, as presented in Figure 3. In that case, Phone Radio Subsystem is designed so that the WPT Physical (PHY) Layer function may request a direct access to control the activation of the narrowband transmitter. Especially, the time-domain interleaving of different functions is important to support co-existence of GSM/WCDMA and WPT signaling.

The architecture of the RF memory tag (Figure 4) is similar to the MINAmI subsystem on the mobile phone. For simple RF memory tags, no network layer implementation is needed to take care of the point-to-point communication between the reader and the tag, and therefore is handled on Medium Access Control (MAC) layer.

As an option for use-cases like data-logging sensor devices, the memory control layer provides a sensor interface. During the sensing, the sensor data is stored to the

 ISBN: 978-1-61208-100-7

Phase-Change Memory (PCM) block and the low data-rate data capturing is powered from a battery or with energy harvested from the environment. For fast downloading of the logged data, the reader powers the sensor tag wirelessly.

A. Hardware architecture

This subsection describes the enabling technologies.

1) Radio Front-end

As presented in [19], very simple super-regenerative transceiver architecture can be used in impulse UWB communication to achieve required data-rates over short distances. In contrast to conventional impulse UWB transceivers [20] there is no need for multipath recovery over the distances below 30 cm. This decreases the requirements set for the UWB transceivers. This is used to minimize complexity and power consumption of the transceivers. In the aforementioned super-regenerative transceiver one super-regenerative oscillator is used alternately both to generate transmitted pulses and to amplify received pulses, and no linear amplifiers are needed. Thus, the architecture utilizes the inherently low duty cycle of the transmitted impulse UWB signal also in reception the receiver being fully active only exactly during the detection of incoming pulses.

Synchronization is often problematic in impulse UWB systems because of the low duty cycle and pseudo-random timing of pulsed signal, and due to frequency drift and differences of reference clocks between the transceivers. In the proposed system the frequency synchronization between the reader and tag is achieved thanks to the mutual narrowband WPT signal, which is also used as the reference clock. The phase synchronization of impulse UWB transceivers is also easier to achieve due to decreased need for pseudo-random time-coding of pulse patterns.

The transceiver structure supports simple On-Off-Keying (OOK) modulation. The data-rate and power consumption is also scalable depending on the power level available for the wirelessly powered tag. Due to the simplified transceiver structure, targeted ultra-low power consumption and partial exploitation (500 MHz) of full UWB band (3.1–10.6 GHz) authorized by Federal Communications Commission (FCC) for unlicensed use, the impulse UWB system referred here is called UWBLEE (UWB Low End Extension).

Altogether, the optimized transceiver architecture makes it possible to achieve required high data-rates with a low power consumption performance (a few mW) suitable for WPT. As a proof-of-concept a complete wirelessly powered RF front-end implementation of the super-regenerative transceiver is presented in [21] by using a single super-regenerative oscillator for transmission and reception. The front-end implementation supports data-rates up to 112 Mb/s with the energy consumption of 48 pJ/bit in reception and 58 pJ/bit in transmission. The feasibility of the ultra low power consumption in high data-rate two-way communication is verified with an integrated RF front-end implementation based on the symmetrical transceiver architecture proposed earlier [19]. A 900 MHz WPT signal is used as a mutual clock reference and the communication is done over an impulse UWB link at 7.9 GHz center frequency. The scalable data-rate of UWB link up to 112 Mbit/s has been demonstrated as well as robustness against narrowband interference.

2) Non-Volatile Memory (NVM) technology

The main reason to pick up PCM in favor of any other memory technology [22] were the benefits of PCM technology, e.g., the estimated high number of read/write cycles as 1×10^6, which consequently results in need of no or just a lightweight wear leveling algorithm, and the bit alterability – lack of need of block erase cycles (as with flash memory) when data should be stored. From the perspective of technology lifecycle PCM stands now between a pure innovative technology and early adopters' stage. There are several 90 nm products [23] on the market already and more to come.

Aggregating main memory characteristics in comparison with NAND/NOR flash technology and DRAM execution memory, PCM stands between those two in terms of cost per die. It is characterized as 5.5 F^2 factor in cell size having the same wafer complexity as DRAM technology. Currently only Single Level Cell (SLC) PCM is available, though Multi-Level Cell (MLC) PCM is on the way out, which can substantially extend the density and, justify the cost structure. Thus, the application range can be quite wide from external usage (cards, keys) and wireless applications (RF memory tags) to high performance computing applications (caches, code execution, data storage). Considering reliability characteristics it is important to note that PCM technology gives more than 10 years retention ratio that can be extended even further, if necessary, by proper bit error management.

PCM has performance characteristics such as read & write latency and read & write endurance almost as good as DRAM, while giving clear benefits through the non-volatile nature of PCM technology. PCM has a low system-wise energy consumption (~0.2 mW/pF read, <1.25 mW write) ~<1 mW/GB of idle power, access time comparable to DRAM (~85 ns), with read latency 50–100 ns, write bandwidth from 10 to 100+ Mbit/s/die, write latency 500 ns – 1 µs, various packaging and die stacking solutions, high-speed low-pin-count low-power interface solutions, and maturity of the technology as such.

The PCM technology highlights provide clear reasoning for the selection of such technology for the RF memory tag application, preserving the opportunity to justify it even further when some other application should be designed.

B. Software Architecture (protocol stack)

The MINAmI software architecture (protocol stack) is designed to be modular and scalable. The protocol stack is based on three layers: Network Layer, MAC Layer, and PHY Layer. The application programming interfaces (API) of the layers are open for 3rd parties. These layers will be presented in the following sections. The protocol stack has been developed taking into account future compatibility with NoTA architecture. Care should be taken to have a clear implementation path towards the final architectural (NoTA) solution.

 ISBN: 978-1-61208-100-7

1) Network Layer

Network Layer will first only provide point-to-point connections regardless of state. In future, also applications using multiple targets could become feasible when MINAmI Subsystem is in active mode. If a point-to-multipoint network protocol is needed, nanoIP is easily implementable [13][24]. However, to get full internet support classical IP protocol may be valid, and more common in networking devices. In the final architecture (NoTA) solution, the network layer will consist of Device Interconnect Protocol (DIP), as a middleware, which guarantees the compatibility with NoTA. In DIP protocol, it is possible to select, which transport mode and network is used. For example DIP TCP L_IN (transport selected) is ready to be used within one device and between several devices in a sub-network as such. Multicasting must be enabled in IP interface in order for device discovery to work. Nodes, which are in different sub-network, cannot be detected [15].

Packet size is an important parameter and depends on what is feasible for MAC and PHY layers. Upper layer packets are segmented and reassembled and this is dependent on what kind of packet sizes the system supports.

2) MAC Layer

The MAC of the novel dual-band radio interface has three different operational modes: the passive mode, where no internal power source is available or used; and the active and semi-passive modes, where internal power source is available. Tags on battery-less objects without power wire connection (e.g., implanted on paper) are passive.

In the active mode, the mobile phone actively searches and selects the target tags, sends the targets the WPT signal for powering and for frequency synchronization of the communication link, reads/writes data on the tags, and closes the connection to the target when active connection is no longer required. This operation can be an automatic feature, or enabled by the user (initiating the application for reading and writing the tag). In the semi-passive mode the phone receives data sent by an outside device, but powers itself, allowing a longer communication range, which would otherwise be limited by the WPT link. In semi-passive mode, however, the initiator device takes care of the synchronization of the I-UWB communication link.

Active mode states are used by battery-powered mobile devices, whereas passive mode states are applied for passive devices and tags. In passive mode, possible connections are powered by an outside device with WPT. In the passive operating mode the default state (when powered by an outside device) is P-IDLE, i.e., ready to receive any data, after the boot-up sequence.

The main operational states of UWBLEE MAC are shown in Figures 5 and 6. In addition to the shown directions of movement from state to state, there need to be possibility of built-in error recovery operation from any operational state to the corresponding idle state (A-IDLE or P-IDLE). For the applications requiring higher security, a protocol based on NFC is applied for the ongoing data transmission.

3) Physical Layer

UWBLEE PHY layer controls both the I-UWB communications and Wireless Power Transfer (WPT)

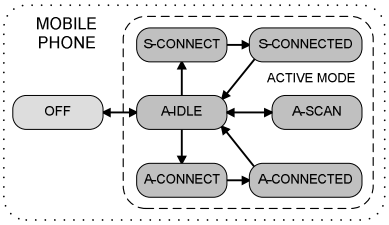


Figure 5. Active (and semi-passive) UWBLEE MAC states on a mobile phone. Active states denoted with A, semi-passive with S.

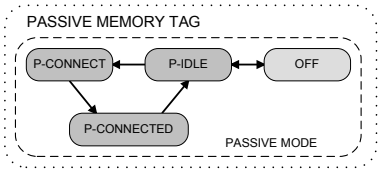


Figure 6. Passive UWBLEE MAC states on a RF memory tag.

TABLE I. UWBLEE PHY IN DIFFERENT MAC STATES

	MAC mode		
	Passive	***Semi-passive***	***Active***
I-UWB	Transmit / receive		
WPT synch	Receive	Receive	Transmit
WPT power	Receive		Transmit
Power source	WPT reception	Battery	Battery
Remarks	Being read / written		Reading / writing other devices

transmission. Depending on the operational mode (active or passive) WPT link is used to send (or receive) power and/or to provide the clock reference signal.

UWBLEE PHY is divided to two sub-blocks: I-UWB PHY and WPT PHY. I-UWB PHY controls the Impulse-UWB radio interface and WPT PHY controls the Wireless Power Transfer interface. I-UWB PHY and WPT PHY are coordinated by UWBLEE PHY so that I-UWB transmission is synchronized with the WPT transmission.

The function performed by UWBLEE PHY is defined by UWBLEE MAC, as shown in Table 1.

C. Packet-level Communication

The MINAmI subsystem communication between active mobile reader/writer and passive RF memory tag consists of periods shown in Figure 7. In the beginning there are no tags within the mobile reader/writer local connectivity coverage. If the mobile reader/writer detects a tag during the powering period, it tries to scan all tags available (in the polling period) and – based on the current selection criteria – choose one with whom to communicate (in the activation period). The right tag is found by scanning the coverage area, synchronizing communications with the tags, and selecting the right tag. After this selection, connection and device configuration is executed in the initialization period to set communication parameters, to specify packet level parameters (e.g., length, memory allocation). The connection period is initiated when connection between mobile reader/writer and selected tag is established. This is followed by the data transmission period, reading and/or writing

 ISBN: 978-1-61208-100-7

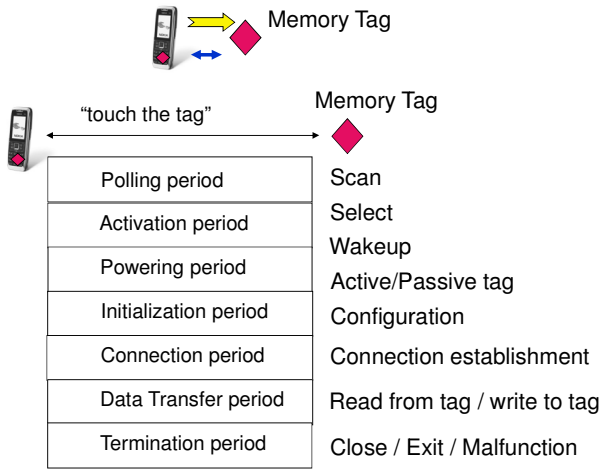


Figure 7. Mobile reader/writer to RF memory tag communication sequence

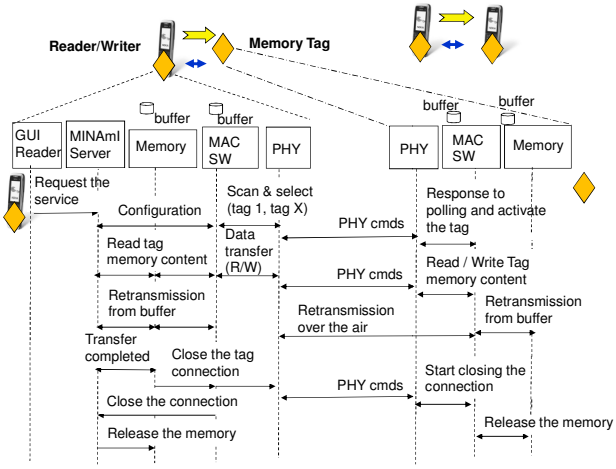


Figure 8. Basic MINAmI subsystem communication setup sequence

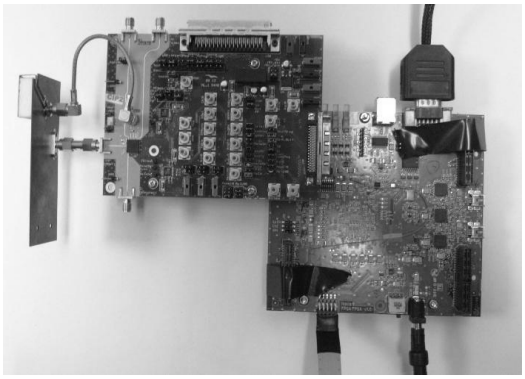


Figure 9. Our UWBLEE implementation

selected content from/to tag. After successful data transmissions, in the termination period, connection is closed or continued with another read/write operation to the tag.

Basic connection procedure between a mobile reader and a tag is described in Figure 8, which also identifies affected internal entities, e.g., MINAmI server, memory management, and communication entity (MAC and PHY layers). For the air interface, the data from/to the non-volatile storage memory (PCM) is buffered into a DPRAM buffer memory equal to the maximum packet size transferred over the air.

D. File System Design

The mobile phone can read tags and with writeable tags the phone can also write all or parts of their contents. The communication capacity between the mobile terminal and the RF memory tag is targeted to exceed 50 Mbit/s (as discussed in Section I). Plug-in software (External memory stack in Figure 1) is required to facilitate seamless use of the tag memory for mobile phone applications.

The memory tag can be used as an extension to the local file system of the reader (e.g., mobile phone). The memory tag can be either a passive and cheap one or an active one, including an own power source and thus being more expensive [11]. Plug-in software in the file system of the reading device handles the connection to the memory tag. Storage space on the memory is mounted on the local file system in the same way as any detachable storage. The volatile nature of the connection causes overhead in maintaining the file system view in the reader/writer device.

Adding a processing element to the memory simplifies the connection. An ultra low-power processing element can process the access requests independently and even provide some more advanced services like metadata-based queries. A service proxy relays the service interface of the memory directly to the applications running on the accessing device. The volatile nature of the connection is not a problem if the server is made stateless and transactions atomic.

Device internal modules need to support NoTA to get full benefit of subsystem independency and still give a fast connection between subsystems. This interconnect architecture allows future extensions for modules within one device.

IV. DISCUSSION

The RF memory tag (i.e., mobile reader/writer and tag) solution was developed and tested in the MINAmI project. Implementation is shown in Figure 9. The development of a RF memory tag sub-system of MINAmI project is based on a flexible, FPGA-based hardware platform. The sub-system takes benefit from the ultra-low power UWBLEE transceiver architecture, which is suitable for data rates required in RF memory tag applications. The technical results are promising and useful for the concept of mobile-phone-readable RF memory tags. The data-rate of 112 Mbit/s has been achieved over the novel radio interface in technical demonstrations [21]. This leaves room for up to 50% protocol and memory access overhead when targeting to 50 Mb/s end-to-end communication. On the PHY and MAC layers short target distance and point-to-point communication efficiently minimize the protocol overhead on packet level. However, efficient pipelining in buffering of the data is in crucial role in optimization of the end-to-end system. The third important factor is the memory access

 ISBN: 978-1-61208-100-7

speed. This is relevant when reading data from the source memory and when writing the data to the target storage memory. As shown in Section III A.2 the continuous development of NVM memories is will provide power-efficient and fast solutions for the target applications. Altogether, the listed factors and the results achieved with the demonstration platform show that mobile reader/writer and the high capacity memory tag is implementable.

A. *Future development*

The UWBLEE wireless connection technology presented in this paper provides data rates significantly exceeding the existing NFC technology already in the market. From technology ecosystem point-of-view there is little sense in developing UWBLEE as an independent technology. UWBLEE can thus be seen as a possible future high-speed extension to existing RFID or NFC technologies.

In a multi-device environment one device can work as a proxy for the memory tag and provide other devices with access to its services [11]. There are also possibilities to have memory tags with their own power sources, which eliminate the need of wireless powering. In that case, the reading range can be extended or power use within the mobile phone can be reduced. The phone can also communicate directly with other similarly equipped phones.

Our RF memory tag solution supports Nokia's *Explore and Share* concept, a new way of transferring content (e.g., multimedia, maps, and applications) to a mobile phone [25].

V. CONCLUSIONS

The evolution of non-volatile memory technologies gives the basis for the vision about RF memory tags. However, the large memory creates a need for a high-speed data connection that can be used to transfer the contents of the tags in a timeframe acceptable for the user. The dual-band radio interface, UWBLEE introduced in this paper provides the required data rate and possibility for future scalability as memory sizes become larger.

Modular architecture is mandatory in the RF memory tag system to optimize performance. For example, latencies common in memory access of centralized systems are not acceptable. Power consumption of the mobile reader/writer is efficiently minimized with an independent sub-system keeping the involvement of the main processor at the minimum. In contrast to conventional radio systems, the main processor only triggers the communication and the independent sub-system handles the transfer and storage of the data. Thus, the main processor does not have to be involved in the low level communication processes.

ACKNOWLEDGMENT

We thank all the partners of MINAmI project, who have contributed to development of this technology. This study is a part of MINAmI project under EU FP6 contract IST-034690. This research was also supported by the Finnish Academy under grant no 129446.

REFERENCES

[1] V. Wu, M. Montanari, N. Vaidya, and R. Campbell, "Distributed RFID tag storage infrastructure". University of Illinois at Urbana-Champaign, IL, USA. Tech. Rep. 2009.

[2] N. Ahmed, R. Kumar, R.S. French, and U. Ramachandran, "RF2ID: a reliable middleware framework for RFID deployment", Proc. IPDPS 2007. IEEE, 2007, pp 1–10.

[3] N. Ahmed, and U. Ramachandran, "Reliable framework for RFID devices". Proc. MDS'08. ACM, 2008, pp. 1–6.

[4] C. Ying, "A verification development platform for UHF RFID reader", Proc. CMC'09. IEEE, 2009, pp. 358–361.

[5] N. Pillin, N. Joehl, C. Dehollain, and M.J. Declercq, "High data rate RFID tag/reader architecture using wireless voltage regulation". Proc. RFID 2008. IEEE, 2008, pp. 141–149.

[6] "HP unveils revolutionary wireless chip that links the digital and physical worlds". HP, Palo Alto, CA, USA, 2006.

[7] J. Nielsen, "Usability engineering". Morgan Kaufmann, 1993.

[8] W. Cui, P. Ranta, T.A. Brown, and C. Reed, "Wireless video streaming over UWB". Proc. ICUWB 2007. IEEE, 2007, pp.933–936.

[9] H. Witschnig, C. Patauner, A. Maier, E. Leitgeb, and D. Rinner, "High speed RFID lab-scaled prototype at the frequency of 13.56 MHz". Elektrotechnik & Informationstechnik, vol. 124, 2007, pp. 376–383.

[10] J. Jantunen, I. Oliver, S. Boldyrev, and J. Honkola, "Agent/space-based computing and RF memory tag interaction". Proc. IWRT 2009.

[11] E. Kaasinen, M. Niemelä, T. Tuomisto, P. Välkkynen, I. Jantunen, J. Sierra, M. Santiago, and H. Kaaja, "Ubimedia based on readable and writable memory tags". Multimedia Systems, vol. 16, no. 1, 2010, pp. 57–74.

[12] MINAmI website. www.fp6-minami.org, accessed 6 July 2010.

[13] I. Jantunen, H. Laine, P. Huuskonen, D. Trossen, and V. Ermolov, "Smart sensor architecture for mobile-terminal-centric ambient intelligence", Sens. Actuators A, vol. 142, 2008, pp. 352–360.

[14] Y. Têtu, I. Jantunen, B. Gomez, and S. Robinet, "Mobile-phone-readable 2.45GHz passive digital sensor tag", Proc. RFID 2009. IEEE, 2009, pp. 88–94.

[15] NoTA World. www.notaworld.org, accessed 6 July 2010.

[16] K. Keinänen, J. Leino, and J. Suomalainen, "Developing keyboard service for NoTA". VTT, Espoo, Finland. Tech. Rep. 2008.

[17] MIPI website. www.mipi.org, accessed 6 July 2010.

[18] S. Asthana, and D.N. Kalofonos, "The problem of Bluetooth pollution and accelerating connectivity in Bluetooth ad-hoc networks". Proc. PerCom 2005. IEEE, 2005, pp. 200–207.

[19] J. Jantunen, A. Lappeteläinen, J. Arponen, A. Pärssinen, M. Pelissier, B. Gomez, and J.A. Keignart, "New symmetric transceiver architecture for pulsed short-range communication". Proc. GLOBECOM 2008. IEEE, 2008, pp. 1–5.

[20] S.R. Aedudodla, S. Vijayakumaran, and T.F. Wong, "Timing acquisition in ultra-wideband communication systems". IEEE Trans. Veh. Technol., vol. 54, no. 5, 2005, pp. 1570–1583.

[21] M. Pelissier, B. Gomez, G. Masson, S. Dia, M. Gary, J. Jantunen, J. Arponen, and J. Varteva, "112Mb/s full duplex remotely-powered impulse-UWB RFID transceiver for wireless NV-memory applications". Proc. 2010 Symposium of VLSI Circuits.

[22] G.W. Burr, B.N. Kurdi, J.C. Scott, C.H. Lam, K. Gopalakrishnan, and R.S. Shenoy, "Overview of candidate device technologies for storage-class memory". IBM J. Res. Dev., vol. 52, no. 4, 2008, pp. 449–464.

[23] "Intel, STMicroelectronics deliver industry's first phase change memory prototypes". Physorg.com, 6 Feb 2008.

[24] Z. Shelby, P. Mahonen, J. Riihijärvi, O. Raivio, and P. Huuskonen, "NanoIP: the zen of embedded networking", Proc. ICC 2003. IEEE, 2003, vol. 2, pp. 1218–1222.

[25] M. Cooper, "Explore and Share – Nokia shows ultra-fast wireless data transfer concept". Nokia Conversations, 23 Feb 2010.

 ISBN: 978-1-61208-100-7

Performance Comparison of Video Traffic Over WLAN IEEE 802.11e and IEEE 802.11n

Teuku Yuliar Arif
Department of Electrical Engineering,
Faculty of Engineering, University of Indonesia
Kampus Baru UI Depok 16424 Indonesia
email: t.yuliar@ui.ac.id

Riri Fitri Sari
Department of Electrical Engineering,
Faculty of Engineering, University of Indonesia
Kampus Baru UI Depok 16424 Indonesia
email: riri@ui.ac.id

Abstract— **This paper reviews the fast deployment of Wireless Local Area Networks (WLANs) and the ability of WLAN to support real time services. Stringent quality of service (QoS) and high throughput requirements has come into force. We compare the QoS support in the IEEE 802.11e to 802.11n standard. The 802.11n frame aggregation mechanism provides a better video traffic transmission performance such as throughput, delay and packet lost. The 802.11e mechanism allows prioritized medium access for applications with high QoS requirements by assigning different priorities to its four access categories. The 802.11n implemented frame aggregation to get high throughput and low delay transmission. We evaluate the performance of both 802.11 standards by implementing real time audio and video traffic using Network Simulator-2 (NS 2) simulation. Parameters such as throughput mean delay and packet lost have been calculated and graphs have been plotted. Simulation results show that 802.11e mechanism provides satisfactory service differentiation among its four access categories. With frame aggregation mechanism in 802.11n, network delay has been effectively decreased to better support real-time audio and video transmissions**

Keywords – Performance Comparison; 802.11e; 802.11n; Throughput

I. INTRODUCTION

In recent years, IEEE 802.11 standard has emerged as the dominating technology and is vastly used in Wireless Local Area Networks (WLANs) [1]. Low cost, ease of deployment and mobility support has resulted in the vast popularity of IEEE 802.11 WLANs. WLAN can be easily deployed in hot-spot zones of airports, hotels, office, and residence homes. With ever increasing popularity of multimedia applications, people want voice, audio and video services through WLAN connections. Unlike the traditional best effort data applications, multimedia applications require quality of service (QoS) support such as guaranteed bandwidth, delay and packet lost. The legacy 802.11, 802.11b, 802.11a/g can provide up to 2 Mbps, 11 Mbps and 54 Mbps data rates. However, the achievable throughput of a WLAN is less than half of the physical layer (PHY) raw data rate because of the protocol overheads, (UDP, TCP, IP, medium access control (MAC), physical (PHY) preamble, interframe spaces (IFSs), acknowledgment (ACK) and backoff time, etc. As both the MAC layer and the PHY layer of 802.11 [2] are designed for best effort data transmissions, the original 802.11 standard does not take QoS into account. Hence to provide QoS support IEEE 802.11 standard group has specified a new IEEE 802.11e standard. IEEE 802.11e supports QoS by providing differentiated classes of service in MAC layer; it also enhances the physical layer so that it can deliver time sensitive multimedia traffic, in addition to traditional data packets [3].

The IEEE 802.11e standard introduces the Hybrid Coordination Function (HCF) as the MAC scheme. While backward compatible with Distributed Coordination Fuction (DCF) and PCF, HCF provides stations with prioritized and parameterized QoS access to the wireless medium. HCF combines aspects of both the contention-based and the contention free access methods, where the contention-based channel access mechanism in HCF is known as the Enhanced Distributed Channel Access (EDCA) and its contention free counterpart is known as the HCF Controlled Channel Access (HCCA).The EDCA is an extension of the conventional distributed coordination function (DCF) [3]. It provides prioritized QoS services which classify all the traffics destined MAC layer to multiple Access Categories (ACs). Also differentiate the chance to get a transmission opportunity (TXOP) using unequal channel access parameters.

In response to the demand for higher performance WLANs to support multimedia applications such as voice and video, the standard group has specified a new IEEE 802.11n standard to provide over 100 Mbps throughput at the MAC data Service Access Point (SAP) via PHY and MAC enhancement [5]. An IEEE 802.11n WLAN can operate with physical layer raw data rate up to 200-600 Mbps by using Multiple-Input Multiple-Output (MIMO) technology, modified encoding and optional channel binding scheme. To efficiently improve the SAP throughput, two main MAC enhancement mechanisms have been proposed to reduce the protocol overhead (1) frame aggregation and (2) bidirectional transmission [6]. These mechanisms eliminate the need to initiate a transmission for every MAC frame in the legacy 802.11 and thus reduce the transmission overheads and improve the throughput efficiency. In our work, we compare the performance of 802.11e to 802.11n on accommodate video traffic.

 ISBN: 978-1-61208-100-7

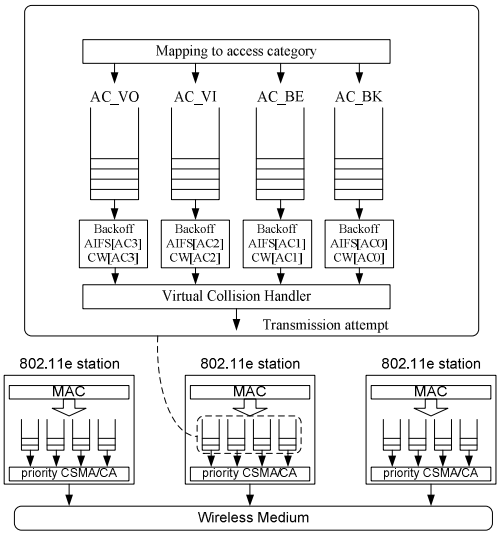


Figure 1. Four access categories in IEEE 802.11e [6].

TABLE 1. 802.11E EDCA PARAMETER SET

Priority	AC	Designation	AIFSN	CWmin	CWmax
3	AC_VO	Voice	2	7	15
2	AC_VI	Video	2	15	31
1	AC_BE	Best Effort	3	31	1023
0	AC_BK	Background	7	31	1023

This paper is organized as follows; Section II describes basic theory of WLAN IEEE 802.11e and WLAN IEEE 802.11n. In Section III, we evaluate video transmission performance over 802.11e and 802.11n using ns-2 simulation and perform the performance comparison evaluation of the simulation results. Finally, Section IV concludes the paper.

II. BASIC THEORY

A. IEEE 802.11e

IEEE 802.11e EDCA is designed to enhance the 802.11 DCF (Distributed Coordination Function) mechanism by providing a distributed access method that can support service differentiation among different classes of traffic. EDCA classifies traffic into four different AC as illustrated in Figure 1 [7]. The four access categories include AC _VO (for voice traffic), AC_VI (for video traffic), AC_BE (for best effort traffic), and AC _BK (for background traffic). To simplify the notations, AC_VO assign as AC3, AC_VI as AC2, AC_BE as AC1, and AC_BK as AC0. Each AC has its own buffered queue and behaves as an independent backoff entity. The priority among ACs is then determined by AC-specific parameters, called the EDCA parameter set. The EDCA parameter set includes minimum Contention Window size (CWmin), maximum Contention Window size (CWmax), Arbitration Inter Frame Space (AIFS), and Transmission Opportunity limit (TXOPlimit). The preferred values of each mechanism parameters that the standard recommends are shown in Table 1 [7].

Figure 2 demonstrates the operations in 802.11e EDCA. To achieve differentiation, instead of using fixed DIFS (Distributed Interframe Space) as in 802.11 DCF, EDCA assigns higher priority ACs with smaller CWmin, CWmax, and AIFS to influence the successful transmission probability (statistically) in favor of high-priority ACs. The AC with the smallest AIFS has the highest priority, and a station needs to defer for its corresponding AIFS interval. The smaller the parameter values (such as AIFS, CWmin and CWmax) the greater the probability of gaining access to the medium. Each AC within a station behaves like an individual virtual station: it contends for access to the medium and independently starts its backoff procedure after detecting the channel being idle for at least an AIFS period. The backoff procedure of each AC is the same as that of DCF. When a collision occurs among different ACs within the same station, the higher priority AC is granted the opportunity to transmit, while the lower priority AC suffers from a virtual collision, similar to a real collision outside the station.

IEEE 802.11e EDCA defines a TXOPlimit as the time interval during which a particular station can initiate transmissions. During this period, defined by a starting time and a maximum duration, stations are allowed to transmit multiple data frames from the same AC continuously within the time limit defined by TXOPlimit [7]. In 802.11e EDCA the higher priority ACs have a longer TXOPlimit, while lower priority ACs have a shorter TXOPlimit. Priority differentiation used by EDCA ensures better service to high priority class while offering a minimum service for low priority traffic [8]. Although this mechanism improves the quality of service of real-time traffic, the performance obtained is not optimal since EDCA parameters cannot be adapted according to the network conditions.

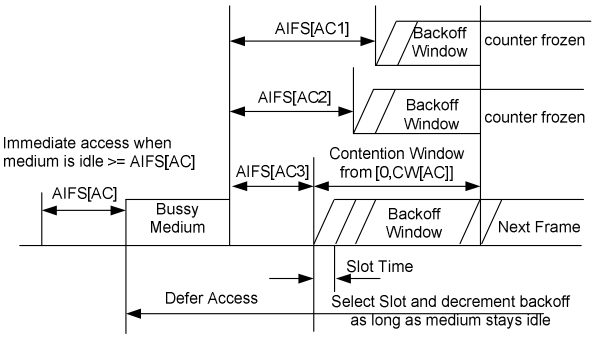


Figure 2. IEEE 802.11e EDCA mechanism [6].

B. IEEE 802.11n

Although 802.11e adds the support of QoS, TXOP and block ACK, the inefficiency of channel utilization in legacy 802.11 MAC is not fully solved. To satisfy the need of the

high-speed wireless network access today, the major target of IEEE 802.11n is to provide high throughput mechanism while allowing the coexistence of legacy 802.11 devices. To meet the requirements of high throughput, two possible methods can be applied. One is increasing the data rate in the physical layer (PHY layer), and the other is increasing the efficiency in the medium access layer (MAC layer) [9]. Based on the foundation of 802.11a/b/g/e, numerous new features in PHY and MAC layers are introduced to enhance the throughput of IEEE 802.11n WLAN [10].

To achieve high throughput in 802.11 wireless networks, the most commonly used method is to increase the raw data rate in the PHY layer. Legacy 802.11 PHY layer uses Single-Input Single-Output (SISO) system in 20 MHz bandwidth channel with one antenna. IEEE 802.11n expands the channel bandwidth to 40MHz to increase the channel capacity, and operates in OFDM scheme with the Multi-Input Multi-Output (MIMO) technique [9].

Aggregation mechanism is the key feature to improve the 802.11 MAC transmission efficiency. Aggregation can enhance efficiency and channel utilization. The aggregation mechanism combines multiple data packets from the upper layer into one larger aggregated data frame for transmission [11]. Overhead in multiple frame transmissions is reduced since the header overhead and interframe time is saved. Aggregation scheme achieves higher system gain for application scenarios with small packets, for example, VoIP.

Some frame aggregation mechanisms are illustrated in Figure 3 [6]. In Figure 3(a), a train of N PHY frames are sent one by one with no IFS. These frames can be transmitted to one or multiple destinations, and each destination station acknowledges the received frame in the same order after a short IFS (SIFS). In Figure 3(b), each destination station sends an ACK immediately after a SIFS when it successfully receives a frame.

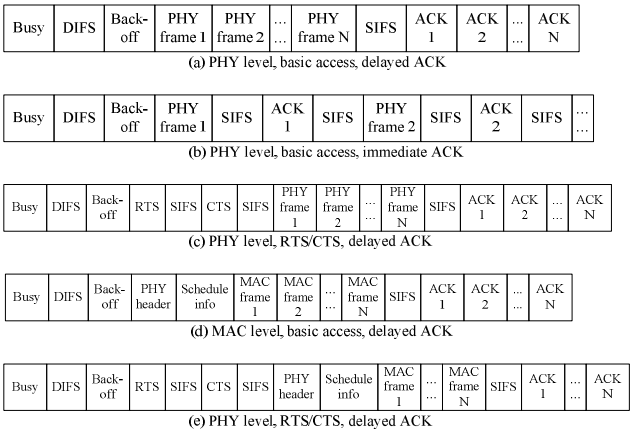


Figure 3. IEEE 802.11n Frame aggregation mechanisms [6].

Maximizing throughput may require a large aggregation frame with length longer than that specified in standard (4095 bytes). On the other hand, it is suggested that the total length of the aggregation frame should be smaller than a threshold since some huge frames may cause unfairness among stations. In addition, long data frames will result in large collision time and thus reduce the transmission efficiency when collision probability is high. In legacy 802.11, the optional Request To Send/Clear To Send (RTS/CTS) is proposed to improve the transmission efficiency when the frame size is larger than a threshold (0–2347 bytes). However, RTS/CTS in legacy 802.11 is employed by a pair of sender and receiver for unicast transmission and is not suitable for the downlink aggregation mechanism which may involve multiple destination stations. Therefore, modified RTS/CTS function can be used with downlink aggregation to reduce collisions resulting from large data frames, as shown in Figure 3(c).

The above three mechanisms are PHY level aggregations. The PHY overhead can be reduced through MAC level aggregations, which are shown in Figure 3(d) and 3(e) for basic access mode and RTS/CTS mode, respectively. With these two mechanisms, N MAC frames for different destinations can be aggregated into one PHY frame [6]. After the (shared) PHY preamble and header, destination stations receive the scheduling information, based on which they can determine the time to receive the MFs if there is any. Using downlink multi-destination aggregation, the AP only needs to contend once to transmit an aggregated frame to multiple MNs, in contrast to multiple contentions and transmissions without frame aggregation.

III. EXPERIMENTAL RESULTS

The purpose of this paper is to evaluate the comparison of the traffic video over IEEE 802.11e and over IEEE 802.11n. All simulation is conducted with ns-2 [12], where to simulate 802.11e we use modules from NKCU Taiwan based on ns-2.28 [13] and to simulate 802.11n we use AFR modules [14] from Hamilton Institute Ireland based on ns-2.30 [15]. Figure 4 shows the topology configuration used in our simulation. The topology consists of a multimedia server that connects to a WLAN Access Point (AP); an AP connects to a mobile node using 802.11e or 802.11n.

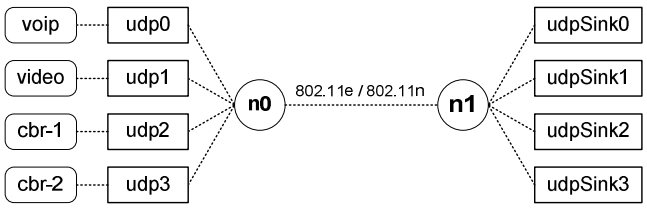


Figure 4. Simulation topology.

The performance of 802.11e and 802.11n can be evaluated from the receiving traffic through the wirelessly connected multimedia server (n0) and mobile node (n1). Several major metrics to compare performance between 802.11e and 802.11n are:

- Throughput, the traffic size through a link in a selected range of time, where :

$$Throughput = \frac{\sum packet}{\sum time} \times 8/1000 \quad Kbps \qquad (1)$$

- Delay, the mean of times in receiving side due the different received for every packet, where :

$$Delay = \sum_{i=1}^{m} \frac{(t_i^{recived} - t_i^{send}}{m} \quad (ms) \qquad (2)$$

- Packet lost, the percentage of lost of packet when received at receiving side compare to transmit packets where :

$$Packetlost = \left[\frac{\sum packet_{send} - \sum packet_{received}}{\sum packet_{send}} \right] \times 100\% \qquad (3)$$

A. Simulation Setup

In this section, we use ns-2 simulator to evaluate the performance of IEEE 802.11e and IEEE 802.11n. We choose 802.11 as the PHY layer, and the PHY data rate is set to 1 Mbps, 11 Mbps and 54 Mbps. The simulation parameters are shown in the Table 2.

In our simulation we have considered three scenarios, namely scenario 1, scenario 2 and scenario 3. In each scenario all the stations are transmitting to the same destination. Scenario 1 and 2 consist of one VoIP connection, one video connection and two connections each of background traffic and best effort data. We use scenario 1 to evaluate the performance of IEEE 802.11e and scenario 2 to evaluate the performance of IEEE 802.11n. In scenario 3 we increased the video transmission rate to be known maximum throughput of 802.11e and 802.11n. The best-effort and background traffics have been created using CBR traffic with the sending rate of 256 Kbps. Consistent with 802.11e specifications, VoIP traffic is carried under AC1, video under AC2, background traffic under AC3 and best effort data under AC4. In every scenario, the video traffic starts at 5 secs,VoIP traffic starts at 0.1 sec, video at 10 secs, BK traffic starts at 20 secs and BE traffic starts at 15 secs.

B. 802.11e and 802.11n performance

We compare the performance of 802.11e and 802.11n mechanism by simulating the scenario 1 and scenario 2, having one VoIP connections, one video connection and two BK/BE connections each.

By comparing the Figure 5 and Figure 6, which plot the throughput of each traffic type, we observe that the throughputs of video and BE/BK data are significantly different from 802.11e and 802.11n, whereas the VoIP traffic is able to maintain its throughput in both cases. In Figure 5, we can observe that the throughput of video traffic drops from around 512 kbps to 400 kbps but still can get the throughput upto 500 kbps. This confirms that the video traffic is well served with the implemented QoS in 802.11e, while many video frames are dropped at the 802.11n where the throughput drops from 512Kbps to around 400kbps because 802.11n does not implement QoS. It can also be seen that the throughput of BE/BK traffic is low in 802.11e as compared to 802.11n because BE/BK have low traffic priority parameters.

TABLE 2. SIMULATION PARAMETERS

	Voice	Video	Background	Best Effort
Transport protocol	UDP	UDP	UDP	UDP
Access Category	AC1	AC2	AC3	AC4
Packet Size	160 bytes	1500 bytes	1500 bytes	1500 bytes
Sending rate	64 Kbps	512 Kbps	256 Kbps	256 Kbps

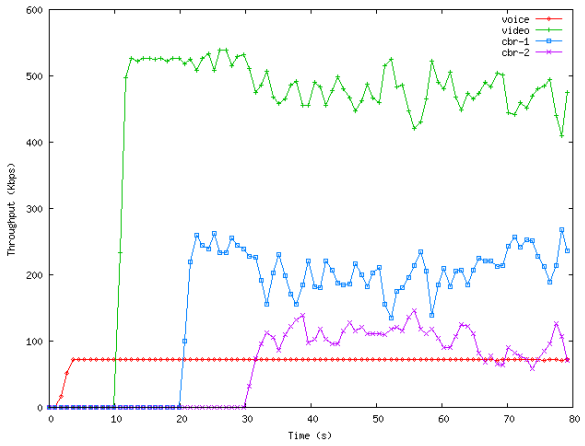


Figure 5. IEEE 802.11e throughput.

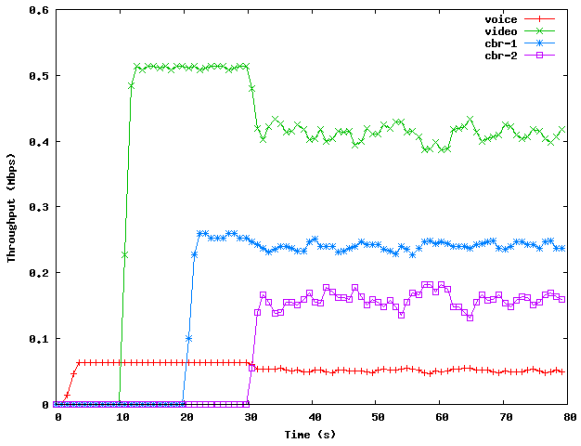


Figure 6. IEEE 802.11n throughput.

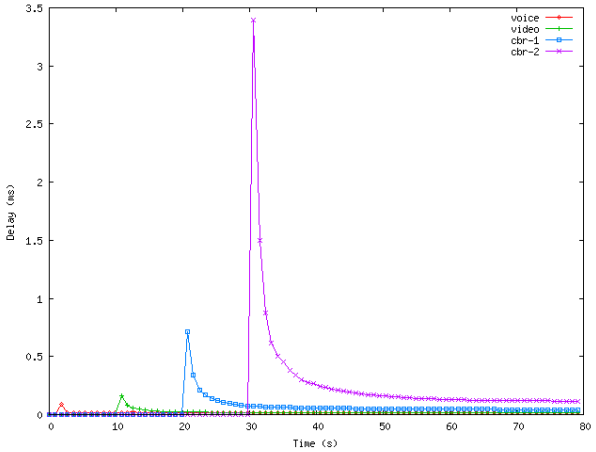


Figure 7. IEEE 802.11e delay.

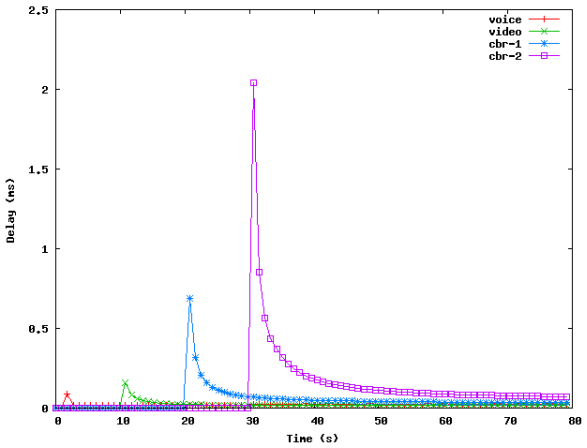


Figure 8. IEEE 802.11n delay.

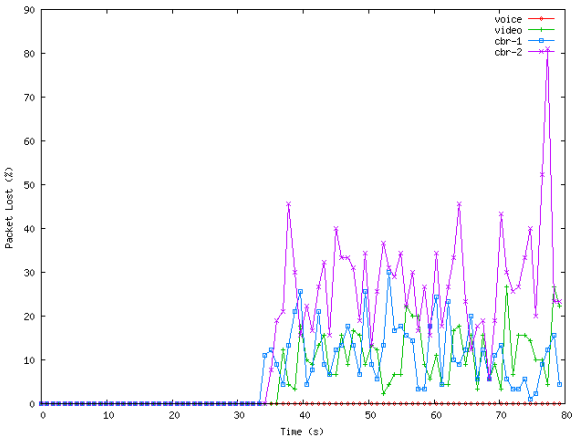


Figure 9. IEEE 802.11e Packet lost.

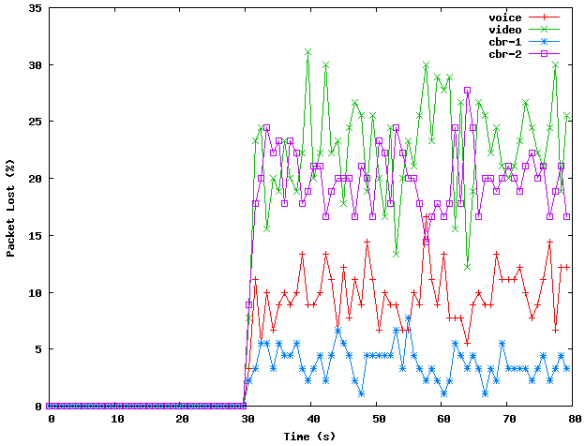


Figure 10. IEEE 802.11n packet lost.

In Figure 7 and Figure 8, we observe that VoIP and video delay performance is improved via 802.11n. We can see that when the BE/BK traffic starts at 20 secs and 30 secs, the voice frame delay and video frame delay have not increased in 802.11n as compared with 802.11e. Note that with 802.11n, the voice frame delay and video frame delay have better performance compared with 802.11e. It can also be seen that the delay for video traffic has improved in 802.11n as compared to 802.11e when all the traffic flows exist in the network. The delay for BE/BK traffic is also better in the 802.11n compared with 802.11e. In Figure 9 we observe that VoIP packet lost via 802.11e is the average zero percent compared to Figure 10 where VoIP packet lost drop to ten percent via 802.11n. Video packet lost is increased via 802.11n and compared to 802.11e.

These simulation results show that there is no service differentiation between the different types of traffic flows in 802.11n, which causes the QoS problem for multimedia applications when traffic load is high. The 802.11e mechanism provides differentiated channel access for different traffic types and can be expected that the 802.11e can support real-time applications with voice and video traffic with a reasonable quality of service.

C. *Comparison Analysis*

First we consider the scenario 1 and scenario 2, consisting of one VoIP connections, one video connection and two connections of each background traffic and best effort data. As mentioned above, the applications were start at different times so as to illustrate the impact of additional traffic streams on existing load. Figure 7 and Figure 8 show the delay performance of these traffic streams. The delay for VoIP frames is small (less than 1ms) from 0s to 0.1 ms, as it is the only traffic in the network so that it does not have to contend the channel with other sources. With the introduction of video traffic at 10ms, the delay for video frames increase to 0.2 ms whereas the delay for VoIP traffic is about 0.1 ms. It can be observed that when the BK/BE traffic is started at 20 secs and 30 secs, the delay for video and VoIP does not increased.

Next we simulate the scenario 3, in which we decrease the data rate from 1 Mbps to 0.5 Mbps. In Figure 11 the impact of decreasing the highest priority video connections can be seen on the delay performance of low priority traffic, when all the traffic streams present. The delay for video

 ISBN: 978-1-61208-100-7

frames increases to 0.25s via 802.11e as compared to 0.15s via 802.11n, also the delay for video traffic via 802.11e at data rate 1 Mbps relatively same as via 802.11n. Thus, the impact of decreasing data rate can be seen on the delay performance of video traffic over 802.11e.

In scenario 3, we decrease the data rate of 802.11e and 802.11n to 0.5 Mbps. In figure 11 we observe that the decrease in low priority traffic does not have any negative impact on the delay of higher priority traffic. It can be seen that the delay for VoIP and video traffic is nearly same for both low BK/BE traffic and high BK/BE traffic. Comparing to data rate load change, decreases data rate in 802.11n load does not affect video delay in Figure 11 compare to decreases data rate in 802.11e, video delay relatively increase.

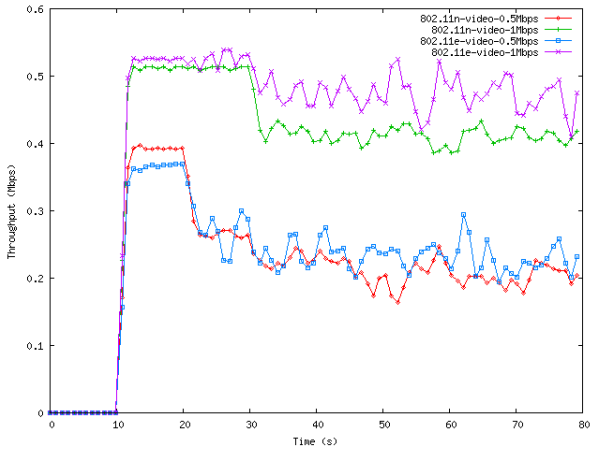


Figure 10. Throughput comparison of 802.11e and 802.11n.

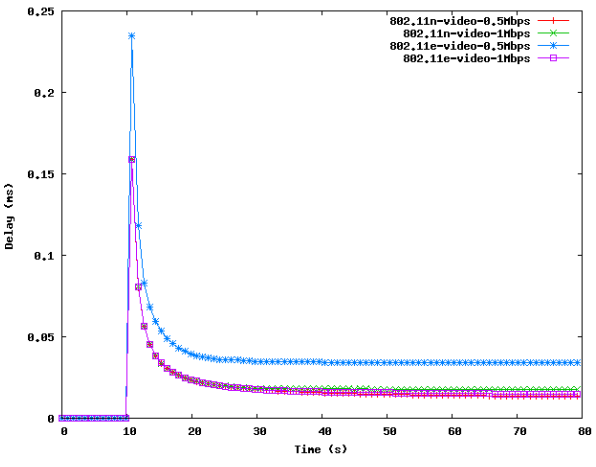


Figure 11. Delay comparison of 802.11e and 802.11n

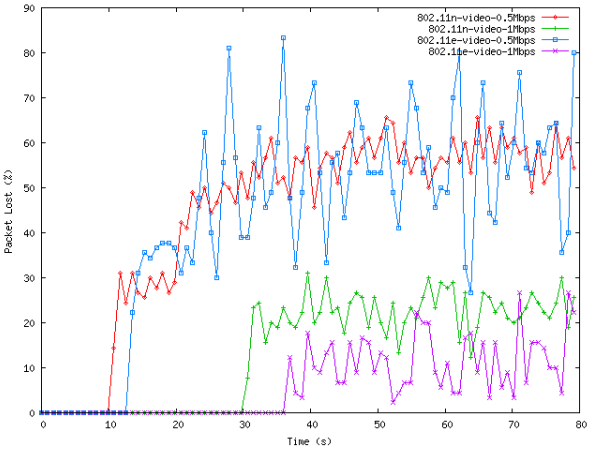


Figure 12. Packet lost comparison of 802.11e and 802.11n.

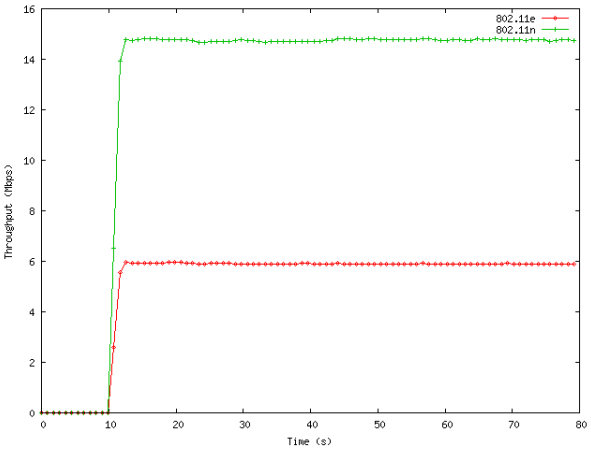


Figure 13. Maximum throughput comparison of 802.11e and 802.11n.

IV. CONCLUSION

In this paper, we have evaluated the 802.11e and 802.11n mechanisms to transmit video traffic over WLAN. Through our simulations, we compared the 802.11e with the 802.11n in order to show that 802.11e provides differentiated channel access for different traffic types and is better equipped than 802.11n to handle real time applications with stringent QoS requirements likes VoIP and video. We conclude that with heavily loaded traffic connections under non-negligible background traffic, the 802.11n mechanism is not able to provide QoS guarantee. However, it can give better delay performance compared to 802.11e.

Our simulation result shows that 802.11n have better total throughput of 15 Mbps compared with 802.11e which only has 6 Mbps. We found out that 802.11n must implement QoS mechanism to support video to get a stable throughput during transmission.

 ISBN: 978-1-61208-100-7

REFERENCES

[1] T.S. Rappaport, A. Annamalai, R.M. Buehrer, and W.H. Tranter, "Wireless communications: past events and a future perspective", Communications Magazine, IEEE, Vol. 40, No. 5, pp. 148-161, May 2002.

[2] IEEE Std. 802.11-1999, Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications, Reference number ISO/IEC 8802-11:1999(E), IEEE Std. 802.11, 1999 edition, 1999.

[3] IEEE Std 802.11e-2005; "Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications Amendment 8: Medium Access Control (MAC) Quality of Service Enhancements", November 2005

[4] H. Dheeraj, P.B. Revoti, and S. Saurabh, "Performance Analysis of QoS supported by Enhanced Distributed Channel Access (EDCA) mechanism in IEEE 802.11e", International Association of Engineers, vol 33, issue 1, Feb 2007.

[5] IEEE 802.11n-2009-Amendment 5: Enhancements for Higher Throughput, 29 October 2009.

[6] L.X. Cai, X. Ling, X. Shen, J.W. Mark, and L. Cai, "Supporting Voice and Video Applications over IEEE 802.11n WLANs", ACM Wireless Networks (WINET), vol. 15, no. 4, pp. 443-454, May, 2009.

[7] Evaluation of video stream quality over IEEE 802.11e EDCF, http://140.116.72.80/%7Ejhlin5/ns2/802_11e/NS-2_80211e.htm, last access 3 May 2010.

[8] R.F. Sari, Y. Maraden, and K. Djunaedi, "Performance Evaluation of IEEE 802.11e EDCA based on Variable Priority Parameters", Proceedings of Quality in Research (QIR) 2007 Conference, Jakarta, 4-5 Desember 2007.

[9] D. Skordoulis, Q. Ni, H.H. Chen, A.P. Stephens, C. Liu, and A. Jamalipour, "IEEE 802.11n MAC Frame Aggregation Mechanisms For Next Generation High Throughput WLANs", IEEE Wireless Commun., vol. 3, no.1, pp. 40-47, Feb. 2008.

[10] C. Wang and H. Wei, "IEEE 802.11n MAC Enhancement and Performance Evaluation," ACM/Springer Mobile Networks and Applications Journal (MONET), Volume 14, Issue 6, Page 760-771, Dec. 2009.

[11] Lin Y. and Wong VWS, "Frame aggregation and optimal frame size adaptation for IEEE 802.11n WLANs", Proceedings of IEEE Global Telecommunications Conference (GLOBECOM), San Fransisco, CA, U.S.A., November 2006.

[12] NS-2, http://www.isi.edu/nsnam/ns/, last access 24 April 2010.

[13] myNS2, http://140.116.72.80/%7Esmallko/ns2/mysetup_en.htm, last access 24 April 2010.

[14] T. Li, Q. Ni, D. Malone, D. Leith , T. Turletti, and Y. Xiao, "Aggregation with fragment retransmission for very high-speed WLANs", IEEE/ACM Transactions on Networking, pp. 591-604, April 2009.

[15] AFR Implementation, http://www.hamilton.ie/tianji_li/afr.html, last access 3 Mai 2010.

 ISBN: 978-1-61208-100-7

Towards Self-Adaptable, Scalable, Dependable and Energy Efficient Networks: The Self-Growing Concept

N. Alonistioti[1], A. Merentitis[1], M. Stamatelatos[1], E. Schulz[2], C. Zhou[2], G. Koudouridis[3], B. Bochow[4], M. Schuster[4], P. Demeester[5], P. Ballon[5], S. Delaere[5], M. Mueck[6], C. Drewes[6], L. Van der Perre[7], J. Declerck[7], T. Lewis[8], and I. Chochliouros[9]

[1] *Department of Informatics & Telecommunications, University of Athens, Greece – {nancy, amer, makiss}@di.uoa.gr*
[2] *Huawei Technologies Duesseldorf GmbH, Germany – {egon.schulz, chan.zhou}@huawei.com*
[3] *Huawei Technologies Sweden AB, Sweden – george.koudouridis@huawei.com*
[4] *Fraunhofer-Gesellschaft zur Förderung der angewandten Forschung e.V., Germany – bernd.bochow@fokus.fraunhofer.de*
[5] *Interdisciplinary Institute for Broadband Technology, Belgium – {pieter.ballon, simon.delaere}@vub.ac.be*
[6] *Infineon Technologies, Germany – {markusdominik.mueck, christian.drewes}@infineon.com*
[7] *Interuniversitair Micro-Electronika Centrum VZW, Belgium – {vdperre, dclerckj}@imec.be*
[8] *Toshiba Research Europe Ltd, United Kingdom – tim.lewis@toshiba-trel.com*
[9] *Hellenic Telecommunications Organisation S.A, Greece – ichochliouros@oteresearch.gr*

***Abstract*— In next generation systems and networks, the incorporation of mechanisms achieving robust, predictable and self-adaptive behavior with minimum cost will be a key requirement. Towards this goal we introduce the notion of the "self-growing network". The latter, in its initial deployment stage, is limited to a single dedicated purpose (energy-efficient networking, spectrum efficient communications, control and surveillance use, etc) but can evolve/grow into a multi-purpose, versatile infrastructure, that serves a broader range of applications by utilising combinations of self-x and cooperating features. The self-growing network paradigm considers (i) mechanisms for energy efficient interaction of the wireless network elements and (ii) mechanisms for the reliable and efficient evolvement towards later lifecycle phases. The Self-growing system incorporates the network(s) as well as user services and applications thus creating self-growing solutions applicable to wide range of purposes and impacting various beneficiaries, such as providers, consumers and end users.**

Keywords-self-growing; self-adaptation; cooperation, distributed systems; low energy

I. INTRODUCTION

In future large-scale distributed systems, the emergence of mechanisms achieving robust, predictable and self-adaptive behaviour will be an important evolution step. At the same time, as systems get more complex in terms of scale and functionality, reliability and dependability are getting increasingly important and self-adaptation techniques for achieving dependable system operation under cost and energy constraints will be a key concept. In this context, key challenges lie in the efficient cooperation of heterogeneous elements in order to provide advanced problem solving capabilities and improved as well as reliable services.

Furthermore, innovations for low energy are considered a fundamental parameter in the efforts to combat climate change and to achieve sustainable economic growth. Low energy solutions create an attractive business case by offering significant benefits in terms of operational cost, long-term product reliability, sustainability, and increased lifetime of wireless or mobile elements. For this purpose, a promising path lies in the study and development of energy-aware distributed and cooperating systems for monitoring and control, in particular based on wireless networks for providing radio and environmental context information.

Current wireless network development is driven by horizontal mass-markets ("one size fits all"). Vertical markets and niche applications demand for (costly) dedicated configurations or developments. Consequently, the evolution of a wireless network often demands for infrastructure and terminal replacement. Extending system and network capabilities, switching services or switching the purpose of an operational network usually requires costly (manual) reconfigurations and upgrades, while usually results in temporary unavailability of system services. Promising solutions for these problems are expected to require foundational multi-disciplinary research, leading to an integration of next generation technologies. Such integration is leveraging on capabilities for spontaneous ad-hoc cooperation between objects, self-adaptive behaviour, exploitation of dynamic information, predictability of non-functional properties (e.g., energy consumption), etc.

Furthermore, the constantly rising complexity of such dynamically changing network infrastructures can only be managed and maintained by highly trained professionals – making the requirements for self-growing and self-adapting wireless infrastructures an absolute must in order to ensure a large scale deployment.

As a summary, *energy efficient and dependable operation at the level of cooperating wireless elements, network compartments and networks as a whole is becoming an increasingly difficult objective, given the ever-increasing complexity in heterogeneous telecommunication environments*. In this context, the evolution of mechanisms to cope with energy-aware and dependable cooperation of wireless elements becomes a fundamental enabler for future heterogeneous large scale distributed systems.

 ISBN: 978-1-61208-100-7

The rest of the paper is organized as follows. The concept of the self-growing network is introduced in Section II. Section III presents the considered uses cases and elaborates on the potential benefits for every use case. Finally, Section IV concludes the paper.

II. THE CONCEPT OF THE SELF-GROWING NETWORK

In order to provide tangible solutions for the challenges discussed in the previous section, it is necessary to progress in two major research directions:

- Solutions for optimised energy consumption, adaptability and dependability in a small scale, purpose-driven network through balancing autonomic and cooperative approaches,
- Mechanisms for the self-evolvement of the network/system, towards a large-scale, multi-purpose network/system.

At the beginning of its lifecycle, defined in this paper as the progression through a series of differing stages of development which can potentially provide different services, a **self-growing network** (Figure 1) is set up on-demand, dedicated to a single purpose. Relevant use cases might be for example monitoring and/or controlling applications with a focus on distributed and cooperating systems, including construction sites, delivering wireless services within a complex home/office environment requiring network parameter negotiation with a multitude of neighbouring networks, etc. These applications usually have strict requirements in terms of energy consumption, thus *solutions for optimised energy consumption* are exploited. During the self-growing networks' lifecycle, it can evolve to serve several different objectives as needed utilizing *mechanisms for self-evolvement.*

The considered evolution may include for example providing general voice and data communications, integrating sensor networks in the vicinity, or supporting safety of life applications under exceptional situations. In the course of this it may coexist and cooperate with other wireless networks of distinct owners and interest groups evolving in the deployment area towards using or augmenting existing capacity. The sensor networks in particular may serve a multitude of purposes, including environmental sensing, radio parameter sensing, etc. Interoperability with existing network infrastructures and wireless standards enables geographical and/or functional on-demand extensions. Towards the end of its lifecycle, the self-growing network may still remain active and serve as a dedicated purpose network or as a failover for applications associated with other networks sharing the same area.

The self-growing concept incorporates both **collaborative** and **autonomic** aspects. Specifically, cooperative behaviour and problem solving is critical in the self-growing initial stage, the small-scale network, as well as in the evolvement to a larger scale network, able to serve different purposes and larger systems. For example, the combination of cooperation paradigms with the inherent redundancy of monitoring and control functionalities in distributed wireless networks not only allows the objects to dynamically adhere to the most energy efficient pattern but also provides error resiliency features, graceful performance degradation in the presence of faults, as well as efficient utilisation of redundancy for fault tolerance that are very important for large scale networks.

Moreover, autonomic capabilities constitute key enablers for self-growing network paradigms. In this context, the degree of autonomicity in an object (network device, network compartment/cluster, as well as a network and the system as a whole) is balanced against the requirement for efficient cooperation in order to maximize the gains in both energy efficiency and dependability. At the end of the day, the mechanisms and the enablers for the self-growing concept from the small scale to the large scale might form a complete toolbox that can be applied to a wide range of application areas or network purposes. It is expected that self-growing capabilities coupled with autonomic and self-adaptation features will pave the way for scalable, energy efficient heterogeneous wireless networks thus impacting various beneficiaries spanning from value providers to end users.

The proposed approach pursues to increase *dependability***,** *cost* and *energy efficiency*, and also *flexibility*, *resilience*, and *robustness* of a heterogeneous wireless network by utilizing reconfigurable wireless communication nodes and distributed cooperative control functions. In contrast, existing solutions that are optimised for a single purpose are expensive and lack flexibility; flexibility would for example allow creating hybrid solutions without significant effort for incorporating additional network and service gateway functions to achieve interoperability.

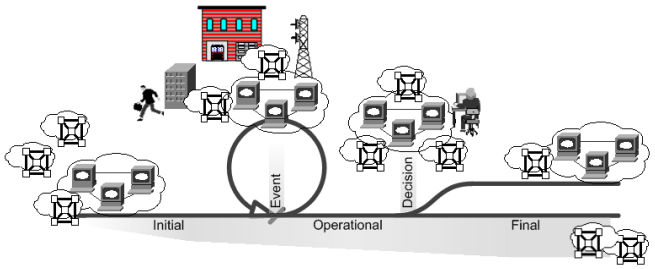


Figure 1: Self-Growing Network Life Cycle.

Regarding **self-evolvement**, the envisaged methodology resembles the structure and approach discussed in the context of self-growing neural networks presented in e.g., [1],[2]. This concept allows initiating a learning process from a very limited training space and supports response to an extension of the training space later on by "growing" the capacity or number of decision elements. A practical implementation of this concept could be repository/ontology based by providing a database of decision making capabilities that is dynamically composed of contributions of participating networks. A similar approach has been presented in the context of self-growing robot control software to respond to changes in service requirements [3]. Leveraging on these notions we will identify and develop a suitable cognitive architecture that will enable participating networks to reconfigure their topologies and optimisation goals on

 ISBN: 978-1-61208-100-7

demand. Furthermore, we will develop and formalize the decision making framework as well as its related baseline functions in order to enable describing, planning and controlling the targeted behaviour of participating networks in the process of self-growing in a concise and scalable way.

III. USE CASES

Three major uses cases are given next to emphasize on potential benefits by applying the approach in certain scenarios and to provide concise examples of possible implementations of the approach. These use cases intend to underline the toolbox nature of the approach mentioned earlier. It is shown that the use cases given here can benefit even though the approach may be applied only to a limited extend.

A. Construction Sites

The first use case addresses both wide-area and in-facility construction sites as well as moving work zones. Network evolution is achieved by deploying heterogeneous equipment and by continuously updating operational policies. This is an example where the network grows by cooperating, collaborating and integrating with neighbouring networks. Network deployment and purposes can be planned and optimised prior to start deploying equipment. Moving work zones are considered a special application of this concept in that they usually rely on short-living configurations with respect to a given geographical area. Road, railway or inland waterway construction scenarios may apply, explicitly requiring the use of mobile nodes and networks. Additionally, first-responder or military scenarios apply where autonomous vehicles pave the way deploying sensors and communication repeaters on their path.

In the initial stage of the deployment of the self-growing network, network equipment is deployed whenever needed for the given purpose and provides sensor network and surveillance services as well as a first step towards a large scale, distributed sensor network. Equipment is mostly dedicated to linking sensor (and actuator) functions, or to provide simple point-to-point and broadcast communications. In case of moving work zones nodes may need to support focused environmental monitoring in addition to RF and network monitoring.

In intermediate stages of deployment, the network is augmented by voice and data communication, positioning, machinery monitoring and control services and supports safety of life applications, e.g., for construction site workers or emergency teams. This can be achieved by integrating with, for example, neighbouring wireless local area networks or even microcells (by spontaneous ad-hoc cooperation between objects). Moving work zones will demand for more network flexibility in terms of environmental or radio scene adaptation capabilities since planning may not be possible to the same degree as applicable to permanent sites.

In the final deployment stage, network equipment might be left embedded in buildings and may provide wireless repeater functions (e.g., through elevator shafts), or may provide sensor network functionality to support facility management and monitoring. The infrastructure released by moving work zones may be utilized later on by other networks evolving in the vicinity or by vehicular networks for example. Network-centric computing with dynamic resource discovery and management will enable a seamless evolution of the network environment whenever new networking components are added.

Expected benefits of the self-growing concept in this use case include the reduced development cost for network nodes, the increased sustainability and flexibility of the system, the sharing of deployment cost, the extended lifecycle of the network, as well as the potential for less effort in planning and managing.

B. Embedded Incident Area Network

This use case addresses the deployment of a network in a limited geographical area utilizing for example a reconfigurable picocell and femtocell infrastructure as a pre-planned implementation of the self-growing network concept. In contrast to scenarios previously discussed that start from a main purpose associated with wireless sensor net functionality, this scenario starts from a general voice and data communications use case serving off-the-shelf end-systems. The network then grows with upcoming requirements from the deployment area. In case of an exceptional event on demand reconfiguration capabilities allow to switch to a purpose focusing on safety of life. The temporary switch of purpose then will designate the network, part of the network, or single network nodes to implement an incident area network until the incident is resolved. This interruption of planned operation will leave the network in a potentially undetermined state and thus requires self-adaptation for returning to pre-planned operation.

In the initial deployment stage nodes provide general voice and data communications only. The network may need to support mesh configurations (e.g., as a wireless backhaul) considering the potential need to cover larger areas with singular attachment points to a wide-area communications infrastructure. Intermediate deployment stages may extend the network into sensor nets and/or safety applications. In this direction, a continuous deployment of (heterogeneous) nodes either providing sensor/actor or communication capacity allows to optimise the coverage of the area in terms of placing functionality where it is needed to, for example, monitor dynamic geological phenomena with a suitable spatial resolution.

Under normal conditions, safety functions in this use case are mainly used in locating/tracking personnel, in health monitoring of personnel, in area monitoring (e.g., detecting a landslide) or in providing emergency call capacity where needed. In an incident case, the network can be reconfigured to guide emergency teams within the area. The reconfiguration might be flexible depending on the type of incident (e.g., focused on the location of the incident) and might be initiated manually by a network management action or automatically triggered by sensors in this location (e.g., after detecting sensor nodes going dysfunctional).

The use of off-the-shelf end-systems is beneficiary especially for this use case: safety functions are gaining from geo-location features provided by state-of-the-art handsets,

 ISBN: 978-1-61208-100-7

and in an emergency situation the network may be allowed to actively place a call, causing an acoustic signal to guide emergency response teams more accurately even in difficult environments. In the final stage of the deployment, the general voice and data communications capacity of the network can be reused, e.g., by transferring it to an operator for establishing a new managed (commercial) infrastructure, or by setting up the initial phase of a new use case scenario, for example a construction site as discussed above.

Expected benefits of the self-growing concept in this use case include the support of service centric, on-demand adaptation capabilities to respond – potentially without requiring human user interaction – to changing application requests. Furthermore, this is achieved without imposing additional cost for network development/deployment to support safety-of-life applications.

C. Self-Growing Home and Office Environment

This use case addresses the deployment of a heterogeneous wireless network in a limited geographical area, such as a home and office environment. The objective is that such a network guarantees the provision of voice and data communication services, but also acts as a large scale, distributed and cooperating system for monitoring and control, possibly incorporating Wireless Sensor Networks. The utilisation of a network entity in a home or office environment may vary a lot during a normal day, for example it may not be used at all when there is no one at home or in the office. The energy efficient parameterisation of the corresponding network is a key challenge, due to both the variation in usage and the high percentage of households that are expected to apply the corresponding concepts in the future. However, this potentially great number of adopters also implies that the inherent potential for energy savings is huge. A further challenge lies in the efficient inter-network parameterisation, to support spontaneous ad-hoc cooperation between objects and exploit network-centric computing paradigms with dynamic resource management.

In practice, the deployment of such a network is expected to be done in various stages. In an initial stage of deployment, households and/or offices will deploy a network designed for an initial estimate of capacity and Quality of Service requirements. For cost reasons, typically common-off-the-shelf equipment is employed without any consideration concerning potential needs for a future evolution of such networks. Also, its integration taking into account corresponding deployments of distinct owners and/or interest groups in the vicinity is typically not considered in such an early phase.

In an intermediate deployment stage, the initial deployment evolves with needs for higher Quality of Service and the interconnection of a constantly increasing number of devices, partially building of novel radio components (for example, UWB components on top of an available WLAN and cellular network environment). The need for an overall network integration becomes apparent taking corresponding deployments in the vicinity into account. This furthermore includes a required evolution towards a distributed and cooperating system being able to support monitoring and control, incorporating wireless sensor networks. Self-growing networks are expected to provide a framework meeting both challenges, through the optimised integration of novel devices into a given network and the efficient overall system configuration taking the overall networking environment in the vicinity into account. Specifically, the approach will provide algorithms that lead to an automated parameterisation of all network components in the managed environment as well as in its vicinity; in particular, an overall change of the network configuration is considered if novel radio components are deployed.

In the final deployment stage, the home/office environment reaches a stable level of a high-density heterogeneous network deployment. The self-growing and self-optimised evolutionary approaches have lead to a stable network deployment, ensuring a low level of overall energy consumption and an optimum exploitation of the available system capacity. Typically, such stable stages exist for some time but eventually the network evolves further, and the network thereby enters the “intermediate stage” again.

Expected benefits of the self-growing concept in this use case include the seamless evolution of a home/office network without any user involvement together with the minimization of efforts to be spent for planning and maintenance. At the same time, the proposed mechanisms facilitate an overall low-power and high capacity optimization of the network taking device deployment and parameterisation in the vicinity into account.

IV. CONCLUSION

In future large-scale distributed systems, the emergence of mechanisms supporting robust, predictable and self-adaptive behaviour will be an important evolution step. Towards this goal we have introduced the notion of a self-growing network that during its lifecycle, can evolve (grow) to serve different objectives. The CONSERN project aims at defining and developing the required mechanisms for realizing the self-growing network paradigm. From the analysis of three major uses cases it is clear that the number of affected entities (e.g. users, operators, manufacturers) is considerable and the potential impact is very significant.

ACKNOWLEDGMENT

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement CONSERN n° 257542.

REFERENCES

[1] J. K. Cios, "Self-growing neural network architecture using crisp and fuzzy entropy", Photo-Optical Instrumentation Engineers (SPIE) Conference Series, 1992.

[2] C. L. Tseng, Y. H. Chen, Y. Y Xu, H. T. Pao, and H.-C. Fu, "A self-growing probabilistic decision-based neural network with automatic data clustering", Neurocomputing 61 (2004): 21 - 38.

[3] H.-M. Koo, and I.-Y. Ko, "A Repository Framework for Self-Growing Robot Software", Asia-Pacific Software Engineering Conference 0, 2005, pp.515-524.

 ISBN: 978-1-61208-100-7

Optimum Cluster Size for Cluster-Based Communication in Wireless Sensor Network

Goutam Chakraborty
Dept. of Software & Information Science
Iwate Prefectural University, Takizawamura, Japan
Email: goutam@soft.iwate-pu.ac.jp

***Abstract*—Clustering of sensor nodes to reduce energy expense during data communication covers a large body of literature. Without clustering energy of sensor nodes near the sink drain fast, which in turn kills more rapidly nodes at further hop distances. Cluster-based routing protocol alleviates this problem. Yet, in cluster-based approaches too, for hop-by-hop communication, power of nodes nearer to the cluster head (CH) are drained more rapidly compared to those at the periphery, as they are more often used as hopping nodes. This is more so when the cluster is big. For too small cluster, there is no meaning in clustering. Uniform dissipation is achieved by reconfiguring the clusters at intervals, which is a big signalling overhead. Most of the previous works are on efficient cluster formation, and on using more than one CH to reduce SNR. In this work, we show that there is an optimum size of a cluster, for which the power dissipation at every node could be made uniform over a time, by transmitting packets at different energy levels. It is a co-operative approach for data transportation, where different portions of packets are forwarded to different nodes towards the CH. This way we can avoid frequent cluster reconfiguration. In this paper, the above goal is formally defined as a constrained optimization problem, for linear array of sensor nodes. It turns out to be a non-linear optimization problem, which is simplified to a linear optimization problem and solved. It is shown that the problem has a solution when the cluster diameter is 6 (in terms of hop count) or less. Cluster of bigger size has no solution. We also formulate the problem, when nodes are uniformly distributed over a plane.**

***Keywords*-Sensor nodes' power decay; Constrained optimization problem; Linear programming;**

I. INTRODUCTION

Regarding sensor network software, energy efficiency is the main motivation of all aspects of researches, ranging from OS [2], data acquisition[3], data dissemination/diffusion [4], query processing [5], media access control [6], communication protocol/routing [8] [7] [10] [11] to network topology [9]. In this work, our motivation is a novel energy-aware communication protocol.

Sensor network installations can be categorized into two classes according to the motivation of use

1) Individual sensor node and information it collects is important. Applications are like fire alarm, icy road condition, frosting of grape bunch (suitable for brewing ice-wine), surveillance camera, structural health of buildings and bridges, etc.. Here the sensor nodes' locations are usually fixed, though sometimes they may be moved in controlled direction remotely.
2) Individual sensor data is not important. Assembled data from a region or cluster is all that is to be delivered. The applications are mainly environmental monitoring for climate changes or like. Nodes may drift due to bad weather or flooding.

In scenario 1), it is obvious that the longevity of every sensor nodes is important. Even in scenario 2), we would like to see that most of the sensor nodes sustain for as long as possible, because information from all segments of the sensor network is equally important.

The main goal of communication protocol for wireless sensor networks (WSNs) should be (1) slower decrement of the average battery power with time, as well as (2) lower variation of the distribution of the remaining battery level. To our best of knowledge, all the works on energy aware communication protocols emphasize on item (1) above. But, in reality both (1) and (2) are important criteria.

A. Existing works and where wediffer

The WSN model used by almost all works is as shown in Fig. 2.

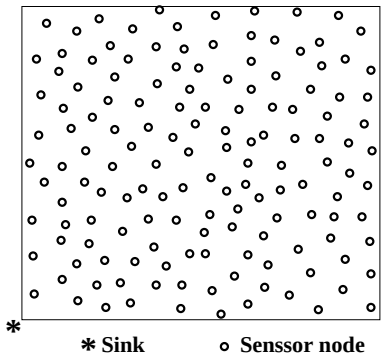


Figure 1. A WSN where sensor nodes are spread with uniform distribution, and the Sink is placed at a corner

We assume that nodes are able to transmit at different distances and control their transmission power accordingly. Assuming the power required to transmit to a distance d

 ISBN: 978-1-61208-100-7

is proportional to d^2, a direct transmission protocol will be very inefficient - considering that all nodes generate data packets at the same rate. Obviously, nodes at further distances from the "Sink" will die soon. There is the class of power aware protocols [1], where packets are routed through intermediate nodes - thus limiting the transmission range, thereby saving energy. Here, nodes near the sink will be overly loaded and die early, cascading the effect to next layers of nodes. The more conventional approach is clustering, where a cluster head (CH) collects packets from all members of the cluster, and transmits to sink through intermediate CHs. Here too, the nodes near to the CH are overloaded. LEACH protocol [10] first proposed to dynamically change the cluster configuration, so that the load is uniformly shared by nodes, over a long time. They showed much longer lifetimes for the sensor nodes.

But reconfiguring clusters is a power heavy task. In this work, we have shown that, if the nodes do not always send packets to their nearest nodes, but transmits different ratios of packets to different distances, towards the CH, the power depletion at different nodes could be made uniform. We have shown that this is possible for clusters of size up to 6 hops in diameter. We formulated this as a constrained optimization problem and solved the required ratios.

The rest of the paper is organized as follows. In Section II, we formally define our goal and corresponding optimization problem, for simple linear network. In Section III, we gave solutions to cluster of radius 2 and 3 hops. We also show that there is no solution for larger cluster that would satisfy the constraints of the problem. In Section IV, we extended the problem definition for WSN spread over a plane. Finally, discussion for further work and conclusion is in Section V.

II. Problem definition, Optimization criterion and constraints

A. Network model and assumptions

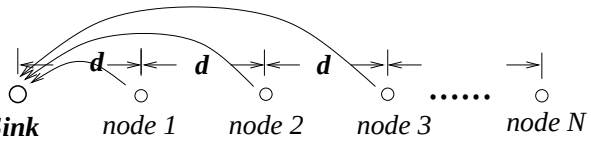


Figure 2. Simple linear node distribution

To simplify the problem, instead of a two dimensional distribution, we will start with a linear network of nodes on a straight line. We will further assume that N nodes are placed equidistant, as shown in Fig. 3. Let us consider the distance between two nodes be d. We assume that the nodes can adjust the transmission power to transmit packets to the target destination. Further, every node creates on an average same number of packets, say m, in a specific interval of time.

B. Defining the Constrained Optimization Problem

We have N-nodes at equal distances d. This includes the sink node, which is node-1. Every node generates m packets. Different proportion of these m packets are forwarded to different nodes towards the sink. In previous works, transmission is always one hop, causing overload of nodes near the CH.

Every node has to service (transmit packets to reach destination) - not only packets it generates, but also packets it receives from nodes further away from the CH. Nodes transmit different portions of its packets (own packets plus those received from behind) to different hop distances towards the sink. This is pictorially explained in Fig. 4.

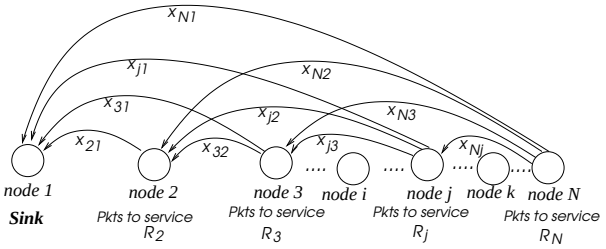


Figure 3. The linear network showing proportion of packets transmitted to different nodes towards the sink

The total number of packets the j^{th} node has to service is, say, R_j. R_j is the sum of packets it generates (i.e., m), plus all the packets it receives from nodes behind it. Thus, node-k has to service R_k packets. Out of that, it sends x_{kj} fraction directly to node-j, where $j < k$. We can write

$$R_j = m + \sum_{k=j+1}^{N} R_k \, x_{kj} \quad (1)$$

$$= m \times \left[1 + \{x_{(j+1)j}\} + \{x_{(j+2)(j+1)} \, x_{(j+1)j} + x_{(j+2)j}\} + \{x_{(j+3)(j+2)} \, x_{(j+2)(j+1)} \, x_{(j+1)j} + x_{(j+3)(j+2)} \, x_{(j+2)j} + x_{(j+3)(j+1)} \, x_{(j+1)j} + x_{(j+3)j}\} \ldots\right] \quad (2)$$

The cost of forwarding all those packets towards the sink is now considered. Say, C_j is the cost of transmitting packets at sensor node-j. Let us suppose that the power required to transmit a packet to a distance d is d^ν, where ν is somewhere between 2 and 3. We can write,

$$C_j = R_j \times \left[\sum_{i=j-1}^{1} x_{ji} \, \{(j-i) \times d\}^\nu\right] \quad (3)$$

The problem is to find all x_{ji}s, where $i < j$. The optimization criterion is to minimize C_js. The constraints are as follows:

$$C_2 = C_3 = \ldots = C_i \ldots = C_N \quad (4)$$

$$0 \leq x_{ji} \leq 1 \quad \text{for all} \quad i < j,\ i \geq 1,\ j > 1 \tag{5}$$

$$\sum_{i=(j-1)}^{1} x_{ji} = 1 \tag{6}$$

Eq. 4 says that the battery power of all nodes should drain equally. Eq. 5 says that the fraction of the packets forwarded to different nodes should lie between 0 to 1. Eq. 6 says that every node needs to transmit all the packets it is needed to service.

A general analytical solution for x_{ji} is not possible. We solved this problem for different values of N, namely for $N = 2, 3, 4, \ldots$. We will further show that there is no feasible solution for $N > 4$.

III. Solving the optimization Problem

We will solve specific problems, when cluster radius is 2, 3 and 4. We also will assume $\nu = 2$, which, though not always true, is accepted in general. For any other value of ν, the method will be the same, though the results would differ. Without any loss of generality, we assume the unit of cost function as $m \times d^2 = 1$ unit. This is only a multiplication factor and is done for making the equations look less cumbersome.

The case for $N = 2$ is trivial as all the packets are to be sent to the sink. Here, the unique solution is $x_{21} = 1$.

A. Case for N=3

Here, in addition to the sink node (i.e., node-1), we have two nodes, node-2 and node-3. Let us first see the total number of packets node-2 and node-3 have to service.

$$R_3 = m \tag{7}$$

$$R_2 = R_3.x_{32} + m = m \times (x_{32} + 1) \tag{8}$$

The optimization function, i.e., minimizing the cost of transmission for each node is expressed as:

$$\begin{aligned} C_3 &= R_3 \times (x_{32}\,(d)^{\nu} + x_{31}\,(2d)^{\nu}) = m\,d^2 \times (x_{32} + 4\,x_{31}) \\ C_2 &= R_2 \times \{x_{21}\,(d^{\nu})\} = m\,d^2 \times x_{21}\,(x_{32} + 1) \\ &= m\,d^2 \times (x_{32}\,x_{21} + x_{21}) \end{aligned}$$

As already mentioned, we set $m\,d^2 = 1$. As, $x_{21} = 1$, we can further simplify the above two expressions as $C_3 = x_{32} + 4\,x_{31}$ and $C_2 = x_{32} + 1$.

The constraints are:

$$C_2 = C_3$$

$$x_{32} + x_{31} = 1$$

$$0 \leq x_{32} \leq 1;\ 0 \leq x_{31} \leq 1$$

Equating C_2 and C_3, we find that $4 \times x_{31} = 1$. So, the solution is unique, namely $x_{32} = 0.75,\ x_{31} = 0.25,\ \text{and}\ x_{21} = 1$.

B. Case for N=4

Here, in addition to the sink node (i.e., node-1), we have three nodes, node-2, node-3 and node-4, as shown in Fig. 5. Let us first see the total number of packets every node has to service.

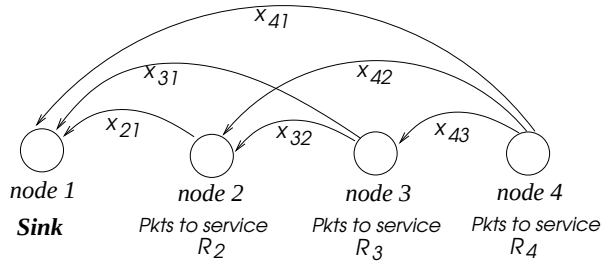


Figure 4. The transmission pattern of the network when N=4

Packets to service at node-4, node-3 and node2 are:

$$R_4 = m \tag{9}$$

$$R_3 = R_4.x_{43} + m = m\,(x_{43} + 1) \tag{10}$$

$$\begin{aligned} R_2 &= R_4.x_{42} + R_3.x_{32} + m \\ &= m.x_{42} + m.(x_{43} + 1).x_{32} + m \\ &= m\,[x_{43}\,x_{32} + x_{42} + x_{32} + 1] \end{aligned} \tag{11}$$

The cost of transmission i.e., C_4, C_3 and C_2, for forwarding packets to nodes towards the sink are,

$$\begin{aligned} C_4 &= R_4 \times \{x_{43}\,(d)^{\nu} + x_{42}\,(2d)^{\nu} + x_{41}\,(3d)^{\nu}\} \\ &= m\,d^2 \times (x_{43} + 4\,x_{42} + 9\,x_{41}) \end{aligned} \tag{12}$$

$$\begin{aligned} C_3 &= R_3 \times \{x_{32}\,(d)^{\nu} + x_{31}\,(2d)^{\nu}\} \\ &= m \times \{x_{43} + 1)(x_{32}\,(d)^{\nu} + x_{31}\,(2d)^{\nu}\} \\ &= m\,d^2 \times (x_{43}\,x_{32} + 4\,x_{43}\,x_{31} + x_{32} + 4\,x_{31}) \end{aligned} \tag{13}$$

$$\begin{aligned} C_2 &= R_2 \times \{x_{21}\,(d)^{\nu}\} \\ &= m\,d^2\,(x_{43}\,x_{32} + x_{42} + x_{32} + 1) \end{aligned} \tag{14}$$

We need to minimize C_2, C_3 and C_4. Here, we assume $\nu = 2$, and used the fact that $x_{21} = 1$. The cost equations, Eq. 13 and Eq. 14, are non-linear. To convert this non-linear problem to linear optimization problem, we replace the variable x_{43} with α, say a constant. As already mentioned, we also set $m\,d^2 = 1$. By that, Eq. 12 to Eq. 14 are modified to

$$C_4 = 4\,x_{42} + 9\,x_{41} + \alpha \tag{15}$$

$$C_3 = (1 + \alpha)\,x_{32} + 4\,(1 + \alpha)\,x_{31} \tag{16}$$

$$C_2 = x_{42} + (1 + \alpha)\,x_{32} + 1 \tag{17}$$

As our motivation is to schedule the transmission so that all the nodes expend equal amount of energy, we made $C_4 =$

$C_3 = C_2$. Combining Eq. 15 and Eq. 16, we get Eq. 18 and combining Eq. 15 and Eq. 17, we get Eq. 19, as follows.

$$4\,x_{42} + 9\,x_{41} - (1+\alpha)\,x_{32} - 4\,(1+\alpha)\,x_{31} = -\alpha \quad (18)$$
$$3\,x_{42} + 9\,x_{41} - (1+\alpha)\,x_{32} - 0\,(1+\alpha)\,x_{31} = 1-\alpha \quad (19)$$

From the constraint $\sum_{i=(j-1)}^{1} x_{ji} = 1$, we get

$$x_{43} + x_{42} + x_{41} = 1$$
$$x_{32} + x_{31} = 1$$

We rewrite the above, replacing x_{43} by α, as

$$1.\,x_{42} + 1.\,x_{41} + 0.\,x_{32} + 0.\,x_{31} = (1-\alpha) \quad (20)$$
$$0.\,x_{42} + 0.\,x_{41} + 1.\,x_{32} + 1.\,x_{31} = 1 \quad (21)$$

From the above we formulate our linear programming problem as follows. We need to find

$$X = [x_{42} \;\; x_{41} \;\; x_{32} \;\; x_{31}]$$

that minimizes the transmission cost (Eq. 14)

$$f = 0.\,x_{42} + 0.\,x_{41} + (1+\alpha)\,x_{32} + 4(1+\alpha).\,x_{31}$$

subject to equality constraint (Copy of Eq. (20), (21), (18) and (19))

$$\begin{aligned} 1.\,x_{42} + 1.\,x_{41} + 0.\,x_{32} + 0.\,x_{31} &= (1-\alpha) \\ 0.\,x_{42} + 0.\,x_{41} + 1.\,x_{32} + 1.\,x_{31} &= 1 \\ 4.\,x_{42} + 9.\,x_{41} - (1+\alpha)\,x_{32} - 4\,(1+\alpha)\,x_{31} &= -\alpha \\ 3.\,x_{42} + 9\,x_{41} - (1+\alpha)\,x_{32} - 0\,(1+\alpha)\,x_{31} &= 1-\alpha \end{aligned}$$

The equality constraint can be written as

$$A_{eq}.\,X = b_{eq}$$

where,

$$A_{eq} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 4 & 9 & -(1+\alpha) & -4(1+\alpha) \\ 3 & 9 & -(1+\alpha) & 0 \end{bmatrix}$$

and

$$b_{eq} = \begin{bmatrix} 1-\alpha \\ 1 \\ -\alpha \\ 1-\alpha \end{bmatrix}$$

In addition, we have upper bound and lower bound constraints for X-variables as follows:

$$0 \le x_{42} \le 1; \quad 0 \le x_{41} \le 1; \quad 0 \le x_{32} \le 1; \quad 0 \le x_{31} \le 1$$

We have assumed $X_{43} = \alpha$, which is again bound between 0 to 1. We changed the value of α from 0 to 1 in steps of 0.05 and solved the above linear programming problem.

Solutions of this linear programming problem exists when $0.47 \le x_{43} \le 0.80$. Corresponding to every value of x_{43}, we had unique solutions for $x_{42},\ x_{42},\ x_{32},\ x_{31}$. The whole set of solutions is shown in the following table. We presented only the important part, deleting rows of essentially similar results.

Table I
COST FUNCTION FOR DIFFERENT VALUES OF x_{ji}S

Flag	Cost	x_{43}	x_{42}	x_{41}	x_{32}	x_{31}
-2	2.6038	0.450	0.5385	negative	0.7347	0.2653
-2	2.6138	0.460	0.5385	negative	0.7366	0.2634
1	2.6139	0.470	0.5252	0.0048	0.7406	0.2594
1	2.6122	0.480	0.5096	0.0104	0.7450	0.2550
1	2.6104	0.490	0.4939	0.0161	0.7493	0.2507
1	2.6087	0.500	0.4783	0.0217	0.7536	0.2464
1	.	.	.	.	.	.
1	.	.	.	.	.	.
1	.	.	.	.	.	.
1	2.5583	0.790	0.0243	0.1857	0.8569	0.1431
1	2.5565	0.800	0.0087	0.1913	0.8599	0.1401
-2	2.5600	0.810	negative	0.1944	0.8619	0.1381
-2	2.5700	0.820	negative	0.1944	0.8626	0.1374

In Table. I, the first column, Flag, denotes whether there is a feasible solution or not. Here, '-2' indicates that the solution is not feasible, and '1' indicates that the solution is feasible. In row 1 and 2, x_{41} s' values are negative, and therefore are not feasible solutions. Similarly, for last two rows, x_{42} values are negative, making them infeasible solutions. The cost, for different values of x_{43} (=α), and corresponding values of other fractions, namely $x_{42},\ x_{41},\ x_{32},\ x_{31}$, are shown. Entries for α from 0.500 to 0.790 are omitted. As the cost function over the whole range of feasible solutions is minimum when $x_{43} = 0.8$, we can write our final solution as,

$$x_{43} = 0.800; \quad x_{42} = 0.009; \quad x_{41} = 0.191;$$
$$x_{32} = 0.860; \quad x_{31} = 0.140$$

Thus, to minimize the transmission cost and to share the load of transmitting packets so that power at all the nodes are equally drained:
Node-4 needs to transmit 80% of its packets to node-3, 1% to node-2, and 19% directly to the sink.
Node-3 needs to transmit 86% of its packets to node-2, 14% directly to the sink

C. Case of $N \ge 5$

In case the number of nodes is 5, to transform the transmission cost function to a linear one in terms of different x_{ji}s, we need to get rid of the three terms $x_{54},\ x_{53},\ x_{43}$ from transmission cost functions. As before, we assigned

them values from 0 to 1, and changed in steps of 0.05. This time we could not get any feasible solution, with all fractions being positive. Thus, there is no solution for $N = 5$. The detail procedure, though not shown here, is exactly the same as in case of $N = 4$.

In case of $N = 5$ there is no feasible solution because, to maintain same transmission cost for all nodes some of the x_{ji} terms has to be negative. Intuitively, similar situation will arise for $N \geq 5$. We thus conclude that there is no solution for $N \geq 5$.

IV. Case when nodes are spread over a plane

In the previous section, we have shown that for a cluster, where sensor nodes are linearly spread, it is possible to transmit different portions of packets to different distances judiciously, so that all nodes dissipate power uniformly. In reality, the sensor nodes are spread over a plane. In this section, we will show that the analysis of the previous section can be extended nodes distributed on a plane.

Let us consider that the sensor nodes are uniformly spread as shown in Fig. IV.

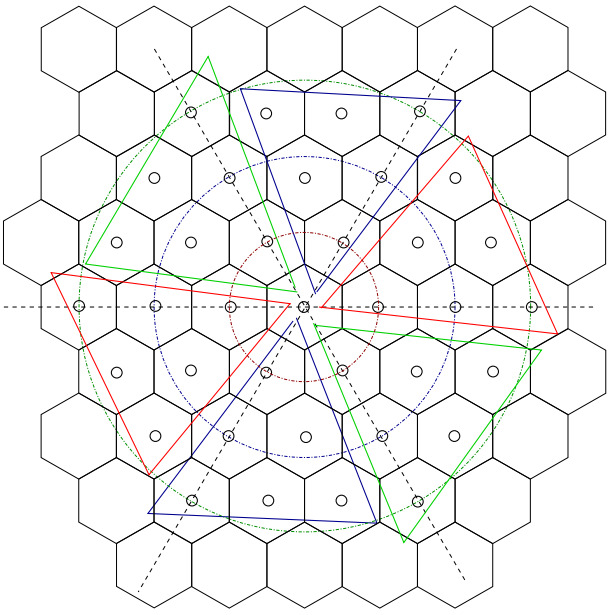


Figure 5. Sensor nodes spread uniformly over a plane

Nodes are put at the center of the imaginary equilateral hexagons. The node density is a function of the hexagon edge length, say a. For uniform distribution with any node density, we can represent the network as in Fig. IV, where with increasing node density the value of a is smaller. Here, CH is at the center. The whole cluster is divided into 6 triangular sections, which together forms the cluster, as shown in Fig. IV.

As in Section III, here too x_{ji} denotes the portion of packets node-j transmits directly to node-i. The ratios are

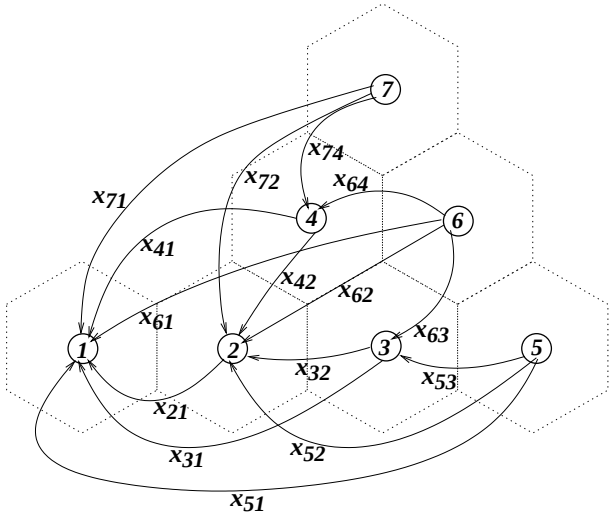


Figure 6. A triangular section showing one-sixth of the cluster

as shown in the following matrix. The columns are receiving node numbers, starting with 1, and the rows are transmitting node numbers, starting with node 2.

$$\mathbf{X} = \begin{bmatrix} x_{21} & 0 & 0 & 0 & 0 & 0 & 0 \\ x_{31} & x_{32} & 0 & 0 & 0 & 0 & 0 \\ x_{41} & x_{42} & 0 & 0 & 0 & 0 & 0 \\ x_{51} & x_{52} & x_{53} & 0 & 0 & 0 & 0 \\ x_{61} & x_{62} & x_{63} & x_{64} & 0 & 0 & 0 \\ x_{71} & x_{72} & 0 & x_{74} & 0 & 0 & 0 \end{bmatrix} \begin{matrix} node\ 2 \\ node\ 3 \\ node\ 4 \\ node\ 5 \\ node\ 6 \\ node\ 7 \end{matrix}$$

Here, $x_{ji} = \frac{packets\ transmitted\ from\ node{-}j\ to\ node{-}i}{total\ packets\ transmitted\ from\ node{-}j}$. Corresponding distances of transmission are shown in the matrix **D** below, where entry d_{ji} is the distance from node-j to node-i. "NA" means the distance is not necessary to consider, because there is no transmission.

$$\mathbf{D} = \begin{bmatrix} \sqrt{3}a & NA & NA & NA & NA & NA & NA \\ 2\sqrt{3}a & \sqrt{3}a & NA & NA & NA & NA & NA \\ 3a & \sqrt{3}a & NA & NA & NA & NA & NA \\ 3\sqrt{3}a & 2\sqrt{3}a & \sqrt{3}a & NA & NA & NA & NA \\ \sqrt{21}a & 3a & \sqrt{3}a & \sqrt{3}a & NA & NA & NA \\ \sqrt{21}a & 2\sqrt{3}a & NA & \sqrt{3}a & NA & NA & NA \end{bmatrix}$$

As before, let us denote the total number of packets serviced (transmitted) by node-j by R_j, which includes both the packets generated at the node (i.e., m) plus those received from nodes further away from the the CH. Therefore,

$$R_7 = R_6 = R_5 = m \tag{22}$$

$$R_4 = m + \sum_{i=5,6,7} R_i \times x_{i4} \tag{23}$$

$$R_3 = m + \sum_{i=5,6,7} R_i \times x_{i3} \tag{24}$$

$$R_2 = m + \sum_{i=3,4} R_i \times x_{i2} + \sum_{i=5,6,7} R_i \times x_{i2} \tag{25}$$

 ISBN: 978-1-61208-100-7

From distances in matrix **D**, we are able to calculate the amount of power the node dissipates for transmitting these packets. At this stage, we ignore (as is done in Section III) the energy required to receive packets, which is lower compared to what is required for transmission, though not zero. As before, we denote the power dissipated at node-j by C_j and make all C_js equal. Formaly, the problem is to find the elements of matrix **X**, where the optimization criterion is to minimize C_j subject to the following constraints:

$$\sum \vec{x_j} = 1 \tag{26}$$

$$C_2 = C_3 = \ldots = C_i \ldots = C_7 \tag{27}$$

$$0 \leq x_{ji} \leq 1 \quad \text{for all} \quad i < j,\ i \geq 1,\ j > 1 \tag{28}$$

Eq. 26 can further be expanded to get 6 simultaneous equations in x_{ij}. From Eq. 27, we get another 6 equations of expanded energy in terms of x_{ij}. By equating them, we get 5 more equations in x_{ij}s. From symmetry, we can assume that $x_{74} = x_{53}$ and $x_{64} = x_{63}$. Fixing values of two unknowns, x_{74} and x_{64}, in small steps, and using optimization criterion to minimize C_is, we can solve x_{ij}s.

the following set of equations:

$$x_{74} + x_{72} + x_{71} = \alpha + x_{72} + x_{71} = 1 \tag{29}$$

$$x_{64} + x_{63} + x_{62} + x_{61} = 2\,\beta + x_{62} + x_{61} = 1 \tag{30}$$

$$x_{53} + x_{52} + x_{51} = \alpha + x_{52} + x_{51} = 1 \tag{31}$$

$$x_{42} + x_{41} = 1 \tag{32}$$

$$x_{32} + x_{31} = 1 \tag{33}$$

$$x_{21} = 1 \tag{34}$$

where, we wrote $x_{74} = x_{53} = \alpha$, assuming $x_{74} = x_{53}$ due to symmetry, and α, a constant which we will vary manually in steps to find its optimum value. Similarly, we wrote $x_{64} = x_{63} = \beta$, assuming $x_{64} = x_{63}$ due to symmetry, and β, another constant which we will vary in steps. We now have 10 unknown x_{ij}s, and 5 equations, Eq. 29 to Eq. 33.

V. CONCLUSIONS

Cluster-based communication protocol in wireless sensor networks had to reconfigure at regular intervals for uniform power dissipation of different nodes. But, dismantling operating clusters and re-arranging a new set of clusters needs transmission of lots of signaling packets To avoid reconfiguration of clusters at regular intervals, we proposed that, the nodes within the cluster do not always transmit to its nearest neighbor, on the way to CH. Instead, they transmit packets to different distances towards the CH, with different pre-assigned ratios. We have shown that, we can then achieve uniform power dissipation for all the nodes within the cluster.

ACKNOWLEDGMENT

Part of this research is supported by Government of Japan, Ministry of Education Scientific Research General Research Fund (C) (2) No. 20500071.

REFERENCES

[1] T. Meng and R. Volkan, "Distributed network protocols for wireless communications." In *the Proceedings of IEEE ISCAS, May 1998.*

[2] P. Levis, S. Madden, et.al.,, "Tiny OS: An operating system for for Wireless Micro-sensor Networks." In *Ambient Intelligence (New York, 2004), Springer-Verlag.*

[3] S. R. Madden, M. J. Franklin, J. M. Hellerstein, W. Hong,"Tiny DB: An acquisitional Query Processing System for sensor networks." *ACM transactions on Database Systems, Vol. 30, Issue. 1, pp. 122-173, March 2005.*

[4] C. Intanagonwiwat, R. Govindan, and D. Estrin, "Directed diffusion: A scalable robust communication paradigm for sensor networks." In *the Proceedings of the Sixth Annual International Conference on Mobile computing and networks (MobiCOM 2000), Boston, Massachusetts, August 2000.*

[5] D. Braginsky, D. Estrin, "Rumor Routing algorithm for sensor networks." In *the Proceedings of Wireless Sensor Network Algorithms (WSNA'02), pp. 22-30, Atlanta, Georgia, September 28, 2002.*

[6] W. Ye, and J. Heidemann, "Ultra-low duty cycle MAC with scheduled channel polling." In *the Proceedings of fourth International conferences on embedded networked sensor systems (SenSys'06), pp. 321-334, 2006.*

[7] Bhaskar Krishnamachari, "Networking Wireless Sensors.", *Cambridge University Press, January 2006.*

[8] W. Heinzelman, A. Chandrakasan, and H. Balakrishnan, "An Application-Specific Protocol Architecture for Wireless Microsensor Networks," *IEEE Transactions on Wireless Communications, Vol. 1, No. 4, pp. 660-670, October 2002.*

[9] A. Cerpa, and D. Estrin, "ASCENT: Adaptive Self-Configuring sEnsor Networks Topologies.", *IEEE transactions on mobile computing, Vol 3, 3, pp. 1-14, 2004.*

[10] W. R. Heinzelman, A. Chandrakasan, and H. Balakrishnan, "Energy-Efficient Communication Protocol for Wireless Microsensor Networks." *in the Proceedings of the Hawaii International Conference on System Sciences, January 4-7, 2000.*

[11] Alex Rogers, Esther David, and Nicholas R. Jennings, "Self-Organized Routing for Wireless Microsensor Networks." *IEEE trans. on SMC - Part A, Vol. 35, No. 3, May 2005*

[12] S. Cui, A. J. Goldsmith, and A. Bahai, "Energy efficiency of MIMO and cooperative MIMO techniques in sensor networks." *IEEE Journal on Selected areas in Communication, pp 1089-1098, 2004.*

 ISBN: 978-1-61208-100-7

Cooperative Communication to Improve Reliability and Efficient Neighborhood Wakeup in Wireless Sensor Networks

Rana Azeem M. Khan and Holger Karl,
University of Paderborn, Paderborn, Germany
{azeem@mail.upb.de, holger.karl@upb.de}

***Abstract*—To maximize lifetime of Wireless Sensor Networks, medium access control protocols usually trade off reliability for energy efficiency. Channel errors, collisions, idle listening, and overhearing further aggravate the problem. Our work investigates opportunities to improve reliability in Wireless Sensor Networks under such constraints. We consider a multi-hop data gathering network in which sensor nodes are deployed around a sink. Nodes periodically sense data and forward it to next hop nodes. For such a network, a Medium Access Control protocol, called CPS-MAC, is proposed. This protocol uses cooperative communication to improve reliability by using overhearing to its advantage. In conventional protocols, overhearing causes nodes to receive packets which are not meant for them. Therefore, these packets are discarded and considered a waste of energy. On the contrary, CPS-MAC intentionally wakes up next 1-hop and 2-hop neighbors to improve their chances of overhearing a packet. The overheard packets are buffered and then relayed to the next hop neighbor, combating channel fading by a cooperative spatial diversity gain. By combining multiple copies of the same packet, next hop neighbor is more likely to recover the original packet. Design challenges such as efficiently waking up neighborhood nodes, minimizing energy overhead, and partner selection are addressed. Simulation results show that CPS-MAC significantly decreases packet error rate without expending additional energy.**

***Keywords*—Wireless Sensor Networks; Media Access Control; Cooperative Communication; Reliability.**

I. INTRODUCTION

Wireless sensor networks (WSN) are used in a wide range of applications, such as target tracking, habitat sensing and fire detection. WSN are particularly useful in situation where an infrastructure network is not present or not feasible. In such conditions, sensor nodes can be deployed around a sink to create a multi-hop data gathering network as shown in Figure 1. The nodes coordinate locally to forward each other packets. The packets travels in a hop-by-hop fashion towards the sink.

As sensor nodes are battery powered, they operate under strict energy constraints. Common WSN protocols such as S-Mac, T-MAC and CSMA-MPS trade off performance for energy efficiency [18], [22]. The nodes use low transmission powers and switch the transceiver between sleep and awake states. Fading and the broadcast nature of the wireless channel results in channel errors, collisions, and overhearing due to which these networks drop a significant proportion of packets.

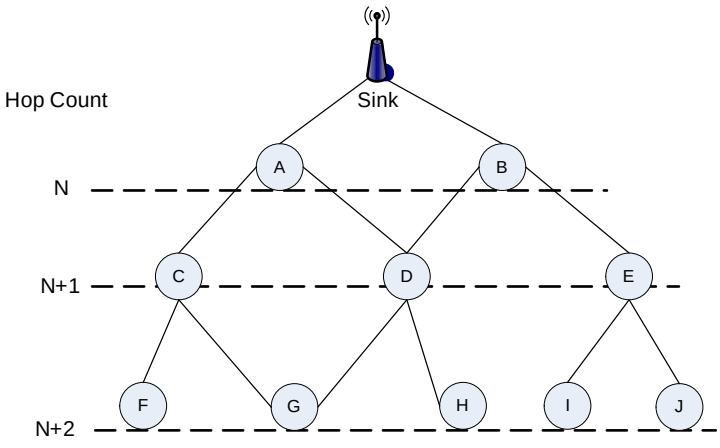


Fig. 1. Data gathering network

Signal fading can be the most severe among these impairments. In a wireless channel, random scattering from reflectors with different attenuation coefficients results in multiple copies of a transmitted signal arriving (and interfering) at a receiver with different gains, phase shifts, and delays. These multiple signal replicas can add together in a constructive or destructive way, amplifying or attenuating the received signals amplitude. Destructive interference results in fading, which causes temporary failure of communication, as the amplitude of the received signal may be low to the extent that the receiver may not be able to distinguish it from thermal noise.

Under such conditions, ensuring reliable communication while conserving energy is a challenging problem. This has motivated us to design Cooperative Preamble Sampling Medium Access Control (CPS-MAC) protocol which can improve reliability without expending additional energy. Our protocol takes advantage of overhearing. Overhearing means that a node will receive all messages in its reception range including those that are intended for other nodes. Considered problematic, specially in dense WSN, these packets are usually discarded and this wastes energy.

We suggest using cooperative communication (CC) [3] to take advantage of these overhead packets. In CC, nodes cooperate to improve the overall performance of the network.

 ISBN: 978-1-61208-100-7

Since a transmission in the wireless channel is overheard by neighboring nodes, these nodes can process the overheard packets and re-transmit them [4]. Figure 2 elaborates a 3-node CC scenario. We refer to this as a cooperative triangle, which consists of a source, partner, and destination node. Destination node here refers to the next-hop node in the cooperative triangle and is used in the same context throughout this paper.

We exemplify a possible realization of a cooperative communication scheme as follows (for alternatives, see [8], [12]–[14]). The source broadcasts a message to the destination in a first phase. Due to the broadcast nature of the wireless channel, the partner station can overhear the source transmission, decode it, and if received correctly, forwards it to the destination in a second phase. We refer to this two phase scheme as one transmission cycle. By combining different copies of the same transmission by source and partner stations, the destination can improve its ability to decode the original packet and exploit spatial diversity and robustness against channel variations due to fading.

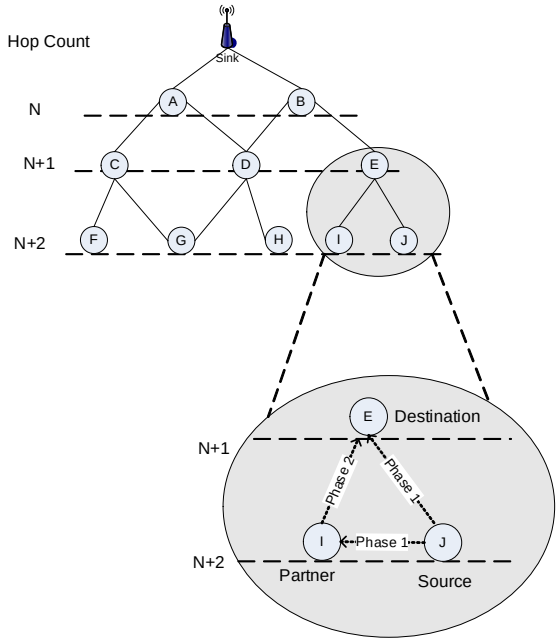


Fig. 2. Cooperative Communication

We propose to realize this concept at the Medium Access Control (MAC) layer, which is responsible for radio usage and scheduling transmission efficiently. Although CC has already been investigated at the MAC layer for traditional wireless networks such as wireless LANs (WLAN) based on the IEEE802.11 standard [8], [9], [13]–[15], integrating CC into MAC layer for WSN has received little attention. It is important to mention here that MAC protocols for WSN differ significantly from MAC protocols for WLAN. In WLAN optimization of performance parameters such as throughput, latency, and fairness is a primary concern. In WSN energy conservation and extending lifetime is essential. Details can be found in Section II.

We briefly outline the challenges faced in developing CC based MAC protocols for WSN, along with solutions proposed in CPS-MAC.

1) MAC protocols such as X-MAC try to conserve energy by maximizing the sleep duration of the nodes [17]. CC on the other hand increases energy expenditure by requiring nodes to be awake more often. In such a situation, improving reliability and conserving energy may seem counter intuitive. CPS-MAC compensates for the additional energy expenditure by reducing the time needed to wake up neighboring nodes and by achieving lower packet error rates.
2) Application of CC in densely deployed WSN can result in multiple nodes overhearing and forwarding a packet and flooding part of the network. In such situations, it could be practical to limit the number of nodes taking part in CC and avoid redundant transmission and energy wastage. For this CPS-MAC includes an addressing scheme which allows source node to select partner and destination prior to transmission. For this, CPS-MAC includes an addressing scheme which attempts to limit one transmission cycle to three nodes and minimizing the number of nodes unnecessarily overhearing the transmission.
3) Under ordinary conditions data would travel in a hop-by-hop fashion during each transmission. Narayanan et al. [10] and Zhu et al. [11] have shown that two-hop forwarding leads to higher total network throughput. Therefore, CPS-MAC attempts to deliver a packet over multiple hops in a single transmission cycle as shown in Figure 3. Notice here that Figure 3 differs from Figure 2. This multi-hop transfer in a single transmission cycle consumes less energy then several single-hop transfers. The protocol uses hop count parameter for this purpose and is explained in section III in detail.

Details of CPS-MAC are presented in Section III.

II. BACKGROUND

A. Medium Access Control in Wireless Sensor Networks

MAC protocols in WSN conserve energy by duty cycling radio which is the main source of energy consumption. Several MAC protocols for WSN have been proposed in recent years, which optimize duty cycle depending upon underlying application requirement and traffic behavior [22]. They can be divided into two main categories namely schedule-based and contention-based.

The schedule-based approach requires nodes to synchronize at some common time of reference such that they can wake up collectively prior to transferring. This approach may seem attractive at first glance because idle-listening and overhearing simply do not occur. However, the need to synchronize sleeping schedules and the control packet overhead make them less feasible. Ideally, a MAC protocol in WSN does not impose a high overhead for exchanging control information. Otherwise, a significant amount of energy will be consumed for it.

 ISBN: 978-1-61208-100-7

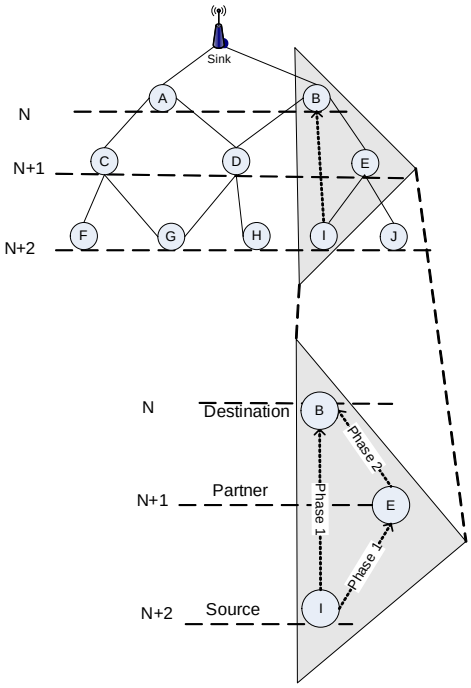


Fig. 3. Cooperation over multiple hops

Contention-based schemes on the other hand does not require synchronization of sleep schedules and are more flexible to handling variable traffic loads [1]. However in such schemes, nodes who wish to transmit must content for the channel and the winner transmits at the risk of collision. Accordingly, these protocols contain mechanism to avoid or to minimize the probability of collisions.

Preamble sampling is one such protocol which is specifically designed for WSN [16] and is particularly useful when the traffic generation is non-periodic. Figure 4 shows the working of the protocol. Nodes switch between sleep and listen (awake) states. When a sender has data to send, it wakes up the receiver by sending a preamble which is longer than the sleep duration of the receiver node. When a receiver node wakes up and switches its radio to listen state, it hears the preamble, uses it to synchronize with the source, and stays awake for incoming transmission. Then, the source initiates the transmission at the end of the preamble. After the transmission is complete, nodes resume duty cycling. As the cost (energy) of waking up is transfered from receiver to sender, and there are more receivers than senders, a lot of energy is saved.

To shorten the preamble length and further minimize energy consumption at both sender and receiver, an improvement to Preamble Sampling was proposed in [17] and [18]. This scheme is known as Minimum Preamble Sampling (MPS) and is shown in Figure 5. Here one long preamble is divided into a series of short preambles interleaved with listening intervals. We refer to these listening intervals as inter-preamble spacing. If a receiving node wakes up and hears the short preamble, it sends an acknowledgment (ACK) packet to the

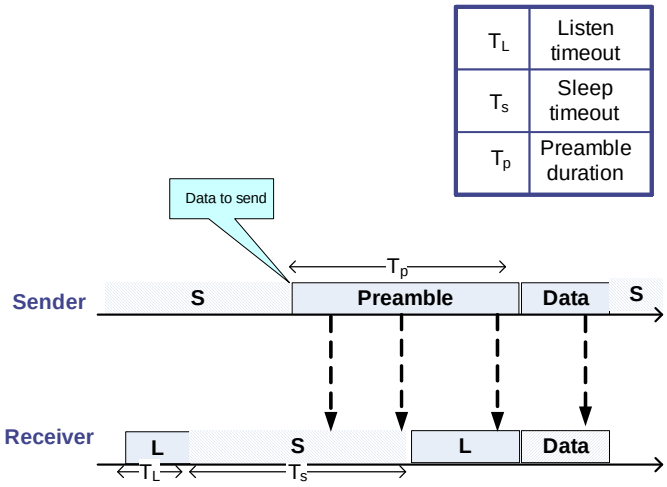


Fig. 4. Preamble Sampling

sender during the inter-preamble spacing. Upon receiving the ACK, the sender initiates the data transmissions.

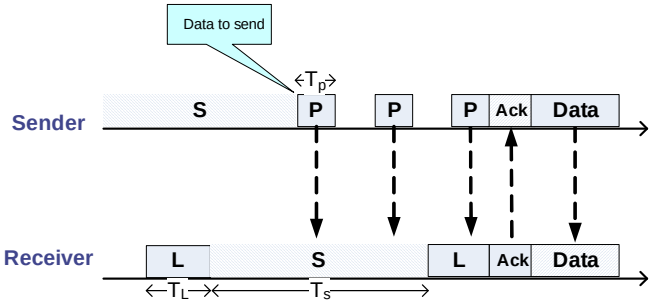


Fig. 5. Minimum Preamble Sampling

B. Medium Access Control Protocols for Cooperative Communication

A significant amount of work has been done on developing CC protocols in wireless networks to combat the effects of channel fading. The initial work focused on physical layer schemes [2], [6]. However, in order to realize the full potential of cooperative communication, it is imperative that the layer directly above the physical layer, namely the medium access control (MAC) layer, must be able to schedule transmissions effectively and efficiently. This has led researchers to investigate the support of cooperative communication in various forms at higher protocol layers including MAC layer [7]. A MAC protocol called CoopMAC illustrates how the legacy IEEE 802.11 distributed coordination function (DCF) [9] can be modified to use cooperative communication thus achieving both higher throughput and lower interference [8]. More cooperative communication protocols based on IEEE 802.11 were proposed in [12], [13], [14] and [15]. However, protocols based on IEEE 802.11 are not feasible in WSN as they have strict energy constraint and limited processing power.

Analyzing the effects of cooperation in legacy MAC protocols for WSN has received little attention. Mainaud et al.

 ISBN: 978-1-61208-100-7

[19] has recently proposed a cooperative MAC protocol for WSN based on preamble sampling. The primary focus of the work is to define a relay node among the neighboring nodes and relaying decision at the link. However, the work does not analyze the effect on energy consumption, a primary concern in WSN.

Motivated by the previous work, we have designed CPS-MAC.The difference between CPS-MAC and prior work is that CPS-MAC addresses a number of design challenges such as addressing scheme, energy efficient wake up, and a scheduling scheme which uses CC. These schemes are integrated together into a low-overhead practical MAC protocol.

III. Protocol Design for CPS-MAC

We consider an ad hoc multi-hop data gathering network where sensor nodes are deployed around a sink as shown in Figure 6. Each node defines its distance from the sink using hop count which is defined as the number of intermediate hops between the node and the sink [20]. The sensor nodes periodically sense the data, wake up the neighboring nodes, and broadcast the data. Neighboring nodes receive the data and the one which is closer to the sink forward it to the next-hop nodes. Data eventually reaches the sink which is responsible for collecting, processing, analyzing, and forwarding the data to a base station.

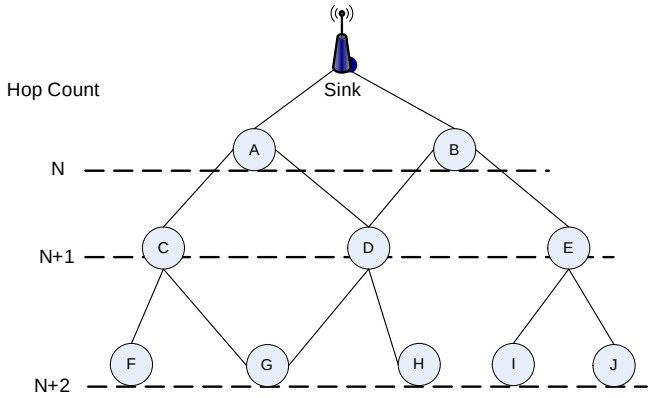


Fig. 6. Data Gathering Network

Once the sensor nodes are physically deployed around the sink, CPS-MAC works as follows.

A. Initialization Phase

In order to make routing decision and address nodes, CPS-MAC uses hop count value and neighborhood information. Hop count is the minimum number of non-cooperative transmissions required to reach the sink from a given node [20]. In order to setup this field, we use a flooding algorithm. An example of such an algorithm is the Cost Field Establishment Algorithm (CFEA) [21]. It is executed during the startup phase of the network and whenever the network topology changes. No CC is used during this phase. Consider the hierarchy shown in Fig 6. Initially, the sink sets its hop count to 0 and nodes set their hop count to ∞. The sink initiates the algorithm by broadcasting an advertisement (ADV) packet. The content of an ADV packet is shown in Figure 7 which would contain nodes hop count, its own addresses, and address of its 1-hop parent nodes. The address of 1-hop parents are needed for addressing and will be explained in the next section.

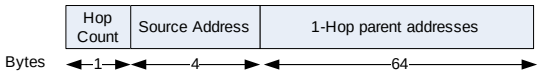


Fig. 7. Advertisement (ADV) Packet

The message propagates down from the parent node to the siblings. We use the term parent and sibling because nodes in the network are deployed in a hierarchy. Whenever a node receives an ADV message, it determines if it leads to a smaller hop count to the sink. If it does, the node resets its hop count and stores the source address as its 1-hop parent and the remaining addresses as 2-hop parent. Then, the node (re-)transmits its own ADV packet.

The 1-hop and 2-hop parent node addresses are stored in a routing table called CoopTable. It additionally stores the addresses of 1-hop sibling nodes. These addresses are obtained by simply overhearing ADV packets on the media and analyzing the hop count value. This is feasible because nodes do not sleep during the initialization phase and can receive all ADV packets in their reception range. Eventually, every node may calculate the optimal hop count to the sink through flooding. Then, the initialization phase stops and nodes start their normal operation; for example, the node D in the hierarchy above would have a CoopTable as follows:

TABLE I
Node D: CoopTable Parent Nodes

1 hop Parent (Hop Count-1)	2 Hop Parent (Hop Count-2)
A	Sink
B	Sink

TABLE II
Node D: CoopTable Sibling Nodes

1 Hop Sibling (Hop Count+1)
G
H

The following section explains how the CoopTable is used to address nodes and select partner nodes for cooperation.

B. Addressing Scheme

A broadcast transmission from a node to the sink over multiple hops can result in multiple nodes forwarding the same packet along different paths and flooding the network. Though it increases the chances of a packet eventually reaching the sink, nodes have to pay the price of energy expenditure and processing overhead. The problem becomes more complicated

 ISBN: 978-1-61208-100-7

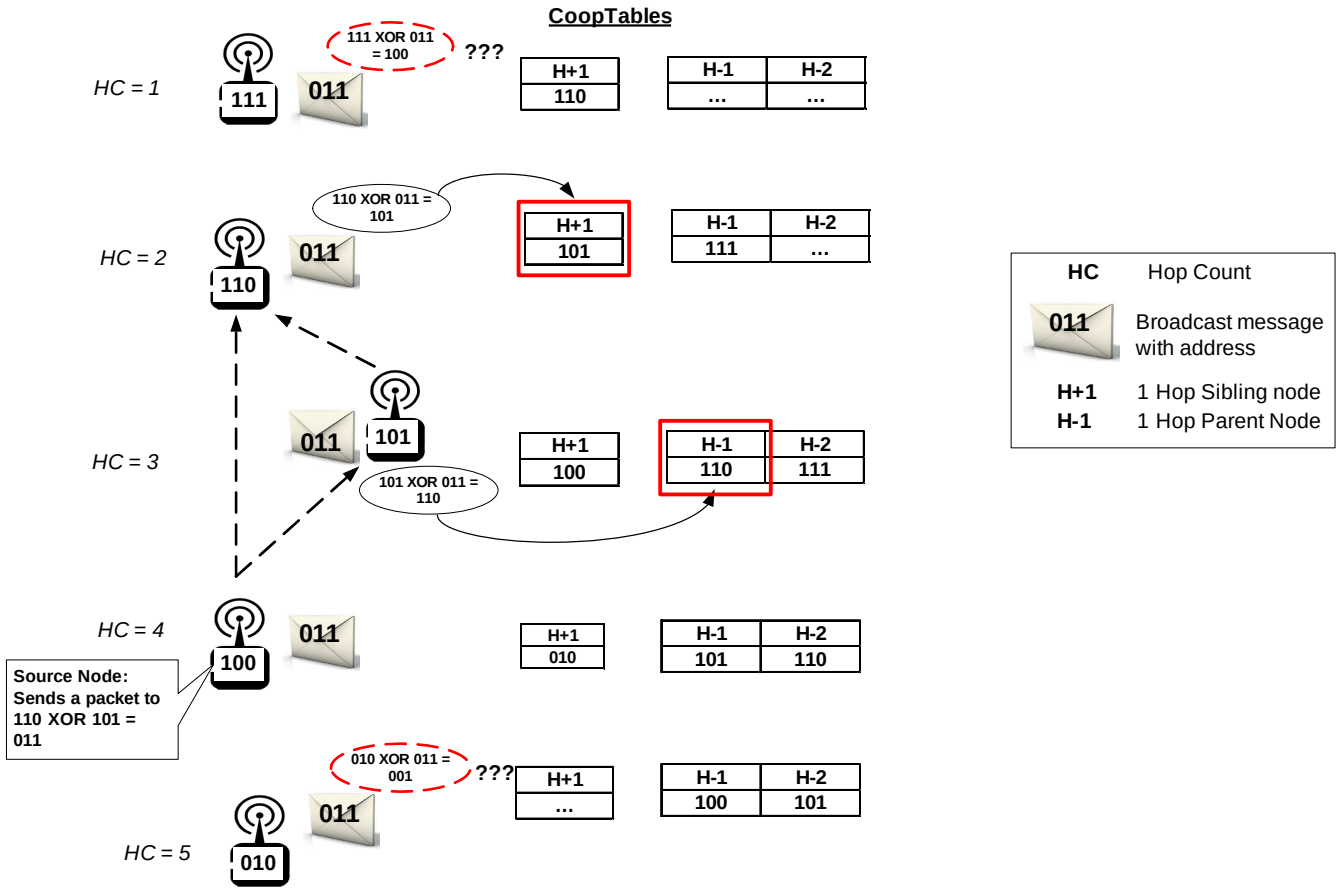


Fig. 8. Addressing Scheme

when we use cooperative communication because it involves a partner node in addition to the source-destination pair. In order to minimize this overhead and limit the cooperative communication to 3 nodes (source-partner-destination) in each transmission cycle, we use the CoopTable mentioned in Section III-A. When a node has data to send, it will select both partner (1-hop parent) and destination (2-hop parent) addresses from the CoopTable. However, instead of adding them as two separate addresses, the node will perform an XOR between them and send it as a single address. Recall from the previous section that every node stores addressing information about its 1-hop and subsequent 2-hop parents and 1-hop siblings in the CoopTable. If multiple partner\destination pairs are possible, the source cycles between them to divide the overhead. Nodes also include their hop count value in the packet. Once the packet is sent, every node that receives it extracts the address, performs an XOR with its own address, and looks up the result in its CoopTable. Nodes also calculate the hop count difference with the source node and then use the following rules to determine its role (partner /destination) in transmission.

1) If the result matches the address of a sibling node and the hop count difference with the source node is 2, the node acts as destination.
2) If the result matches the address of a parent node and the hop count difference is 1 with the source node, the node acts as partner.
3) If either the result does not matches an entry in the lookup table or if the hop count difference is greater than 2, the node takes no action.

For example, in Fig 8, the node with Identifier (ID) 100 sends a packet to node 101 and 110. The XOR of their address is 011, which is included in the data packet. Assuming that all nodes in the neighborhood correctly receive the packet, they decode the address using XOR with their own address. The lookup in the CoopTable for the node 110 and 101 matches the above mentioned rules and they define their roles as destination and partner respectively. The node 111 and 010 are not able to find the resulting address in the CoopTable and therefore do not take part in Cooperation. In this scheme, there is a probability that the result from the XOR operation might result in collision, i.e., the resulting address can map to a value in the CoopTable even though the node was not addressed, especially when the number of bits used for node identifiers is small. However, the probability significantly reduces when the identifier is large (e.g. 48, bit MAC address).

 ISBN: 978-1-61208-100-7

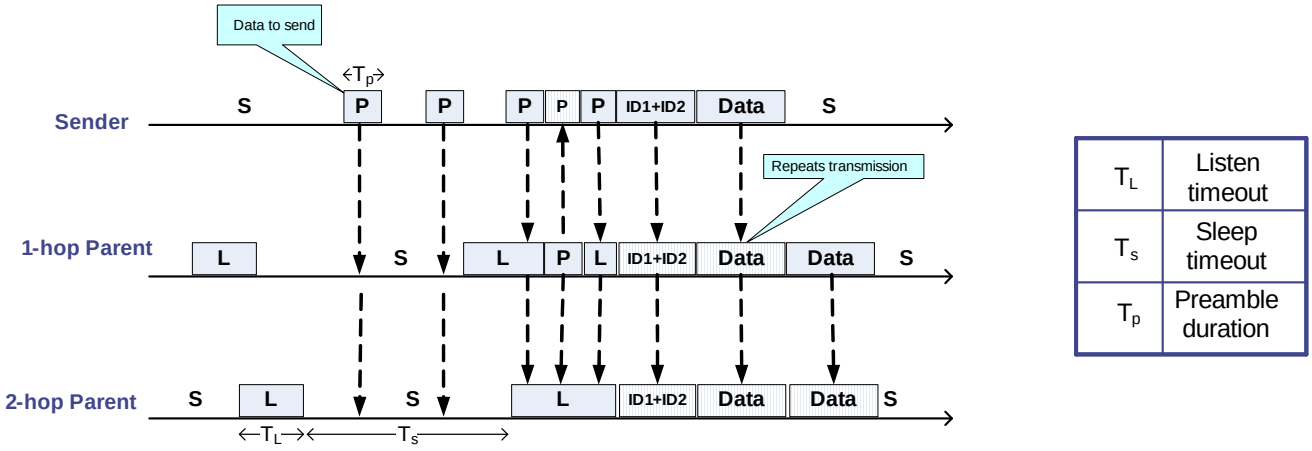


Fig. 9. CPS-MAC

C. Medium Access Control Layer

We propose a MAC protocol that uses cooperative communication to increase the probability of correct transmission while reducing energy consumption. Usually a broadcast transmission can be received by nodes which are multiple hops away from the source but they are discarded as they suffer from bit errors due to fading and attenuation. Our motivation is to utilize even these corrupt packets. The idea is to form cooperative triangles in the network where each triangle consists of source, partner, and destination as shown in Figure 3. Nodes cooperate in this triangle to deliver multiple copies of the packet to the destination where packet combining [4] is used to recover the original packet. However, for such a scheme to work, it becomes challenging to wake up nodes which are multiple hops away before initiating a data transmission. To solve this, we propose a wake up scheme which is based on minimum preamble sampling explained in Section II-A [16].

Figure 9 elaborates the working of the protocol. When a source node has data to send, it transmits a strobed preamble packet containing synchronization bits and the node's hop count value at the end. The strobed preamble is repeated until the source receives an acknowledgment (ACK) preamble from a neighboring node. When a neighboring node wakes up and receives the preamble, it analyzes the hop count value. If the receiver is not a parent node, it discards the preamble and immediately returns to sleep state as it cannot help the source to forward its data to the sink. 1-hop parent nodes that receive the preamble contend for the media and the successful node sends an ACK preamble. As no addressing is used in preamble, any 1-hop parent node can send the ACK preamble. This ACK preamble serves two purposes. First, it will act as wakeup preamble sequence for the next-hop parent. Second, the source will know that nodes in 1-hop neighborhood are awake. After receiving the acknowledgment preamble, the source sends the address packet. Nodes analyze the address packet as explained in Section III-B. If a node cannot define its role, it will return to sleep state to conserve energy. After this, the source broadcasts the data packet. The transmission is heard by both partner and destination yet it is unlikely to be received correctly by both nodes at once. After receiving the packet, the partner uses decode and forward (DAF) [4] to decide if it should again broadcast the packet. In DAF, the partner decodes a received packet to check for bit errors and erroneous packets are discarded. Only if the packet is received correctly, the partner again broadcasts the received packet to the destination. Thus, the destination receives two copies of the same data packet. The two packets are combined using maximum ratio combining (MRC) [4] to recover the original data. In its simplest form, MRC is modeled by adding the instantaneous signal-to-noise ratio (SNR) of the two packet received from source and partner. This accumulation of the instantaneous SNR increases the rate at which the destination can reliably decode the packet. After the transmission, nodes may return to sleep or listen state. The recipient of the data packet will schedule a transmission for further propagation of the data packet towards the sink.

IV. Results

In this section, we present simulation results for CPS-MAC. Simulations are conducted using Mobility Framework for the OMNET++ discrete event simulator [23]. Our purpose is to show how the protocol behaves and reacts to typical WSN conditions such as fading channels, extended periods of low data flow, and their effect on power consumption. This gives us a good understanding of how deployment on real sensor nodes would perform.

The performance of CPS-MAC is compared with MPS-based MAC protocol mentioned below. This means that the nodes use MPS for waking up neighboring nodes prior to data transmission. For comparison purpose, we have implemented the following network configuration.

1) Direct-MPS: This scenario consists of two nodes, source and destination. The source transmits directly to the destination and uses MPS to wake up the destination node.

 ISBN: 978-1-61208-100-7

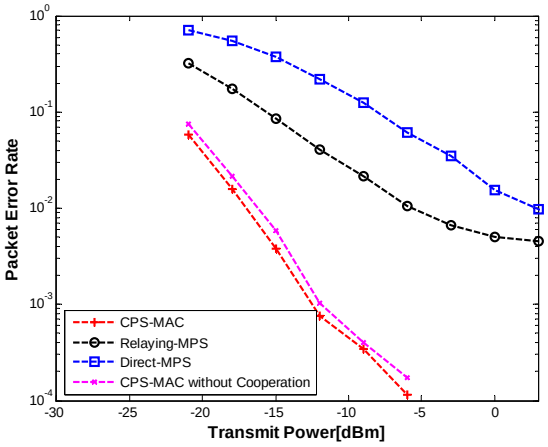


Fig. 10. Packet Error Rate

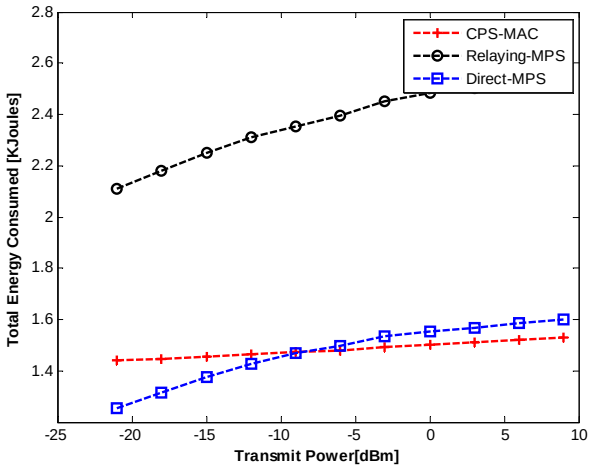


Fig. 11. Total Energy Consumed

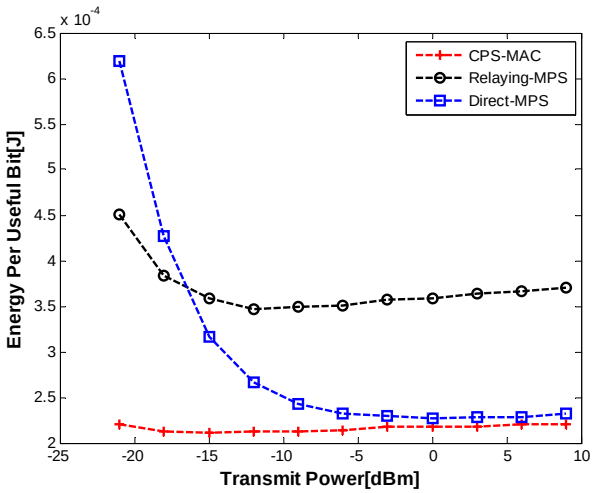


Fig. 12. Energy Per useful bit

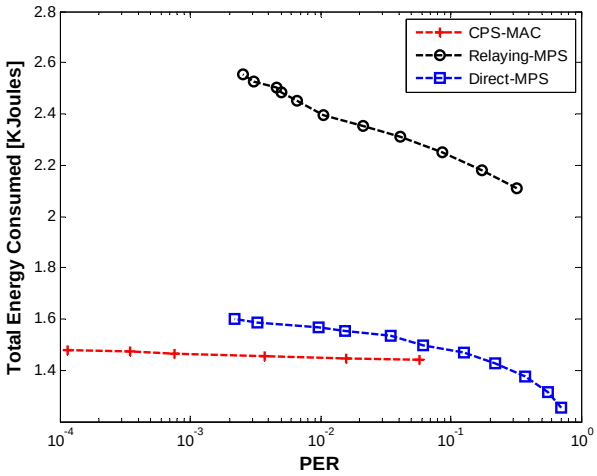


Fig. 13. Total Energy Consumed vs Packet Error Rate tradeoff

2) Relaying-MPS: In this scenario, an intermediate node is introduced between source and destination. The source first wakes up the relay using MPS and transmits the packet. The relay node then wakes up destination and forwards the packet, if correctly received from the source. If a node receives correct packets from both the source and relay, it discards the duplicate packet. This is done by keeping a sequence number of correctly received packets in a table.
3) CPS-MAC: This scenario uses our proposed protocol for a 3 node scenario as shown in Figure 3. We use cooperation to exploit both the source-destination and source-partner-destination channels.
4) CPS-MAC without cooperation: This scenario is similar to the previous one (CPS-MAC) however, cooperation for data packets is disabled. This gives us an idea of how many packets are lost in the absence of cooperation.

Figure 10 shows the Packet Error Rate(PER) for varying transmission power. CPS-MAC here achieves better PER as compared to direct and relaying MPS protocols. We have evaluated CPS-MAC performance both with and without CC. This performance improvement over MPS based protocol is attributed to the CPS-MAC wake up scheme. Repeating the preamble from the partner node increases the chances of the destination node waking up prior to data transmission. This process is similar to CC but here, preamble packet is repeated at the partner station instead of data packet. Thus, the destination would receive multiple copies of the preamble packet, increasing its chances of overhearing the preamble. CPS-MAC-without-cooperation shows the performance of CPS-MAC in the absence of cooperation. The difference in PER between CPS-MAC and CPS-MAC-without-cooperation represents the diversity gain achieved by CC and MRC for data packets. The total energy consumed by the whole network for the entire simulation duration is shows in Figure 11. The energy consumption of CPS-MAC is comparable to direct-MPS and significantly less then relaying-MPS. This is because CPS-MAC is able to wake up the 2-hop destination nodes in a single transmission cycle using repeated preambles from 1-

 ISBN: 978-1-61208-100-7

hop partner node. As the amount of time for waking up the node is significantly larger than the data transmission phase, size and number of preambles is a primary factor contributing to the energy expenditure. By reducing both the number of preambles sent and the time needed to wake up the nodes, CPS-MAC is able to reduce the energy utilization, making it comparable to direct-MPS.

Figure 12 shows the energy consumed per useful bit (EPUB) for the three configurations. The EPUB metric takes into account the energy consumption of all the nodes in the topology. For high transmission power, EPUB for CPS-MAC and direct-MPS is almost the same. However, at low transmission power, the improved PER pays off and CPS-MAC achieves significantly lower EPUB. Figure 13 shows the trade-off between total energy consumption and PER. For a given PER value, CPS-MAC consumes less energy then both Direct-MPS and Relaying-MPS. One thing to notice here is that the Direct-MPS is more energy efficient at very low transmission power, however, the high PER value makes it infeasible for applications where better reliability is desired.

V. Conclusion and Future Work

This work has shown the possible benefits of using cooperative communication to increase the reliability and reduce energy consumption in WSN. We propose CPS-MAC, which improves reliability by using overhearing to its advantage. The improvement is realized by forming cooperative triangle in densely deployed WSN, where channel errors, collisions, idle listening, and overhearing significantly effect the performance. In duty cycling MAC protocols for WSN, the wakeup scheme has a big effect on the packet error rate at the destination. Repeating the preamble in a cooperative manner significantly increases the probability of destination waking up prior to data transmission. Results show that destination is better able to receive and decode packets under this scheme as compared to conventional MPS protocols.

By using CC for data packets, CPS-MAC delivers multiple copies of packet to the destination. Packet combining using MRC further helps CPS-MAC in combining and decoding erroneous packets and reducing the PER. By reducing the number of preambles and time needed to wake up the nodes and transferring data over multiple hops, the network can achieve significant reduction in energy expenditure. This behavior is important in preamble sampling MAC protocols as energy used in sending and receiving preambles is the dominant factor in such protocols. Simulation results show that energy expenditure of CPS-MAC is comparable to direct-MPS protocol and outperforms relaying-MPS.

We are currently planning the performance evaluation of CPS-MAC in a larger WSN configuration. For such a network, in addition to energy utilization, additional paramters such as end-to-end latency and network throughput would also be evaluated.

References

[1] H. Karl and A. Willig, “Protocols and Architectures for Wireless Sensor Networks,” Wiley, May 2005.

[2] J. N. Laneman, D. Tse, and G. W. Wornell, Cooperative Diversity in Wireless Networks: Efficient Protocols and outage Behavior, IEEE Transactions on Information Theory, vol. 50, pp. 3062-3080, December 2004.

[3] A. Nosratinia, T. Hunter, and A. Hedayat, Cooperative communication in wireless networks, Communications Magazine, IEEE, vol. 42, 2004, pp. 74, 80.

[4] A. Meier and J.S. Thompson, Cooperative Diversity in Wireless Networks, 3G and Beyond, 2005 6th IEE International Conference on, 2005, pp. 1 -5.

[5] A. Sendonaris, E. Erkip, and B. Aazhang, User Cooperation Diversity - Part I: System Description, IEEE Transactions on Communications, vol. 51, pp. 1927-1938, November 2003.

[6] A. Sendonaris, E. Erkip, and B. Aazhang, User Cooperation Diversity - Part II: Implementation Aspects and Performance Analysis, IEEE Transactions on Communications, vol. 51, pp. 1939-1948, November 2003.

[7] P. Liu, Z. Tao, Z. Lin, E. Erkip, and S. Panwar, Cooperative Wireless Communications: A Cross-Layer Approach, IEEE Communications Magazine, Special Issue on MIMO Systems, August 2006.

[8] Pei Liu et al., CoopMAC: A Cooperative MAC for Wireless LANs, Selected Areas in Communications, IEEE Journal on 25, no. 2 (2007): 340-354.

[9] Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications, ANSI/IEEE Std 802.11, 1999 Edition, 1999

[10] S. Narayanan and S. Panwar, To Forward or not to Forward - that is the Question, Wireless Personal Communications, vol. 43, Oct. 2007, pp. 65-87.

[11] H. Zhu and G. Cao, rDCF: A Relay-Enabled Medium Access Control Protocol for Wireless Ad Hoc Networks, IEEE Transactions on Mobile Computing, vol. 5, 2006, pp. 1201-1214.

[12] L. Yi and J. Hong, A New Cooperative Communication MAC Strategy for Wireless Ad Hoc Networks, Computer and Information Science, ACIS International Conference on, Los Alamitos, CA, USA, IEEE Computer Society, 2007, pp. 569-574.

[13] W. Ji and B. Y. Zheng. A novel cooperative MAC protocol for WSN based on NDMA. In Proc. of The 8th International Conference on Signal Processing (ICSP06), pp 16-20, 2006.

[14] R. Lin and A.P. Petropulu. A new wireless network medium access protocol based on cooperation. IEEE Trans. on Signal Processing, 53(12):4675-4684, Dec. 2005.

[15] S. Moh, C. Yu, S.-M. Park, H.-N. Kim, and J. Park. Cd-mac: Cooperative diversity mac for robust communication in wireless ad hoc networks. In Proc. of the International Conference on Communications (ICC 07), pp. 3636-3641, June 2007.

[16] A. El-Hoiydi, Aloha with preamble sampling for sporadic traffic in ad hoc wireless sensor networks, 2002 IEEE International Conference on Communications. Conference Proceedings. ICC 2002 (Cat. No.02CH37333), New York, NY, USA, pp. 3418-3423.

[17] M. Buettner, G.V. Yee, E. Anderson, and R. Han, X-MAC: a short preamble MAC protocol for duty-cycled wireless sensor networks, Proceedings of the 4th international conference on Embedded networked sensor systems, Boulder, Colorado, USA, ACM, 2006, pp. 307-320.

[18] S. Mahlknecht and M. Bock, CSMA-MPS: a minimum preamble sampling MAC protocol for low power wireless sensor networks, Factory Communication Systems, 2004. Proceedings. 2004 IEEE International Workshop on, 2004, pp. 73-80.

[19] B. Mainaud, V. Gauthier, and H. Afifi, Cooperative communication for Wireless Sensors Network : A Mac protocol solution, Wireless Days, 2008. WD ’08. 1st IFIP, 2008, pp. 1-5.

[20] M. Rossi and M. Zorzi, Integrated Cost-Based MAC and Routing Techniques for Hop Count Forwarding in Wireless Sensor Networks, IEEE Transactions on Mobile Computing, vol. 6, 2007, pp. 434-448.

[21] F. Ye, A. Chen, S. Lu, and L. Zhang, A scalable solution to minimum cost forwarding in large sensor networks, Proceedings Tenth International Conference on Computer Communications and Networks (Cat. No.01EX495), Scottsdale, AZ, USA, pp. 304-309.

[22] K. Langendoen, Medium Access Control in Wireless Sensor Networks, Medium Access Control in Wireless Networks, H. Wu and Y. Pan, Eds., Nova Science Publishers, Inc., 2008, pp. 535-560.

[23] OMNET++ discrete event simulator, http://www.omnetpp.org, June 2010.

 ISBN: 978-1-61208-100-7

Topological Cluster-based Geographic Routing in Multihop Ad Hoc Networks

Emi Mathews
Heinz Nixdorf Institute
University of Paderborn, Germany
Email: emi@hni.upb.de

Hannes Frey
University of Paderborn, Germany
Email: hannes.frey@uni-paderborn.de

***Abstract*—Existing geographic routing algorithms face serious challenges due to location errors, nonplanarity issues and overhead of location service. To solve these issues, we propose Topological Cluster Based Geographic Routing that combines topology-based routing and geographic routing. It is a localized routing scheme where the geographic routing is performed on an overlay network of topological clusters. Preliminary results from simulations show that the overlay graph created by topological clustering has the potential to create planar graphs even with realistic wireless models. Hence, the typical Greedy-FACE-Greedy protocol used in geographic routing works in these overlay graphs and makes the geographic routing applicable in realistic wireless networks. Moreover, due to the topology-based multi-hop clustering which we apply in this work, the proposed routing has the potential to subside node localization errors and reduce the overhead of the location service.**

Keywords*—geographic routing; multi-hop clustering; overlay graph; graph planarization; location fault tolerance.

I. INTRODUCTION

Efficient routing in mobile ad-hoc networks is a challenging task due to their highly dynamic network topology, bandwidth constrained links and resource constrained nodes. Routing protocols developed during the earlier stages were based on the topology information, which used the knowledge about the links between the nodes to establish and maintain end-to-end paths for routing. Topology-based routing has significant overhead in maintaining up-to-date paths between source and destination, if the topology changes frequently. Moreover, it has scalability issues, as the overhead increases according to the number of nodes in the network.

With the availability of small, inexpensive, and low-power Global Positioning System [1] receivers or other localization systems, routing based on the geographic information gained significant attention. In geographic routing, a node uses the position information of the sender S, its neighboring nodes, and the destination node D, to move the packet further toward the destination at each hop.

The simplest form of geographic routing is to greedily forward the message to the node which minimizes a local forwarding metric. A simple example of such a metric is the distance to the destination. In this case, the message is forwarded to the neighbor node which minimizes the distance to the destination.

In all greedy routing variants the message may arrive at a node, which compared to its neighbors is the best with respect to the routing metric applied. In this case, greedy routing has to stop and an alternative routing mode called void handling has to be used to circumvent the void regions.

Void handling based on routing in planar geometric graphs attracted a major amount of research effort due to its stateless property and guaranteed delivery in planar graph models [2], [3]. The planar graph algorithms uses graph traversal rules (right hand rule and face changes rules) to find a path from the source to the destination on planar graphs along the boundaries of the void regions.

Existing geographic greedy routing and void handling methods do not require the establishment and maintenance of end-to-end paths. Thus, compared to topological routing, changes in topology have less impact. However, the following shortcomings of geographic routing algorithms are known:

Localization errors: Location measurement is often noisy and incurs errors. Even small errors can lead to incorrect, non-recoverable geographic routing with noticeable performance degradation. Kim et al. [4] shows that location inaccuracy ($\leq 20\%$ of radio range) caused forwarding loops, and packet drops reaching up to 54 percent.

Void handling issues: The planar graph-based void handling guarantees message delivery only if the graph is planar. Hence a planarization step, prior to routing is required. However, localized planarization processes that work in realistic wireless models do not exist. Moreover, localization errors as discussed above, causes incorrect edge removal and may disconnect the planar subgraph [4]. Hence an efficient localized planarization algorithm that works in realistic wireless models with location fault tolerance is needed.

Location service: Geographic information of the destination node is required to send packets to the destination. Typically a location service maintains the up-to-date positions of the nodes in the network and the source directly asks the location service for destination position information. The overhead in maintaining a location service with up-to-date positions of all nodes in the network is very high; especially at high node mobility [5].

To address the above mentioned problems, we propose Topological Cluster Based Geographic Routing (Topogeo), which is a combination of small scale topology-based and

 ISBN: 978-1-61208-100-7

large scale geographic-based routing. The principal idea behind the Topogeo routing is presented in Section II. Preliminary results of the Topogeo performance analysis on graph planarization are discussed in Section III. Finally, Section IV summarizes the findings of this work and provides an outlook on possible future research.

II. Topological Cluster Based Geographic Routing

Topogeo combines topology-based routing and geographic routing. Here, the geographic routing is performed on an overlay network of topological multi-hop clusters. The cluster heads constitute the vertices of the overlay graph. The cluster heads collect information about their neighboring clusters and establish links towards the neighboring cluster heads. These links constitute the edges of the overlay graphs. Since the cluster hop counts are limited and the cluster head keeps information only about the neighboring clusters, Topogeo remains a localized routing algorithm.

Using topology-based multi-hop clustering to create a robust overlay graph for geometric routing, is a novel idea. Geographic routing performed on an overlay graph of cluster heads (CH) was already reported; for instance in [4], [6]–[8]. Contrary to these works that create geographic clusters or geographic cells, our work is based on the idea of topological clustering. Terminode Routing [9] combines hierarchical and geographic routing. The hierarchy created in Terminodes is by using anchor nodes, where, as in Topogeo the hierarchy is created by topological clustering. Moreover, Topogeo is a localized geographic routing, where as Terminodes needs a global map of the anchor nodes or path discovery protocols to obtain the anchored path.

Topological Cluster Based Geographic Routing has three different processes, namely 1) Multi-hop clustering and cluster head selection 2) Overlay network formation and 3) Topogeo routing. Topogeo uses any multi-hop clustering algorithm, e.g, k-hop connectivity ID algorithm [3], for clustering and cluster head selection. After the clustering process, virtual nodes are created at the center of gravity of the clusters. These nodes are chosen as the vertices of the overlay graph, as they are more location fault tolerant than the real cluster heads. In principle, a node close to the location of virtual cluster head (VCH) or the real cluster head itself serves the functions of the VCH. Now, virtual links are created between neighboring VCHs. After the overlay graph formation, the Topogeo routing process is executed.

Topogeo routing uses topological information of the cluster member nodes for routing within the clusters. During the clustering process, each cluster head collects topology information about its member nodes. The cluster member nodes keep information about the path to reach their cluster head. This information is used for the topology-based intra-cluster routing. For routing beyond the source node's cluster, Topogeo uses geographic routing. In geographic routing, the typical Greedy-Face-Greedy type protocol [2] is used to route packet along the edges of the overlay graph. In the greedy-forwarding mode of the protocol, each cluster head forwards the packets to its neighboring cluster head based on any forwarding criterion that guarantees loop-free paths; e.g., the most-advance based forwarding [10] criterion. When the packets get stuck at voids during greedy-forwarding, they use the planar graph traversal based on Face routing [3] for void handling. This helps the packets to circumvent the void region. Once they circumvent the void region, the routing is switched back to the greedy forwarding mode. When the packets reach the cluster head of destination node, the local topology routing protocol is used to forward the packets to the recipient node.

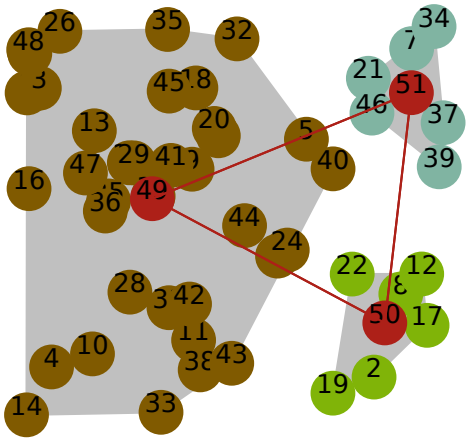


Fig. 1: Overlay graph created with 2-hop clustering for illustration

Planar-graph-based void handling guarantees message delivery only if the graph is planar. Hence a planarization step that works in realistic wireless models is needed. Topogeo creates an overlay graph that is almost 100% planar, even with realistic wireless models such as Log normal Shadowing model [11] as observed in our experiments. The planarity of the overlay graph is achieved with the multi-hop clustering step.

Figure 1 shows a small network created with 50 nodes and average node degree 9.4 (3π) for illustration. The edges of the original network are not shown in the figure for better visualization of the overlay graph. On performing 2-hop clustering, three clusters are formed with the cluster member nodes displayed in three different colors, along with their node ids. The silver colored area shown in the background of clusters, is the convex hull region of their member nodes. The vertices of the overlay graph are shown in brick-red colored circles and the edges in brick-red colored line segments.

Topogeo addresses the shortcomings of localized geographic routing discussed in Section I as follows. It is location fault tolerant, as Geographic routing is performed at the cluster head level rather than at the node level. Hence the location and link errors of individual nodes do not directly reflect on the routing decisions in the overlay graphs; especially in dense networks having several gateway nodes between each clusters or cells. Moreover, the virtual cluster head creation and large cluster formation, subside the discrepancy in the location of nodes which will be validated in our future experiments.

 ISBN: 978-1-61208-100-7

In Topogeo, the location service keeps and maintains the position information only for virtual cluster heads. For each individual node in the network, now only the node's cluster head id is stored in the location service. Hence the location service need not update information of each node in the network while it has a position change. Location information about a node only needs to be updated when it changes its cluster. Moreover, individual node movements do not change the position of virtual cluster heads significantly and hence the virtual cluster head position update is less frequent.

The overlay graph created by Topogeo has almost no intersections even with realistic wireless models. Hence the non-planarity issues of planar-graph-based void handling in realistic wireless models is solved and FACE routing works with guaranteed delivery in such overlay graphs.

III. Experimental Analysis and Preliminary Results

We have implemented the multi-hop clustering algorithms such as k-CONID (k-hop connectivity ID) algorithm [3] and Max-Min D-Cluster algorithm [12] for clustering and cluster head selection process. After this process, the virtual nodes and virtual links of the overlay graph are created. The third process, the Topogeo routing, works on any existing topology-based and geometric routing algorithms for intra-cluster and inter-cluster routing respectively. The evaluation of Topogeo performance on its ability to solve the shortcomings of localized geographic routing discussed in Section I is independent of the implementation of Topogeo routing process.

To make statistical estimations on the planarity of the overlay graphs, we conduct various experiments on different test networks. We use the ShoX network simulator [13] for network creation and overlay graph formation. Networks with field sizes 500×500 containing 100 to 500 nodes, and another one with 2500×2500 containing 2500 to 12500 nodes, are used in these experiments. For each field size, 100 different network configurations are created for a specific average node degree value. We use more than 1000 different networks in total for the evaluation.

We use the Log normal shadowing (LNS) model [11] for our experiments. The path loss at a distance d expressed in decibel by the LNS model is given as:

$$PL(d)[dB] = PL(d_0)[dB] + 10\gamma \log_{10} \frac{d_0}{d} + X_\sigma[dB], \text{ where} \quad (1)$$

γ is the path loss exponent, X_σ is a zero-mean Gaussian random variable with variance σ^2, and $PL(d_0)$ is the reference path loss at a reference distance d_0. We choose realistic values $\gamma = 3.25$ and $\sigma = 2.5$ for our experiments. We assume $P_{tx} = 15\ dBm$ and the reference path loss $PL(d_0) = 40\ dBm$ for the reference distance $d_0 = 1\ m$. We set the receiver sensitivity at 80 dBm.

The preliminary results obtained on the planarity of the overlay graphs show that overlay graphs created from multi-hop clustering are almost 100% planar especially at higher hop counts. We are working on the strategies to create overlay graphs that are planar even at lower hop counts.

To compare the performance of the Topogeo planarization, a Gabriel Graph (GG) [2] based planarization algorithm is also implemented. GG creates a planar graph locally from the full graph. It provably yields a connected planar graph on Unit Disk Graphs. When the GG based planarization algorithm is performed on the LNS model, more than 80% of the test cases produced are nonplanar graphs, where as almost 100% graphs are planar with Topogeo.

IV. Conclusion and Future Work

We proposed Topological Cluster-based Geographic Routing, a new routing scheme which combines topology-based routing and geographic routing. The geographic routing is performed on an overlay network created from multi-hop clustering. For routing within the clusters, Topogeo uses topology information and for inter-cluster routing, it uses the typical Greedy-FACE-Greedy routing scheme. The Greedy-FACE-Greedy routing guarantees message delivery only if the network graph is planar. Nevertheless, localized planarization processes that work in realistic wireless models have not yet been found. The preliminary results obtained from simulations show that the overlay graphs created by the multi-hop clustering are almost 100% planar even with realistic wireless models like Log Normal Shadowing model. Hence the Greedy-FACE-Greedy routing works in these overlay graphs and makes the geographic routing applicable in realistic wireless networks. More investigation on multi-hop clustering algorithms that create overlay planar graphs using local information need to be done. This work shows that the geometric properties of the overlay graph produced from topological clustering, is an area worth exploring more.

Due to the multi-hop clustering, it is evident that individual node localization errors do not affect the performance of Topogeo routing and the overhead of a location service is reduced significantly. We are currently working in this direction to provide more quantitative results in future. We are also working towards the implementation of Topogeo routing process.

Acknowledgment

Emi Mathews acknowledges with appreciation, the research fellowship from the International Graduate School, Dynamic Intelligent Systems, University of Paderborn, Germany.

References

[1] B. Hofmann-Wellenhof, H. Lichtenegger, and J. Collins, *Global Positioning System: Theory and Practice*. Springer-Verlag, 1997.

[2] B. Karp and H. T. Kung, "GPSR: Greedy perimeter stateless routing for wireless networks," in *MobiCom '00: Proceedings of the 6th annual international conference on Mobile computing and networking*. New York, USA: ACM, 2000, pp. 243–254.

[3] G. Chen, F. G. Nocetti, J. S. Gonzalez, and I. Stojmenovic, "Connectivity-based k-hop clustering in wireless networks," in *HICSS '02: Proceedings of the 35th Annual Hawaii International Conference on System Sciences*, vol. 7. Washington, DC, USA: IEEE Computer Society, 2002, p. 188.3.

 ISBN: 978-1-61208-100-7

[4] Y. Kim, J.-J. Lee, and A. Helmy, "Modeling and analyzing the impact of location inconsistencies on geographic routing in wireless networks," *SIGMOBILE Mobile Computing and Communications Review*, vol. 8, no. 1, pp. 48–60, 2004.

[5] D. Chen and P. K. Varshney, "Geographic routing in wireless ad hoc networks," in *Guide to Wireless Ad Hoc Networks*. Springer London, 2009, pp. 1–38.

[6] H. Frey and D. Görgen, "Planar graph routing on geographical clusters," *Ad Hoc Networks*, vol. 3, no. 5, pp. 560–574, 2005.

[7] J. Lin and G.-S. Kuo, "A novel location-fault-tolerant geographic routing scheme for wireless ad hoc networks," in *IEEE International Conference on Vehicular Technology Conference*, vol. 3. Melbourne, Australia: IEEE Computer Society, 2006, pp. 1092 – 1096.

[8] J. Cao, L. Zhang, G. Wang, and H. Cheng, "SSR: Segment-by-segment routing in large-scale mobile ad hoc networks," in *IEEE International Conference on Mobile Adhoc and Sensor Systems Conference*, vol. 0. Los Alamitos, CA, USA: IEEE Computer Society, 2006, pp. 216–225.

[9] L. Blažević, S. Giordano, and J.-Y. Le Boudec, "Self organized terminode routing," *Cluster Computing*, vol. 5, no. 2, pp. 205–218, 2002.

[10] T. Melodia, D. Pompili, and I. Akyildiz, "Optimal local topology knowledge for energy efficient geographical routing in sensor networks," in *INFOCOM 2004: Twenty-third AnnualJoint Conference of the IEEE Computer and Communications Societies*, vol. 3, Hong Kong S.A.R., People's Republic of China, 2004, pp. 1705 –1716.

[11] H. Karl and A. Willig, *Protocols and Architectures for Wireless Sensor Networks*. John Wiley & Sons, 2005.

[12] A. Amis, R. Prakash, T. Vuong, and D. Huynh, "Max-min d-cluster formation in wireless ad hoc networks," in *INFOCOM 2000: Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies*, vol. 1, 2000, pp. 32 –41.

[13] "Shox – a scalable ad hoc network simulator," accessed on May 1, 2010. [Online]. Available: http://shox.sourceforge.net

 ISBN: 978-1-61208-100-7

Optimizing Parameters of Prioritized Data Reduction in Sensor Networks

Cosmin Dini
University of Haute Alsace
34 rue du Grillenbreit 68008 Colmar - France
France
cosmin.dini@uha.fr

Pascal Lorenz
University of Haute Alsace
34 rue du Grillenbreit 68008 Colmar - France
France
lorenz@ieee.org

***Abstract*— Wireless Sensor Networks are popular and proven useful in various service areas. Energy optimization, processing optimization and storage optimization are the main challenges. While there is a debate for complete or partial data extraction from sensors, having special data process functions and operation primitives proves useful for sensor operating systems. To deal with robustness and reliability, data processing at the network/sensor level satisfies some of the reliability requirements, especially when communications are not operational. There are situations where data reduction is an alternative when storage is no longer available and data is accumulating, especially when some sensor links are not operational. Using predictions and optimized parameters to prioritize data reduction is a solution. In this article, we define special heuristics for data reduction using a set of data processing primitives and special data parameters. We apply these heuristics via a methodology that enables various factors, both internal and external to the sensor, to influence the data aging process and the data reduction operations.**

Keywords – sensor, data management, data sensor storage, data priority optimized parameters, prediction models.

I. INTRODUCTION

Requirements posed by unattended data collections in remote areas become very challenging for traditional network deployments. The main problem is raised by the fact that users might look for full collected data, while effective business models take into consideration a small fraction of it.

Most of the WSNs (Wireless Sensor Networks) also perform the essential functions for data processing; one of the most important, in special cases of uncontrolled link availability, is data reduction under several the constraints driven by the nature of the data, the relevance of the data, the data dependency, and the business model using such data. A sensor on a node captures a time series representing the evolution of a sensed physical variable over space and time. Reducing the amount of data sent throughout the network is a key target for long-term, unattended network sensors. A second target, equally relevant, is defined by unattended networks with unreliable links. In this case, gathered data may be rapidly aging and could exceed the storage availability on a given node. Data reduction mechanisms are used to partially handle these cases [1][2][15].

Unnecessary communication as well as appropriate data reduction techniques can be modeled in the case of physical phenomena with a pre-defined, application-dependent accuracy [15]. If an accepted measurement error is bounded as [-*e*, +*e*], only values exceeding the predicted one by +/-*e* will be considered. Similarly, if the errors of the gathered values are within the bonded interval, data reduction can be further simplified in the context of repeated equal measurements.

In this paper, we present a series of heuristics on predictions used to summarize collected data. These heuristics are based on past experience in parameter variation and on the intended use of the data. At the two extremes of data usage are refinement and discovery. Data refinement approaches collection from a perspective of pure prediction where more data is collected around already confirmed scenarios. Data discovery on the other hand tends to ignore known data value patterns by putting more weight on unpredictable corner case scenarios. The difference between the two situations determines what reduction rules are used and how data importance is computed.

The rest of the paper has the following flow. Section 2 introduces the state of the art concerning data management and predictions in sensor networks. Section 3 revisits the model used to reduce collected data. Section 4 includes heuristics for prediction on data processing. Section 5 concludes and identifies future work.

II. RELATED WORK

In this section, we summarize a data processing model introduced in [1][2] and prediction approaches for data reduction [3].

In the past, the database community pushed different data reduction operators, *e.g.*, aggregation and reduction, with no enough flexibility to handle extracting complete raw sensor readings (i.e., using "SELECT *" queries).

There are two specific needs to perform in-network data processing, *i.e.*, (i) to significantly reduce communications costs (energy), and (ii) to deal with link-down situations. In-network aggregation was proposed in [5][6], while data reduction via wavelets or distributed regression in [7][8]. All these techniques do not provide the desired data granularity, as requested by network users.

Managing data in a storage-centric approach was studied in different approaches, based on a reliable connection [11],

 ISBN: 978-1-61208-100-7

additional buffering [9], or collaborative framework [10]. Details on collaborative storage are provided in [2].

OS primitives acting on recurring and non-recurring data collection have been proposed in [1]. Mainly, compression, thinning, sparsing, grain coursing, and range representation were used to deal with data aging in a pessimistic and optimistic approach. As a note, data deduplication was not considered in the above model. To optimize data reduction, concepts of data units, data importance, and compensation factors were introduced in [2]. Mainly, measurements are partitioned into contiguous intervals; data importance relates to the business semantics, while the compensation factor affects the importance in business computation of a given data unit after it has undergone some type of data reduction. The model considers that there is data that cannot be reduced in any circumstance. A mechanism for data dependency between different data units was presented in [2]. Further division of the data units (leading to more flexibility) was not considered at this point. Associated with the new concepts, the following functions were introduced: interval production function (only for recurring data), default compensation function, data importance function, and data reduction function. Solutions based on data redundancy (leading to more robust deployments and measurements) were not considered.

A data reduction specification use case considering data dependency across data units was presented in [2]. Consequently, appropriate values for data importance can be derived considering the constraints on the data importance computation as pertaining to two categories: (i) internal constraints and (ii) external constraints. External constraints are caused by factors over which input data has no effect. Such factors are data age and inherent interest in the data depending on the exact purpose of the data collection. Internal constraints represent inter- and intra-data dependencies.

A prediction model for approximate data collection is presented in [3]. The techniques are based on probabilistic models (BBQ system, [12]). We apply the prediction models to the framework proposed in [1][2] considering also the approximation scheme providing data compression and prediction [13] and predictive models from [14].

In this paper, we consider the PDR components introduced in [2] (Figure 1) to derive appropriate data reduction considering known correlations (spatial, temporal, etc.) and prediction models.

III. BASIC MODEL FOR OPTIMIZATION

We consider a simplified sensor model introduced in [2] with the following modules:

- Storage Engine (SE)

The SE is concerned with writing data to the node's storage. It makes no judgment as to the relevance or importance of the data itself. It simply follows data collection rules established by the business case and sends them to the node's permanent storage. At this time, there may enough space on the storage device in which case the data is simply recorded, or there isn't enough space at which point some data reduction occurs: either on the incoming data, existing data, or both

- PDR Engine (PDRE)

The PDRE contains all the data reduction rules, which are a direct reflection of the business case. They are not constantly applied, but at specific times and with specific space recovery objectives as dictated by the PDR Controller

- PDR Controller (PDRC)

The controller is responsible for monitoring the state of the available storage, and, if dictated by the business case, triggers the PDRE to perform data reduction operations. Deciding what data to target and how much to reduce it is again subject to the business requirements.

Figure 1 shows the interaction of the components shown above: green represents data flow and red represents control paths.

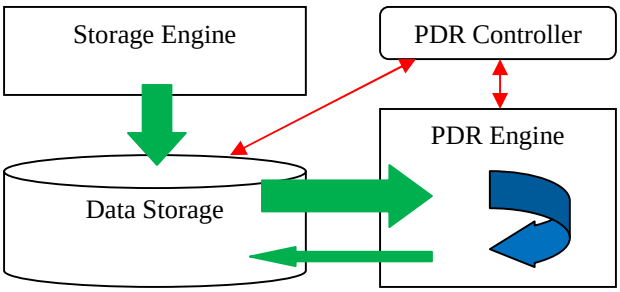


Figure 1. Interaction of main PDR components [2]

Two collection modes are allowed, *i.e.*, Recurring Data Instance (RDI) that represents a series of measurements taken at specified intervals, starting at a given time and ending at a given time, and (ii) Non-Recurring Data Instance (NRDI), representing a single measurement of a parameter at a specific time.

We introduced the following primitive operations that are the basic actions that the PDRE employs to actually decrease the amount of data. Here is a quick review of what they are and how they operate:

- *compression*: a simple data compression algorithm is applied which reduces the amount of space used, but any

 ISBN: 978-1-61208-100-7

further data reduction cannot be applied to the compressed data

- *thinning*: in the case of an RDI, a section of data corresponding to a time interval is discarded
- *sparsing*: in the case of an RDI, for a specific interval, the data sampling rate is decreased and excess data is discarded
- *grain coarsing*: the resolution/precision of a data instance is decreased
- *range representation*: an entire interval of an RDI is replaced by a tuple reflecting on the data that was discarded: minimum value, maximum value, and the average during that interval form the tuple.

To include a complete characterization of the data reduction mechanisms, a few concepts were introduced:

- *Data Units:* Data collections can be both recurring and non recurring. The non recurring collections generate data that is stand alone and considered atomic. It makes sense to consider a non recurring data record as a single data unit (**DU**). Recurring data collections have several values over a potentially long period of time.
- *Data Importance:* Data importance, denoted as **I**, is a value that numerically reflects the relevance of a data unit for the business case.
- *Compensation Factor:* The compensation factor, **K**, is an importance modifier that reflects the data reductions that have already been performed on the specific data unit.

Data importance depends on the business model. In this section, we identify input factors that can be used to establish the importance of a data unit, such as age/collection time, self values, other data units of same instance, and other data instances [2].

The following fucntions are needed to handle the newly introduced concepts for RDI and NRDI

- *interval production function (for RDI only)*
- *default compensation factor*
- *data importance function*
- *data reduction operation function*

In the next section, we present different predictive heuristics for data processing, using the model exposed in Section III.

IV. OPTIMIZATION AND PREDICTION HEURISTICS

Let us assume a deployment of sensors that have the ability to measure the UV Index. The UV Index is a measurement of ultraviolet rays intensity and has a value between 1(low) and 11+(extremely high). The value of this index is collected for the purpose of gaining precise insight into variations during the course of the year.

Expected results are already available:

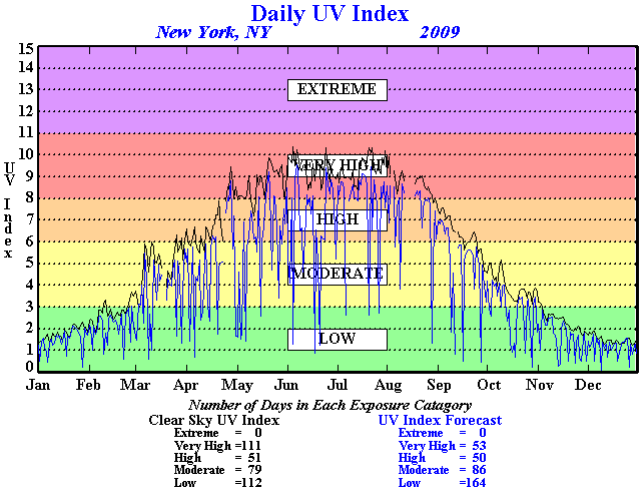


Figure 2. Data UV Index for New York, 2009 [16]

Figure 2 presents the UV Index reading for New York during 2009. We observe that the UV Index has lower values during winter and higher values during summer. There is a natural difference that is expected during a 24 hour cycle, but there are also less obvious dips in the UV Index values. The probable cause for this would be overcast conditions. This naturally interferes with the data collection and is of no interest.

On the other hand, let's consider the example of CO_2 concentration collection in an isolated forest area on a somewhat active volcano. The point of this is not to refine data, but to monitor CO_2 and possibly offer an explanation to unexpected values. CO_2 increase could be caused by a fire in the area or volcanic activity. In such a case, correlations are to be made with seismic sensors, and with temperature and visibility sensors.

In Figure 3 we have a section of interest in a deployment where the objective is not data refinement, but data discovery. We seek correlation between different parameters and possibly seek causality relationships.

 ISBN: 978-1-61208-100-7

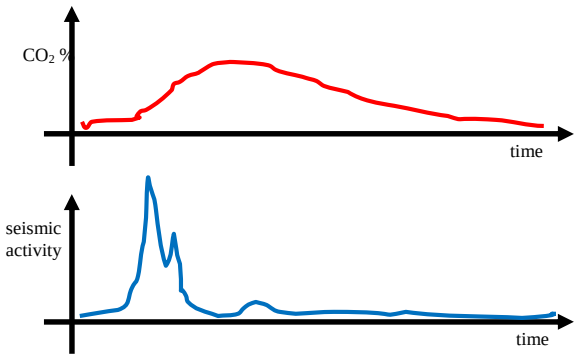


Figure 3. Parameter correlation

The cases presented in Figure 2 and Figure 3 are opposite ends of a spectrum of cases. In Figure 2 we seek to confirm and gain better resolution with respect to expected parameter values, while in Figure 3 we seek to explain observed but not expected parameter values. These cases require different approaches with respect to data reduction.

The data reduction rules that are deployed need to give to the PDR Engine (Figure 1) possibilities to select a reduction approach. Each possibility has different expected outcomes with respect to freed space, effect on the compensation factor, and amount of data importance parameters to be re-calculated.

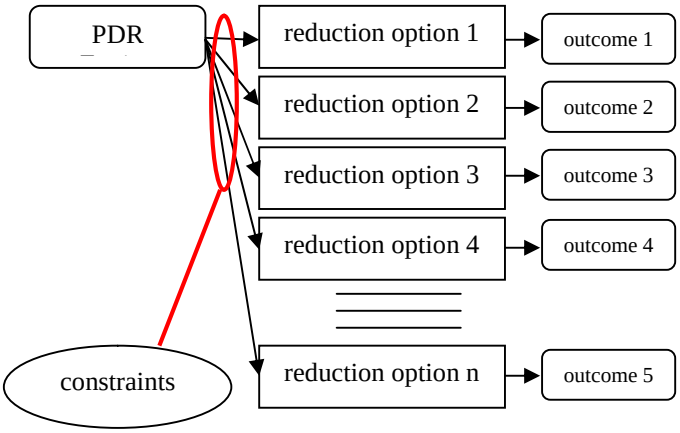


Figure 4. Selecting reduction operation

To investigate optimization of data reduction we define an optimization function as follows:

{fi} = {reduction option 1, . . , reduction option n}

{gj} = {interval production function, compensation factor, data importance, data reduction}

F (x, y) = {fi x gj | constraints}, (1)

where

constraints represent the dependability relations among correlated data.

constraints ::= {context | bounding}

We define the *context constraints* as follows; let *x* and *y* be two variable to gather the values for, and "*oo*" an operator that is defined by

oo = {same, opposite, nil} with the following semantic

x same *y* = when the collection of x is more frequent, then the collection of y should increase too
x opposite *y* = when the collection of x is more frequent, then the collection of y should decrease
x nil *y* = collection of x and y are independent

We define the *bounding constraints* as following:

{x | [-*e*, +*e*], with *e* in *R+*} → p(x/e, y/e') → (2)
{y | [-*e'*, +*e'*], with *e'* in *R+*},

with the semantic: a computation F is not necessary when x's values hold in the bound interval defined by +/-e with the prediction that y's values are bound by +/-e'.

p(x/e, y/e') is the probability that the y's variations hold in the given interval when x's variations are in a given interval; p(x/e, y/e') is derived from the prediction model.

The optimization function triggers the computation of K according to the primitive operations applied for data reduction assuming the prediction model holds. This is an important decision for saving computation power and energy.

These formalisms are used to specify the behavior of the PDR controller mentioned in Figure 1.

We can improve the presented data reduction function considering an average model, i.e., an average ~X of {xi} in the given approximation bounds +/-e predicts a ~Y of {yi} in the +/-e' bounding interval. Computation of ~X and ~Y using in-network data aggregation reduces the consumed computational resources for F, with a reasonable approximation.

Other parameters have to be considered in the data reduction, based on the fact that:

- Some sensing features exhibit typical correlations, *e.g.*, correlation preciseness is inversely proportional with the distance between sensors (temperature)

- The data units can be smaller, especially when working with average values.

 ISBN: 978-1-61208-100-7

Extensive simulations for identifying correlations patterns are needed. Most of the time, in sensor network, and approximate answer is more useful; in special applications, extensive experiments should be conducted to evaluate the performance of data reduction under error threshold define by the bounding constraints. Considering the datasets mentioned in [15], we conclude that a reasonable identification of a correlation pattern requires a sensing period of about one year with a number of samples exceeding 7-8 thousands. For some applications, a tolerated prediction error can relax the bounded interval [$-e$, $+e$].

V. CONCLUSION AND FUTURE WORK

Data reduction in unattended sensor networks with intermittent or non reliable connections is an important computation when considering data storage and data aging.

In this paper, we proposed an optimization function using prediction to map the use of data reduction primitives, optimization parameters (K, I) and dependency constraints (contextual or bounding). The model considers a probability that a variation of a variable is correlated with a variation of another variable. A variant of average values can also be considered.

A simple use case had shown the nature of dependencies and computation challenges for two correlated readings.

Accurate evaluation of the model requires extensive simulations, where combinations of the primitives and data parameters are combined with various type of constraints. We estimate that finding some correlation patterns will favor the use of average values, leading to a reasonable computation effort.

REFERENCES

[1] C. Dini and P. Lorenz, “Primitive Operations for Prioritized Data Reduction in Wireless Sensor Network Nodes”, Proceedings of the 2009 Fourth International Conference on Systems and Networks Communications, September 2009, Porto, Portugal, ICSNC 2009, pp. 274-280

[2] C. Dini and P. Lorenz, “Prioritizing Data Processing in Wireless Sensor Networks”, Proceedings of the 2010 Sixth International Confenrce on Networks and Services, March 2010, Cancun, Mexico, ICNS 2010, pp. 23-31

[3] D. Chu, A. Deshpande, J. M. Hellerstein, and W. Hong, “Approximate Data Collection in Sensor Networks usingProbabilistic Models”, Proceedigns of the 22nd International Conference on Data Engineering. ICDE 2006, Atlanta, USA http://www.cs.umd.edu/~amol/papers/icde06.pdf [Retrieved: August 20, 2010]

[4] A. Mainwaring, D. Culler, J. Polastre, R. Szewczyk, and J. Anderson, “Wireless sensor networks for habitat monitoring”, International Workshop on Wireless Sensor Networks and Applications, September 28, 2002, Atlanta, Georgia, WSNA’02, pp. 88-97

[5] C. Guestrin, P. Bodik, T.R., P. Mark, ans S. Madden, “Distributed regression: an efficient framework for modleing sensor network data”, IPSN, 2004

[6] S.Madden, M.J. Franklin, J.M. Hellerstein, and W. Hoing, ”Tag: a tiny aggregation service for ad hoc sensor networks” , SIGOP Oper. Syst. Rev. 36(SI):131-146, 2002

[7] S. Nath, P.B. Gibbons, S. seshan, and Z.R. Anderson, “Synopsis diffusion for robust aggregation in sensor networks”, Proceedings of the 2nd ACM Conference on embedded Networked Sensor Systems, SenSys 2004, Baltimore, MD, USA, 2004

[8] J. Hellerstein and W. Wang, “Optimization of in-network data reduction”, DMSN, 2002

[9] M. I. Khan, W. N. Gansterer, and G. Haring, “In-Network Storage Model for Data Persistence under Congestion in Wireless Sensor Network:, First International Conferenece on Complex, Intelligent and Software Intensive Systems, April, 2007, Viena, Austria, CISIS’07, pp. 221-228

[10] S. Tilak, N. B. Abu-Ghazaleh, and W. Heinzelman, “Collaborative storage management in sensor networks”, International Journal of Ad Hoc and Ubiquitous Computing, Volume 1, Issue 1/2 (November 2005), pp. 47-58

[11] Y. Diao, D. Ganesan, G. Mathur, and P. Shenoy, “Rethinking Data Management for Storage-centric Sensor Networks”, Proceedings of the Third Biennial Conference on Innovative Data Systems Research (CIDR), Asilomar CA, January 7 - 10, 2007.

[12] A. Deshpande, C. Guestrin, S. Madden, J. Hellerstein, and W. Hong, “Model-driven data acqusition in sensor networks”, VLDB, 2004

[13] I. lazaridis and S. Mehtotra, “Capturign sensor-generated time series with quality guarantees”, ICDE, 2003

[14] S.M. Graham Cormode, M. Garofalakis, and R. Rastogi, “Holistic aggregates in a network world: Distributed tracking of approximate quantiles, SIGMOD, 2005

[15] Y.-Ae¨ . Le Borgne, et al., “Adaptive model selection for time series prediction in wireless sensor networks”, Signal Process. (2007), doi:10.1016/j.sigpro.2007.05.015

[16] National Weather Service Climate Prediction Center [Retrieved; August 28,2010] http://www.cpc.ncep.noaa.gov/products/stratosphere/uv_index/gif_files/jfk_09.png

 ISBN: 978-1-61208-100-7

Is the Mashup Technology Mature for its Application in an Institutional Website?

Serena Pastore
INAF – Astronomical Observatory of Padova
Vicolo dell'Osservatorio 5 – 35122
Padova, ITALY
e-mail: serena.pastore@oapd.inaf.it

Abstract—**Web design for general-purpose websites today requires framework software infrastructure built from different components. Among web server software as the front--end needed to process HTTP/HTTPS requests, the top-down layers consist of different software middleware of which CMS (Content Management System) technology is a primary component. However, with the proliferation of websites and the advent of Web 2.0 philosophy, much distributed information has become structured in interoperability file formats (i.e., XML, RSS, JSON, etc.), and thus, new technologies such as mashups have been developed to collect this information in a dynamic way. The document approaches mashup technologies in order to evaluate their application not only in an end-user environment to create a personalized start page that embeds different sources of information relevant to a specific user, but also in generic areas such as the development of an institutional website in terms of creating a business mashup.**

Keywords - Mashup applications; web widgets; website platform; XML formats; CMS technology; Web 2.0 tools.

I. INTRODUCTION

Most websites today are built upon a software infrastructure made of different components of which Content Management System (CMS) technology [1] is the primary one. Among the variety of commercial and open source solutions, the common aspect is the presence of a database management system (DBMS) where content is stored and an engine that queries such a system transforms content in an HTML page linked to stile sheets following some template. So, these days, communicating over the web means, in most cases, the setup of a website using a web CMS software and its integration in several ways with one or more of the so-called Web 2.0 tools [2], i.e., web systems able to allow a single user to interact by presenting videos stored on YouTube [3] or creating a link with Facebook [4], Twitter [5] and so on. The applicability in a broader context is called Enterprise 2.0 [6], a term that describes an organizational model for companies and the application of technologies aimed at distributing information and services through a collaborative lens.

Most distributed content is available in an interoperability format such as XML technologies [7] or related structures specific for some functionalities (i.e., content-specific markup languages such as RSS [8], Atom [9], JSON [10], KML [11], etc.) that share a markup language used to describe the content.

From a technical perspective, this has become a simple operation: an organization could implement a website by using web CMS software (i.e., Wordpress [12], Joomla [13], Drupal [14], Plone [15], etc.) which, in a few steps, allows a website to install a database management system as a backend. The graphical aspect is provided through numerous templates, while content editors take advantage of a simplified method for uploading files and sending commands through a web interface. This revolution in website development has undoubtedly enhanced a webmaster's work, but it has contributed to a shift in the attention from other aspects of a website (graphical presentation, the presence of feeds, videos and changing images on the homepage) away from content. Web design has, in some cases, been reduced in its importance as a roadmap made from specific steps whose final objective is to communicate information.

In our opinion, most of these websites result in appealing entry pages, with poor content since, if dynamically updated, it is reduced to a list of news. Most modern websites appear as web journals collecting the same chunks of information. Each organization has its own website, even if a great many have poorly designed content or show a duplication of information, instead relying on the force of an appealing interface and connection with Web 2.0 social tools. Among these, web feeds are a common method of distributing news [16]. Since, in a web infrastructure, information is distributed and no longer centralized, is a general-purpose website still useful for communicating over the web? Is it perhaps more useful to take advantage of web technologies (i.e. XML, REST [17], Ajax [18], etc.) to create a mashup application that uses web widgets in order to create a homepage that could aggregate and present information coming from different sources? After these considerations, a web project manager could approach a different solution in a restyling of a website and try to aggregate and present content that is already available on the web, taking whatever attention necessary to follow all the steps required for good web design.

Such technologies are normally used in an end-user environment (i.e., the iGoogle approach [19], the user's Google startup page that allows a user's customization of merging different feeds in a single page). This paper describes research we conducted for a project to restyle our website in order to evaluate if an institutional website could be reduced to a homepage made up of mashups and web widgets as interactive graphical elements and that adheres to web standards in terms of accessibility laws.

 ISBN: 978-1-61208-100-7

In our opinion, such as solution could contribute to reducing the duplication of information and making web content relevant again. This paper starts by describing our need for re-designing the actual website, and presents an overview of technologies related to mashups and the ability to collect data and services that should appear in an interoperability manner. Then, the paper discusses a prototype solution that could be adapted to the case study, outlining the issues related to collect data from different sources and publishing them in a user-friendly manner for different categories of users.

II. THE ISTITUTIONAL CONTEXT AND THE STATE OF ART

When, in 2005, we began to develop the new website for the INAF [20] institute (see Fig. 1), we focused our attention on the type of content that we should publish and the necessity of defining a publication workflow that starts from a web editor inserting content and then passes to a redactor and then a reviser.

Figure 1. How actually appears the homepage of the INAF institute

The choice fell [21] on Plone software as a CMS technology, which allows to define, in detail, the types of roles that different users can have and the actions the users can perform. Even though the project objective was satisfied, this website suffers from the fact that it does not truly represent the whole institute. This institute is, in fact, made up of 19 astronomical observatories and other institutes, together with two facilities hosting telescopes. Each of these entities, geographically distributed in Italy, Spain, and the U.S.A (in regard to facilities) has its own website, (i.e., the INAF, Astronomical Observatory of Padua [22]) implemented in a different manner from the others and with its own graphical aspects, etc.

For these reasons, most information is duplicated (present on both the national site and on the local site), making it difficult to recognize the single entities as part of the same institute. Moreover, in the last period, another website called "Media inaf" [23] dedicated to the general public as the Fig. 2 shows, has been developed to distribute information.

Figure 2. The start page of the INAF website dedicated to the public

Because the content is spread among different websites, related only by links or by entirely duplicating data, information updates are not frequent (especially for some sections of the site) and there are duplicate web servers, bandwidth costs and webmasters needed to manage them.

Moreover, this structure does not promote efficiency, since each server relies on its own software and backup scheduling. Considering the amount of information, a single server developed as a web server farm could replace all those sites in order to provide reliability and load balancing.

From this aspect, the required restyling of the website, necessarily implies some consideration of the possibility of using a general-purpose server. Considering that the information for publishing already exists in a distributed way among the local websites, from a technical point of view, modern technologies could help to aggregate and dynamically present the needed information. Mashups are an emerging technology in the Web 2.0: in the words of Wikipedia, a mashup is a web site or web application "that seamlessly combines content from more than one source into an integrated experience". Content comes from different sites and is aggregated into a start page, which could be an innovative starting point for the institute.

The idea is to transform the website into a single homepage that collects and presents information as a result of careful analysis. The visual design project, developed in a second phase, could provide an appealing aspect while still adhering to web standards and web accessibility guidelines with respect to laws and to maintaining the original idea of the web, that of being devoted to all users, especially those with disabilities. The paper approaches the feasibility study in order to verify if it is possible to use existing tools to develop and create such a page as chunks of information visualized in sections that are dynamically updated? Mashups are used in a single homepage and in an end-user environment. Why not introduce them into the official homepage of an institute or an organization? What is the difference between these two technologies in different contexts? Both a generic websites developed with CMS technology and those developed with web mashups focus on web content. Content is essentially structured in a database

 ISBN: 978-1-61208-100-7

format (essentially, relational or object database) or in an XML-like technology to be imported or transformed onto an HTML/XHTML page

III. MASHUP TECHNOLOGIES AND EXPORTED DATA FORMATS

A Mashup [24] is a hybrid web application that integrates data and functionality using web technologies to create new services and can in general run either on the server or on the client side.

The term, as Fig. 3 shows, implies fast integrations, APIs (Application Programming Interfaces), and data sources. Categorization divides a consumer mashup aimed at the general public to provide personalization of data/viewing according to users' needs, data mashups that combine similar types of media from multiple sources and information into a single representation, and business mashups that focus on a single presentation.

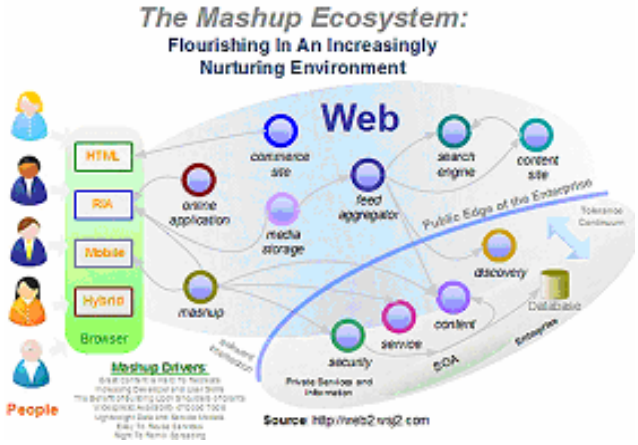


Figure 3. The Mashup ecosystem: involved components and technologies

Architecturally, there are two styles of mashups: web-based mashups that use a web browser to combine and reformat the data and server-based mashups that analyze and reformat the data on a remote server and transmit the data to the user's browser into its final form. For server side development, some technologies enable retrieving and processing data dynamically (i.e., PHP [25], Ruby [26] programming languages, etc.) and technologies implemented or supported in web browsers like Ajax, RSS, Adobe Flash [27], Microsoft Silverlight [28], etc.).

Both mashups and web applications generally use a browser as a client-side solution component that provides a user interface (UI) to a given application. They are centrally managed from a server in the enterprise, and both are deployed to the user's browser when the URL is used to access a web resource on the server. Business logic can be executed on systems, and data can be retrieved from databases, public data sources, or external services. Code that defines and makes up the UI and information extracted from internal and external systems is rendered within the browser, providing users with the means to view and act on the data. The content sent to the browser can include Javascript and scripting libraries that execute in the browser's runtime to allow custom logic to be run locally in the browser. A mashup can appear to be a variation of a Façade pattern, that is a design pattern that provides a simplified interface to a larger aggregate of different feeds with different APIs. A mashup can be used with software provided as a service (SaaS) [29]. Publishing and syndication formats require special attention since they are the glue that links mashed sites together [30].

Mashup selection can be performed dynamically identifying the situational set of widgets available at runtime. A widget is a software component used a by mashup assembler that encapsulates disparate data, user interface markups, and behavior into a single component that generates HTML code fragments. They are implemented as chunks of code that can be easily embedded in pages. Widgets are applications developed through RSS or Rich Internet Applications (RIA) [31] as means of combining web technologies (HTML/XHTML [32], CSS [32], ECMAscript [33], etc.).

Widgets are available on webtops (with drag and drop buttons) and by inserting source code or HTML pages. Mashups and widgets are available to end-users because most web 2.0 tools release example coding (usually as script in JavaScript programming) that can be integrated into a website. The technology takes advantages from structuring content in an XML-like format. Re-using and combining information and services is possible with a specific format (i.e., RSS or Atom format) using APIs, or web services developed in REST-style technology that have are based on HTML protocols. Even if the final goal is information publishing and thus HTML/XHTML and CSS language is used to visualize on a web browser such content, if information is exported in XML, transformation into whatever format is possible thanks to all XML technologies. In fact, most mashups include news feed available in RSS [8] or ATOM [9] format (which uses a markup language like XML) or Google maps, which make use of KML[11] format to structure information. The most common data formats that are used for exporting data are RSS and ATOM, even if XML languages or Google related formats are also used.

A. *Interoperability file formats*

JSON [10] is a lightweight data-interchange format based on a subset of standard ECMA-262, originally specified in RFC 4627 with an official Internet media type (application/json) and used for serializing and transmitting structured data over a network connection. JSON is primary used to transmit data between a server and a web application serving as an alternative to XML. JSON is supported by most browsers (i.e., Mozilla Firefox, Microsoft Internet Explorer, Opera and Webkit-based browsers such as Google Chrome, Apple Safari, i.e., browsers based on the webkit layout engine designed to allows web browsers to render web pages), and it is part of most popular Javascript libraries [34] like Prototype, jQuery, Dojo Toolkit, Mootols, Yahoo! UI library. XML is family of markup languages used to describe structured data and serialize objects with the primary goal for data interchange. When data are

 ISBN: 978-1-61208-100-7

encoded in XML, the result is typically larger in size An examples used in such a context is KML (Keyhole Markup language) [11], an XML-based languages used to work with geographic data such as those used by Google Maps. KML uses a tag-based structure with nested elements and attributes (and an extension .kml o .kmz as zipped format). Finally, RSS and ATOM are the most frequently used as web feed formats to publish content as blog entries, news headlines, audio, and video. ATOM is specifically an XML language used for web feeds and it has a related protocol (the Atom Publishing protocol AtomPub or APP [35]) that is a simple HTTP-based protocol for creating and updating web resources. This format was published as an IETF proposed standard in RFC 4287, while the protocol in RFC 5023.

B. *User web mashups*

The primary application of mashups is in the area of end-users, since these activities involve little programming if a user chooses an already-publicized service. A webtop such as iGoogle is an example of a mashup that helps user to individually configure a start page with specific information. Users can propose favorite preferences and compose various widgets in a mashboard. This result in a new sort of interactive web applications whose effect gives new values to the end user. Web mashups use either public APIs from existing sites and/or web scraping to collect data and combine them on a single web page. Examples are mashup web applications using Twitter web services and Google maps services to display locations mentioned. Several examples exist in literature by using different programming languages [36] for data extraction on the server side, even if an alternative approach is to use visual tools [37].

Any public web applications that allows users to enter account information for other web sites needs to take suitable precautions safeguarding this information, and this is difficult dealing with thirdy-party web portals since all information should be stored in plain-text (i.e, Google and Twitter credentials). In this case, it is necessary to adopt standards for open authentication APIs like OAuth [38] a protocol that allows single sign-on solutions for authentication on web sites.

The literature and websites report many examples of how to create a mashup, but all of them concern the inclusion of so-called web 2.0 tools (i.e., Facebook profiles, Twitter connections, and so on) by using different visual mashup editors. Unfortunately, most initially available have been dismissed (i.e., Microsoft Popfly or Google Mashup editor that is migrated to Google App Engine [30]). Yahoo pipes [40] remains one of the tools useful for visually merging content from different sources and in some way “free” even if it requires a Yahoo authentication. IBM QEDWiki has graduated to IBM Lotus Mashup [41] as a commercial tool.

IV. MASHUPS FOR AN INSTITUTIONAL WEBSITE

The use of mashup to address institutional website requires available data sources. The great obstacle to this solution, whose resolution could free up many activities in the actual web management and could contribute to reducing the duplicating of information, is that each website needs to publish information in a format that can be easily collected. Actually, APIs available to include photos on Flickr [42] or sales item in eBay [43] refer to commercial sites and need a sign-up and thus specific credentials. In fact, usually mashup-maker tools use some programming languages for data extraction from an external data source on the server side and combine data together. Most of data that will be combined should be available in an XML-like format (i.e., RSS/Atom feed publishing and syndication formats that are used for exporting data) to be retrieved and processed dynamically. The problems are both how to export information to single sites in this way and how to collect them in order to design the homepage of a national institute. Until now, all mashup technologies have been devoted to end-users, since the aim is to create a user startup page suited to each user in a manner similar to portals. However, its use for an institutional website should be the collection of specific content, available as widgets, in order to form a single page

This represents a great constraint: most sites export data in the RSS feed format, but only for limited information, such as news content. Creating a dynamic start page for an institute involves different types of content being combined with static information (i.e., information about the main features of the institute and its organization, which are not constantly updated). However, using such software requires an availability of exported data and is difficult when integrating static information. The final start page should be a composite aggregation of dynamic and static information devoted to all kinds of people that use such a site. And content should be differentiated if the user is an internal employee, a professional astronomer, or simply a person who likes astronomy.

The mashup for this goal is necessarily server-based that combines sources from different websites. Simple techniques use cutting and pasting snippets of Javascript, using RSS feeds and XML to connect the various parts together and even on-line js inclusion that can pull in and integrate a powerful external component.

A first idea on the schematic layout of the homepage is presented in Fig. 4. This first prototype includes some external APIs from major thirdy-party sources (i.e., Google Maps, Flickr, Facebook, etc.) and several data from different RSS/ATOM feeds. This solution could be easily implemented since all single local sites export al least some information as RSS feed, especially those regarding all news (internal and external) about local activities and projects.

If such a technology could be easily used to merge content available on different sites, there should be savings in hardware, software and human resources.

Moreover, mashup technology, even if born around the single user and his or her ability to create his or her own

 ISBN: 978-1-61208-100-7

content, functionalities, and services could be adapted to reach a group of users (i.e., a different category of users to whom a website is devoted).

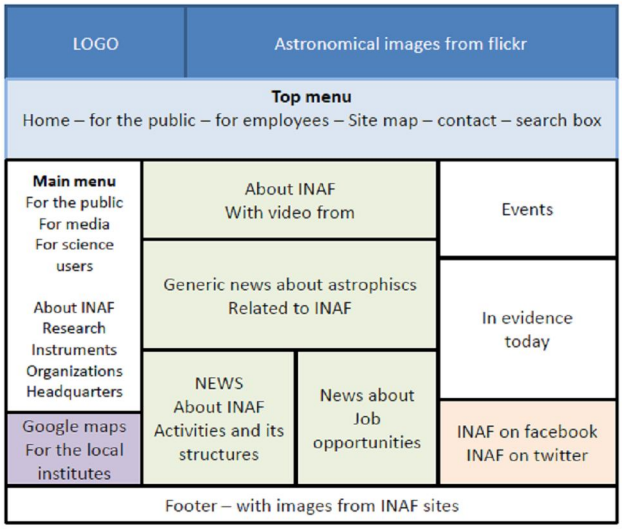


Figure 4. A draft of a webpage layout

A mashup could be a way to extend the concept of using Web 2.0 tools in an enterprise context [44] being a tool to create a innovative website.

A. *Combining user partecipation in website and the need of controlling the published information*

In our opinion, Web 2.0 is a way to allow end-users to insert their own content and functionalities in a website that could became a work tool that uses of a set of web technologies related to the fact the content should in some way imported and published in a customized manner. Close to the concerns of exporting data and services through the network, there is the problem of allowing different users to choose data and services to be published and at the same time avoiding transferring full control of the information to the end-users. We have outlined several categories of users who will be involved with our website. It is logical that internal users require information different from external users (science users or the media for example). Details of users is shown in Fig. 5.

The concern is to export data as feeds or in other XM-like format for all relevant information (i.e., customized different news) and to create some web services able to include other types of content in the page such the multimedia content that is non-necessary stored on a Web 2.0 websites such as YouTube or Flickr.

This is realized with a registration by e-mail in the iGoogle approach or by including a special key to be used in the exported API. However, this approach seems to be too user-customizable in an institutional website, and a tradeoff should be found to mix these requirements. The state of art of mashup technologies and literatures do not seem to face with this problem probably because the main aim of the instrument devoted to end-user.

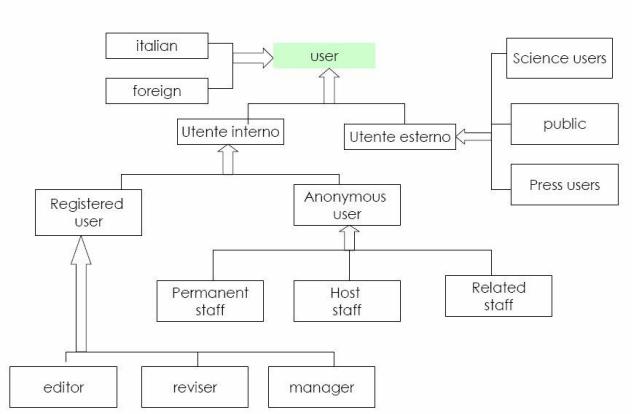


Figure 5. Users' categorization according an analysis of thw website's visitors

But, we think that this solution will provide many savings from technical and content viewpoint, and thus this project will be the aim of next months. Probably the main concern will be to have all main content structured in an XML-like format in order to apply XML technologies [45] (i.e., XSLT) helping in processing files and trasforming them in a format useful to be published on a web page.

V. CONCLUSIONS

This work is a preliminary discussion of the possibility of proposing a mashup as a substitution for a generic website. The need to avoid information duplication and for a website that is not used by its employees has lead to an analysis of the possibility of applying a different approach in the design of a website. Mashups and widgets could possibly be applied in such a solution, but it is necessary to make a further study in analyzing how to actually implement such a technology and its tools. One such solution is the use of mashup-makers for composing personal homepages, but could this tool also be applicable for a general homepage, such as that of a research institute? And how it is possible to define from which part of the homepage to collect information? For example, an important step is to understand how content should be exported in order to be collected (in an XML format, such as an RSS or Atom feed) and if a local website could contribute by implementing a web service interface through an API. All these requirements will be taken into consideration by implementing specific tools able to create a composite startpage. However, the first prototype that includes some features peculiarly to a web mashup has been developed. It collects the web feeds available from different website in order to compose a list of information about news in the astrophysical area and content about the institute's activities. Moreover some web services exported by famous commercial sites (i.e., Google, Facebook, and so on) could easily be embedded in the page. A Google map could help locate several institutes, and Google Docs could be useful for

 ISBN: 978-1-61208-100-7

managing administrative documents before and after the publication on the website. A problem here is that information is strictly related to this site (i.e., there is the need for a specific Google key in order to import such a service) and related to business activities.

REFERENCES

[1] D. Addey, J. Ellis, P.Suh, and D. Thiemecke, Content Management Systems (Tool of the Trade), Apress, 2003.

[2] T. O'Reilly, What Is Web 2.0: Design Patterns and Business Models for th next generation of software, O'Reilly Press, 2005.

[3] YouTube site, http://www.youtube.com, 10.07.2010

[4] Facebook site, http://www.facebook.com, 10.07.2010

[5] Twitter site, http://twitter.com, 10.07.2010

[6] D. Tapscott, Enterprise 2.0, how social software will change the future of work, Gower Publishing, 2008.

[7] XML (eXtensible Markup Language) Activity Statement, http://www.w3.org/XML/Activity, 10.07.2010.

[8] RSS (Really Simple Syndication) board, http://www.rssboard.org, 09.07.2010.

[9] The Atom Syndication Format,.http://tools.ietf.org/html/rfc4287. 09.07.2010.

[10] JavaScript Object Notation (JSON), http://www.json.org, 09.07.2010.

[11] KML (keyhole markup language Tutorial http://code.google.com/apis/kml/documentation/kml_tut.html, 09.07.2010.

[12] Wordpress site, http://wordpress.org, 10.07.2010.

[13] Joomla!, http://www.joomla.org, 10.07.2010.

[14] Drupal, http://drupal.org, 10.07.2010.

[15] Plone CMS, http://plone.org, 10.07.2010.

[16] R. Jee, Pro Web 2.0 Mashups, remixing data and web services, Apress, 2008.

[17] R.L. Costello, Building web services the REST Way, http://www.xfront.com/REST-Web-Services.html, 10.07.2010

[18] Ajax and other "rich" interface technologies, http://www.owasp.org/index.php/Ajax_and_Other_%22Rich%22_Interface_Technologies, 10.07.2010.

[19] N. Conner, Google Apps: The missing manual, Pougue Press, 2008.

[20] Te National Institute of Astrophysics (INAF) website, http://www.inaf.it, 09.07.2010

[21] C. Boccato and S. Pastore, "The Web Information System of the National Institute for Astrophysics: different actors contributing to disseminate information". In Current Research in Information Sciences and Technologies, Multidisciplinary approaches to global information system, Proc. Of Int. Conf. on Multidisciplinary Information Sciences and Technologies (INSCIT2006), Merida, Spain, October, 2006, Vol. I, pp. 507-511.

[22] INAF OaPD, http://www.oapd.inaf.it, 09.07.2010.

[23] Media INAF website, http://media.inaf.it, 09.07.2010.

[24] D. Merrill, Mashups: the new breed of web app. An Introdution to mashups, IBM developerworks, 2009.

[25] PHP: Hypertext Preprocessor, http://php.net, 09.07.2010

[26] Ruby progamming language, http://ruby-lang.org, 09.07.2010

[27] Abobe Flash Platform Technologies, http:// http://labs.adobe.com/technologies/flash/, 10.07.2010.

[28] Microsoft Silverlight, http://www.silverlight.net, 10.07.2010

[29] Tim O'Reilly, Open Source Paradigm Shift, O'Reilly Media, http://tim.oreilly.com/articles/paradigmshift_0504.html, 13.07.2010.

[30] M. Watson, Scripting Intelligence: Web 3.0 Information Gathering and Processing, Apress, 2009, pp.269

[31] Putting the power of Web 2.0 into practice. How rich Internet Applications (RIA) can deliver tangible business benefits, IBM white paper, 2008.

[32] HTML &CSS, http://www.w3.org/standards/webdesign/htmlcss, 09.07.2010.

[33] ECMAScript programming language, http://www.ecmascript.org/, 09.07.2010.

[34] Javascript librieries, http://javascriptlibraries.com 09.10.2010.

[35] The Atom Publishing Protocol, http://www.ietf.org/rfc/rfc5023.txt, 09.07.2010.

[36] O. Beletski, "End User Mashup Programming Environments", Helsinky University of Technology, Telecommunications Software and Multimedia Laboratory, Seminar on Multimedia, 2008.

[37] A. Spoerri, "Visual Mashup of Text and Media Search results, in Proceedings of the 11th International Conference Information Visualization," IEEE Computer Society, 2007, pp. 125-134.

[38] OAuth Community site, http://oauth.net, 09.07.2010.

[39] Google App Engine, http://code.google.com/appengine, 10.07.2010.

[40] Yahoo pipes, http://pipes.yahoo.com/pipes, 09.07.2010

[41] IBM Mashup Center, http://www-01.ibm.com/software/info/mashup-center/, 10.07.2010.

[42] Flickr Photo Sharing, http://www.flickr.com, 10.07.2010.

[43] eBay, http://shop.ebaby.com, 10.07.2010.

[44] A. McAfee, Enterprise 2.0: new collaborative tools for your organization's toughest challenges, Harvard Business School Press, 2009.

[45] XML Technology, http://www.w3.org/standards/xml/, 13.07.2010

.

 ISBN: 978-1-61208-100-7

Building the Web of Things with WS-BPEL and Visual Tags

Web of Things using Service-oriented Architecture standards

Antonio Pintus, Davide Carboni, Andrea Piras
ICT-Information Society
CRS4
Pula, Sardinia, Italy
e-mail: pintux@crs4.it, dcarboni@crs4.it, piras@crs4.it

Alessandro Giordano
Dip. di Ingegneria Elettrica ed Elettronica
Università degli Studi di Cagliari
Cagliari, Sardinia, Italy
e-mail: alegiordy@tiscali.it

***Abstract*— The Web of things is an emerging scenario in which everyday objects are connected to the Internet and can answer to HTTP queries with structured data. This paper presents a system that allows users to build networks of everyday objects using visual tags as proximity technology. The system backend is based on Service-oriented Architecture languages and tools for the runtime composition of "things" establishing connections we call hyperpipes.**

Keywords - Ubiquitous computing, Web services, Web of Things, SOA

I. INTRODUCTION

The Web of things is an emerging scenario in which every object is connected to a pervasive wireless/wired network and can answer to a HTTP query with structured data. Everyday surrounding objects like phones, domestic appliances, advertisement billboards, musical instruments become the nodes of the Web of things. Nevertheless, merely putting real objects into the network is nothing without a logic that creates a net value. One key is to compose objects together [10] and to put the orchestration in the hands of the final user. Simple mechanisms to connect "things" can foster a huge number of unpredictable applications. Towards these objectives, users, objects and networks are the ingredients to build a Web of things in which users become seamlessly the "programmers" [9]. The Web of things has the potential to become the next killer application and it must seamlessly emerge from existing Web infrastructure taking to the limits all the Web related technologies and providing new use cases that will improve the definition and the adoption of new standards and protocols.

The Web is basically built on two metaphors: the hypertext and the hyperlink. The former is just a digital reification of human language in its written form, while the second is a mechanism that non-digital forms of writing (like writing on paper) could not provide. The two metaphors are on the basis of the Web of pages while in a Service-oriented Architectures (SOA) not only pages are linked together but are also linked with information services. We want to extend the pages and services interlink from digital objects to real ones.

This paper, after an analysis of some related works and an introductive classification of "things" based on their capabilities, presents a "things" composition framework called hyperpipes. Finally, a prototype and conclusions along with indications for future work are provided.

II. RELATED WORKS

This work is located in the main stream of ubiquitous computing, and more precisely in a subset of the "Internet of Things" based on Web protocols instead of ad-hoc and special purpose transport and application protocols.

We think that in the Web of things all kind of services (WS-* and REST [5]) provided by objects must be orchestrated together but for practical reasons in this work we sketch a design solely based on Web services Description Language (WSDL) and Simple Object Access Protocol (SOAP) Web services. The use of SOA Web standards for the Internet of things is not new: the SODA project [4] goes toward the definition of an architecture where devices are viewed as services in order to integrate a wide range of physical devices into distributed IT enterprise systems. A SOA approach for embedded networks is also persuaded by other projects, such as SIRENA [8] and SOCRADES [13]. Our work distinguishes from the others above because we experiment the direct generation of new process definitions according to user selection and pointing of real objects in the environment. We postulate that physical objects must expose in a formal specification the set of operations they can perform and the data they can exchange with a precise contract in a way that they can be composed and orchestrated using existing standard languages like the Web services Business Process Execution Language (WS-BPEL) [1]. This assumption makes the choice of SOAP and WSDL 1.1 of practical use for our actual implementation. In principle, the inclusion of RESTful services in orchestration is possible with the support of WSDL 2.0 but in practice this standard cannot be effectively adopted yet. In the meanwhile RESTful services could be proxied by ad-hoc SOAP services and orchestrated in WS-BPEL but in this work we do not address this issue.

The proximity of users to objects is another fundamental aspect that must be considered in pervasive computing. One of the peculiar points of our work is that process definitions for object pipelining are created by users on demand. In [11] the authors use Bluetooth as option for providing connectivity, and propose RFID technology to enhance the Bluetooth connection establishment procedure.

 ISBN: 978-1-61208-100-7

Our approach to proximity is that after an object and in a given situation and with a given mood a person can have the idea to build something new. The subsequent action in our scenario is to build a connection. If we imagine the world as a giant sketch board we just want a way to draw a line from an object to another and build something useful as result. In [3] a similar interaction pattern is depicted but the system architecture and the data formats are described at a general level while in this work we focus on architectural aspects with formal specification and adoption of SOA standards. Simple but effective rules, applied to a multitude of objects tend to form a complex system [7]. In our scenario millions of real objects can simply be connected through hyperpipes with natural gestures in the real environment without sitting in front of a PC screen.

III. ASSUMPTIONS ON OBJECTS CAPABILITIES

One of the main assumptions in a Web of things is that objects can communicate at HTTP level and above. This assumption is a weak one because technical solutions to achieve this result are already discussed and designed in literature [2], [12], [14] and some projects are ongoing. Thus, if the connectivity is lacking in the real world this is due to a lack of infrastructure and not to a lack of know-how.

Nevertheless, it is useful to sort "things" according to the level in the communication stack they can be connected:

- at the top of this sorting we have bare virtual goods and services like Web sites, e-mail boxes and 3D models, just to mention some. These objects can be easily wrapped and then referenced in a HTTP addressing space like resources (REST) or like services (WSDL).
- At a second level we find appliances with a complete HTTP stack like wireless printers or networked screens.
- In a upper intermediate level we find objects that are not equipped with a complete HTTP stack but can still communicate at a TCP/IP or UDP/IP level. For those objects is straightforward to build a HTTP wrapper.
- In a lower intermediate level we find objects that cannot communicate over IP networks, but still can communicate with different protocols like ZigBee, Bluetooth or X10. For those objects a proxy can be deployed to present these objects in the HTTP addressing space.
- Finally, there are bare physical objects. For those a digital counterpart must be built and published online. For example, a real book has a virtual counterpart like a Web page in a online bookstore.

Let us consider any object (physical or digital) like a process exposing a set of operations. We classify the operations according to their ability to produce data (sources), process data (processors), and consume data (sinks). This classification is useful to distinguish between sensors, actuators and processors. Operations are public, thus their names are globally known.

IV. HYPERPIPE FRAMEWORK BASED ON SOA LANGUAGES

Considerable challenges are related to connecting a large set of information sources and sinks together. When only existing protocols and data formats are used, the communicating parties must be matched based on the descriptions describing their capabilities. To get a balance among the generality of purposes and the need to implement a system really able to work, we made some choices that drive the design of our work. First, we choose to adopt WSDL as the formalism to describe what an object is able to provide. In this way, an object can be considered as a SOAP Web service. Another issue is related to the type of communications between objects and the definition or the adoption of a suitable related protocol. In our design objects are allowed to exchange data without strict type checking (automatic type adaptation is a feature), and communication may be either synchronous or asynchronous. Both these interactions can be easily modeled and implemented using WSDL and adopting SOAP over HTTP protocol for messages exchange. Another type of logical connection to include in the design is multimedia streaming between objects. For instance, the user selects a MPEG camera source and a wall screen as sink. Embedding multimedia streaming in SOAP messages is not an efficient implementation, thus another protocol should be used instead. In concrete, the main assumption we have to make is that, in order to actively play a role in a pipe, an object must be able to connect to the network and to run a Web service stack. This general capability can be accomplished basically in two ways: the object itself is powerful enough to satisfy the previous requirements or it has to be connected and "driven" by a proxy computer, which satisfies the requirements. We choose WS-BPEL for concrete representation of pipes. WS-BPEL is an XML-based language born to define executable business processes as orchestration of Web services. WS-BPEL orchestrations expose a service interface described using WSDL: in this way, from the point of view of a client, WS-BPEL process is a Web service itself. Expressing pipes using WS-BPEL brings two main benefits to our vision: first, it is possible to associate a well-defined functional interface to each pipe, in our case modeling that in order to expose Video Cassette Recording (VCR)-like functionalities: start, pause and stop, which are the public available operations for a generic pipe control.

Three basic patterns of "in-Pipe" communication emerge:

a) synchronous, on an object A is invoked an operation src, the result is adapted and then passed as argument to an operation snk of an object B;

b) asynchronous, the pipe registers itself as a listener for an event produced by an operation src on object A. When the event is fired, the data attached to the event is adapted and the sent to the sink.

c) streaming, an object B receives from an object A a stream of data (for instance video mpeg from a camera to a screen). Given that binary real time data encapsulation inside SOAP messages is not an efficient implementation, rather Real-time Transport Protocol (RTP) or equivalent

 ISBN: 978-1-61208-100-7

real time media transmission protocols should be used instead, using the SOAP messaging only to initialize the session for handshaking.

For the first two patterns, we created two different WS-BPEL document templates, which define all the required activities, message exchange and service orchestration for the execution of each of them in a WS-BPEL engine. In particular, pattern (a) is a typical Web service orchestration scenario with a subsequent invocation of services; pattern (b) basically is an orchestration in which the WS-BPEL document describes an asynchronous invocation of a service (the event producer) using a callback mechanism, which invocation triggers an event causing a delivering of the event to the other service (the event listener). The pattern (c) uses WS-BPEL only for protocol negotiation and handshaking between the two services, in this way, after these steps, the objects can instantiate streaming sessions in an independent way using the suitable chosen protocol. From a WS-BPEL point of view the pattern (c) is equivalent to pattern (a) but the data exchanged are Session Description Protocols (SDP) instances and the work of establishing a streaming is completely delegated to endpoints.

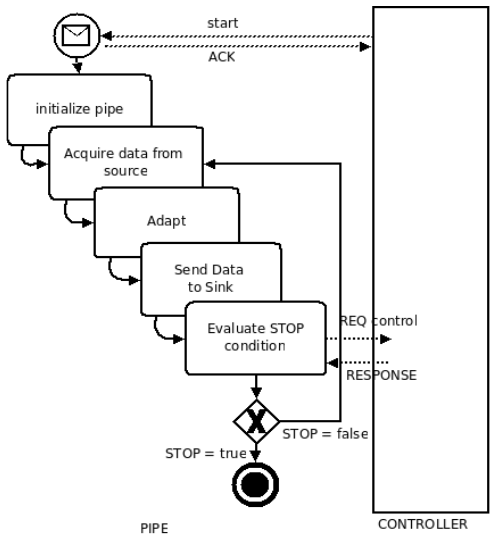


Figure 1. BPMN-like notation for a synchronous pipe from a source to a sink operation.

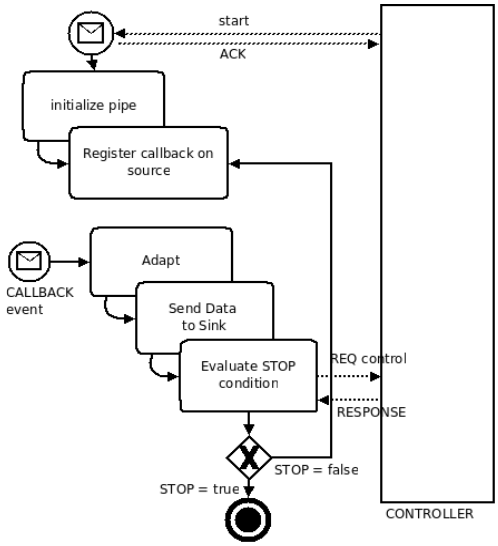


Figure 2. BPMN-like notation of an asynchronous pipe between a source and a sink. The callback endpoint is invoked when data can be consumed.

The difference between the (a) and the (b) template is that in the second the data source asynchronously emits a data and requires that a callback endpoint is registered in order to consume data when data are ready. The different design is depicted in Figs. 1 and 2 where a BPMN [15]-like notation is used instead of showing the XML code, which results too verbose to fit the limit of this article.

V. PROTOTYPE

The objective of our prototype is to show a living system that allows users to select real objects in a room and to compose them building a real time orchestration starting from the user interaction.

2D barcodes systems like Datamatrix and QR are attached with no cost to any object in order to "augment" their features realizing a virtual connection with a one its digital pair. Appropriate programs can recognize codes and download linked information from the Internet. To implement our point-click-and-compose interactive paradigm, we adopt QR barcodes so the user can point an object and retrieve what that object is able to perform. Given the verbosity even of a simple WSDL document, we choose to encode in the barcode only a URL to reference it. The user points a smart phone against the barcode, then the phone decodes the visual tag and asks an online server to parse the WSDL document to obtain the list of operations. Selecting two different actions from two different objects (or even from the same object) a pipe can be constructed. The WS-BPEL templates are filled with real endpoints, deployed on the WS-BPEL engine and then activated. To implement a prototype we needed some "things" to become endpoints of pipes. Thus, we instrumented normal objects with some SOAP messaging abilities deploying personal computers and notebook to simulate sources and sinks.

VI. FUTURE WORKS AND CONCLUSION

Capabilities of objects are well expressed with WSDL and translated into human readable lists of actions in the phone user interface. The main advantage is the ability to generate WS-BPEL at runtime and to create new executable processes (the hyperpipes) with the point-select-and-compose interaction.

The overall design results well conceived for the transmission of "data as documents" between different objects while data streaming is less supported by the Web services stack and SOAP is only used for exchanging session descriptions and that commuting to other protocols in the communication stack. The choice to model objects like opaque components able to perform operations poses some issues in the seamless connection with other Web resources. It is clumsy to make a pipe having as endpoint a Web page or a RSS feed because even if these are digital objects, they

 ISBN: 978-1-61208-100-7

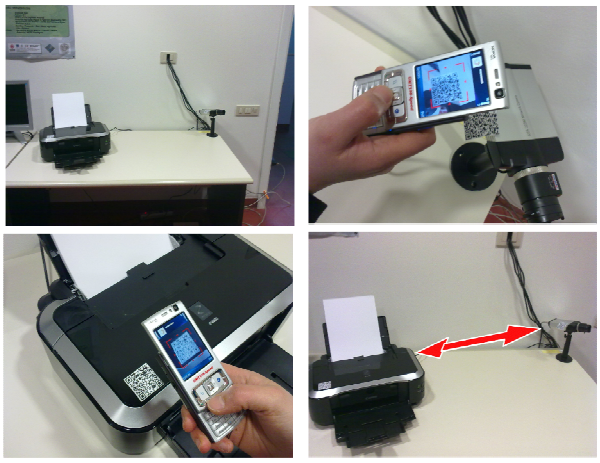


Figure 3. Selecting two different actions from two different objects (or even from the same object) a pipe can be constructed.

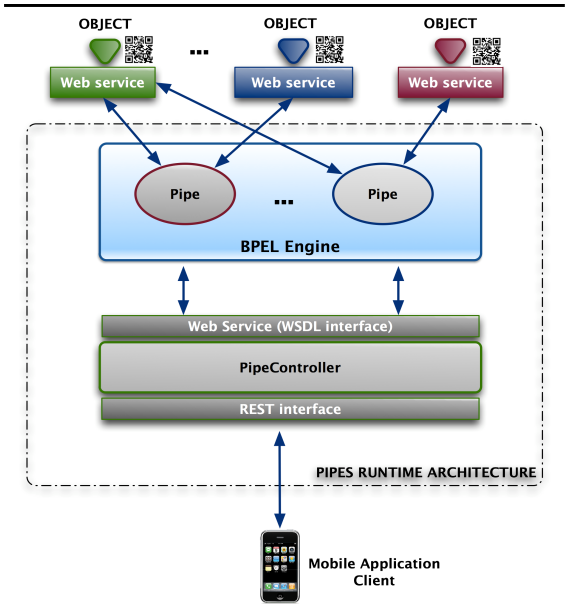


Figure 4. Hyperpipes at runtime. Objects are instrumented by Web services and Pipes are implemented as WS-BPEL processes created and deployed at runtime. A mobile phone is used to point objects in the environment and to retrieve the WSDL specification.

need to be wrapped by a WSDL interface and decorated with a service implementation. The resource-oriented nature of the Web is somehow in contrast with the procedure-oriented architecture of Web services. Some authors [6] consider RESTful the only choice for a Web of things architecture and advocate some motivations related to the programmatic complexity of Web services and according to their experience, not well suited for end-user to create ad-hoc applications. Among the motivations is that discovery of Web services via Universal Description Discovery and Integration (UDDI) is not suitable for sensors or devices because the UDDI-based discovery has not context information (e.g., where a sensor is placed). In our work the problem of discovery is simply by-passed by the fact that services are discovered by users when they are close to an object using some proximity technology (the QR tags in our work) and the programmatic complexity is totally hidden to end-users by the automatic generation of WS-BPEL executables. As mentioned before, we think that orchestrations should include the largest set of element types, both real and virtual, and represented by either stateful processes (WS-*) or stateless resources (REST). One next achievement is to build such a universal orchestration starting from user interactions in the environment.

Regarding the user interaction we conclude that building a pipe between two objects results as a straightforward task. Composing multiple pipes with processor in cascade is somehow less intuitive and requires the user to know how the underlying process is created. The use of QR has revealed to be a practical choice very easy to implement and quite easy for users to manage. Nevertheless, the conclusions on human interaction here reported are merely qualitative and based on the experience of few test users. We plan to make a more accurate usability evaluation in a future work.

REFERENCES

[1] OASIS Web Services Business Process Execution Language (WS-BPEL) TC. http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=wsbpel. Last accessed on 01-06-2010

[2] QuadraSpace. http://www.quadraspace.org/. Last accessed on 01-06-2010

[3] Carboni, D. and Zanarini, P. Wireless wires: let the user build the ubiquitous computer. Proceedings of the 6th international conference on Mobile and ubiquitous multimedia, (2007), 169–175.

[4] Deugd, S.D., Carroll, R., Kelly, K., Millett, B., and Ricker, J. SODA: Service Oriented Device Architecture. IEEE Pervasive Computing 5, 2006, 94-96, c3.

[5] Fielding, R.T. Architectural Styles and the Design of Network-based Software Architectures. University of California, Irvine, 2000.

[6] Guinard, D., Trifa, V., Pham, T., and Liechti, O. Towards physical mashups in the web of things. Proceedings of INSS, (2009).

[7] Holland, J.H. Emergence: from chaos to order. Addison-Wesley Longman Publishing Co., Inc., 1998.

[8] Jammes, F. and Smit, H. Service-oriented paradigms in industrial automation. IEEE Transactions on Industrial Informatics 1, 1 (2005), 62–70.

[9] Kindberg, T., Barton, J., Morgan, J., et al. People, places, things: web presence for the real world. Mob. Netw. Appl. 7, 5 (2002), 365-376.

[10] Ragget, D. The Web of Things: Extending the Web into the Real World. SOFSEM 2010: Theory and Practice of Computer Science. Springer Berlin / Heidelberg, (2010), 96-107.

[11] Salminen, T., Hosio, S., and Riekki, J. Enhancing Bluetooth Connectivity with RFID. Proceedings of the Fourth Annual IEEE International Conference on Pervasive Computing and Communications, IEEE Computer Society (2006), 36-41.

[12] Sommer, S., Scholz, A., Buckl, C., et al. Towards the Internet of Things: Integration of Web Services and Field Level Devices. .

[13] de Souza, L.M., Spiess, P., Guinard, D., Kohler, M., Karnouskos, S., and Savio, D. Socrades: A web service based shop floor integration infrastructure. Lecture Notes in Computer Science 4952, (2008), 50.

[14] Trifa, V., Wieland, S., Guinard, D., and Bohnert, T. Design and implementation of a gateway for web-based interaction and management of embedded devices. Submitted to DCOSS, (2009).

[15] White, S.A. Introduction to BPMN. IBM Cooperation, (2004), 2008–029.

 ISBN: 978-1-61208-100-7

Integrated E-Learning Web Services

Alina Andreica, Florina Covaci, Daniel Stuparu, Árpád Imre, Gabriel Pop
IT Department and Chemistry Department
Babes-Bolyai University
Cluj-Napoca, Romania
{alina, florina.covaci, dstuparu}@staff.ubbcluj.ro, aimre@chem.ubbcluj.ro, gabrielp@staff.ubbcluj.ro

***Abstract*— The present paper focuses on means of creating integrated web e-learning services by providing learning and dedicated information systems facilities within a web portal. The information system facilities are obtained by integrating into a global web portal the dedicated services and synchronizing databases based on various technologies (php / postgresql, asp / MS sql). The portal is based on MS technology and provides, as learning services, management content and e-learning facilities for various user categories, together with dedicated information system facilities.**

Keywords: web services; system integration; database synchronization; e-learning services.

I. INTRODUCTION AND WORKING FRAMEWORK

In the framework of the knowledge based society, information technologies strongly impact on the learning processes [16] and organizational management by means of dedicated information systems. Information system integration has been tackled in the literature especially for business and organizational processes [11]. System interoperability has also been dealt from a semantic point of view [12] .

This paper presents a framework for integrating an e-learning system with dedicated information systems for managing organizational processes for a higher education institution. The proposed framework enables data and service integration that may be further exchanged within federated web services [10], [19].

The paper aims at describing a system integration framework for providing web services into a global portal, including web-based facilities offered by dedicated information systems. The portal we describe has mainly learning purposes (but they may be adapted to various information sharing & communication needs), is based on Microsoft - MS technology and provides means of integrating various information systems, using different technologies (php / postgresql, asp / MS sql). In this respect, we describe an integrated architecture using a ILM - Identity Lifecycle Management [25] server, and additional interface modules, used in order to integrate the dedicated information systems into a web portal that also provides e-learning facilities, based on SharePoint Portal functionalities[*]. Based on the principles regarding the way in which data types are structured in organizations' databases and used by different components of integrated software systems, ILM can synchronize and optimize data access and delegate processing means to the appropriate dedicated software components.

Section 2 describes integration principles that we have designed and are in train of being implemented. In Section 3, we present the web services that are provided within the portal: learning facilities, virtual labs and dedicated information facilities, while Section 4 focuses on the web portal and its assessment from the administrating and maintenance points of view, in order to prove its flexibility and adaptability advantages.

II. SYSTEM INTEGRATION ARCHITECTURE

The web-services that we deal with have mainly educational purposes and are made available within the integrated e-learning portal that we are in train of implementing. In order to ensure the integration of our e-learning portal with the dedicated information systems (AcademicInfo, ManageAsist, Research Management System – see 3), we have designed an advanced system integration framework that we further describe.

Integration principles are based on an integrated authentication solution, which maps facilities from the dedicated information systems into the portal, for each user category [6]. The authentication server associates, to each user group, the facilities that correspond to their permissions in each of the dedicated information systems AcademicInfo, ManageAsist, Research Management System, in order to make them available within the portal – see figure 1.

[*] The present work is supported by the EU funded grant, within the European Fund for Regional Development, "CCE 124/323/31.08.2009 SMIS 4424 - Sistem electronic aplicativ integrat de educaţie al Universităţii Babes-Bolyai" – Integrated applied electronic system for education of Babes-Bolyai University - BBU, contracted by BBU with the Romanian Ministry of Communication and Information Society, Organismul Intermediar pentru Promovarea Societăţii Informaţionale (the Intermediary Structure for Promoting the Information Society), during 31-08-2009 – 31-08-2011

 ISBN: 978-1-61208-100-7

The MS architecture managed by an ILM type server is used in order to ensure single sign-on capabilities and uniform interface to the dedicated information systems. In this respect, we are in train of designing interface modules in order to map the portal authentication into each dedicated system.

A global synchronized database is in train of being created, the most important common information being the human resource & organization chart ones, retained in the dedicated tables [6]:

User[userid, account, password, unitid]

Unit[unitid, unitname, …]

Organization_chart[unitid, superior_id, horiz_id]

This common database, used by the ILM server, will also contain user and group authentication information, together with dedicated permissions in each of the information systems, in order to ensure access to corresponding permissions in AcademicInfo, ManageAsist, Research Management Systems.

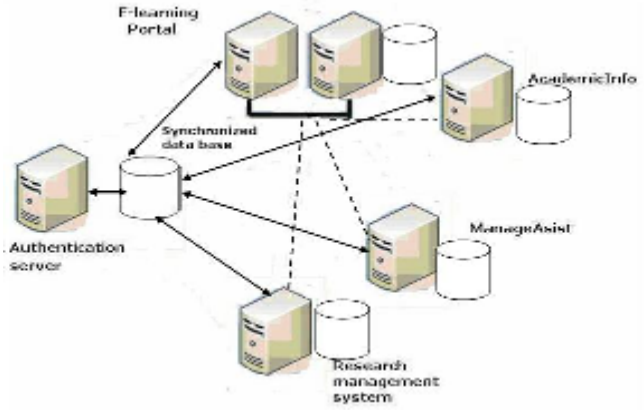


Figure 1: Framework for advanced system integration

The integration solution is also designed to ensure database synchronization among AcademicInfo, ManageAsist, Research Management System and Portal databases based on matching the following data [6]:

- ◊ portal – AcademicInfo: users (all categories), curricula, study contracts, grades, fees
- ◊ portal – AcademicInfo – ManageAsist: organization chart, human resources, managers, financial information
- ◊ portal – AcademicInfo – Research Management System: research activities, PhD Students
- ◊ portal – ManageAsist – Research Management System: organization chart, units, human resources, grants & corresponding financial information

A major issue in the implementation of the global database synchronization is that the applications use 2 database management systems (PostgreSQL for ManageAsist and Research Management System and MS SQL Server for AcademicInfo), while the synchronization scheme is a multi-master one, each of the databases requiring bi-directional synchronization with the master database – see figure 2.

On the other hand, this solution ensures significant autonomy functional advantages for information systems, compared to a direct Active Directory integration & mapping.

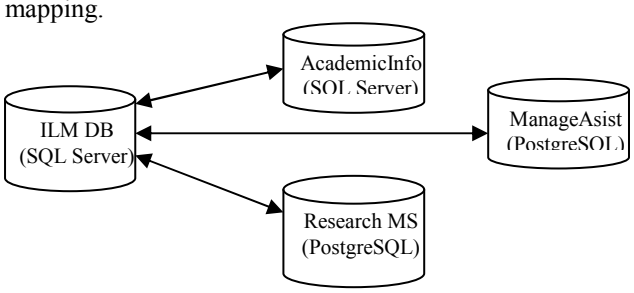


Figure 2: Synchronization scheme

The MS SQL - PostgreSQL synchronization has been tackled in literature [18]. Although a certified multi-master synchronization is fairly complex, in our case, the information that has to be synchronized / replicated between all databases is reduced to a few tables (see the description above for details): users, organization chart, units, people, with a fairly low modification rate, and therefore inducing a moderate network load.

We are currently implementing a synchronous database replication, in order to obtain real time synchronization. Since such a process is quite resource demanding, we also intend to explore some asynchronous mixed solutions if the run time of the global synchronizing & monitoring system tends to increase over a reasonable limit. The asynchronous solution transfers the whole database, having a larger data load, but the moment in time will be chosen in respect with the applications' low workload or even stand-by state.

In order to implement the authentication server we use MS Identity Lifecycle Management server, which has advanced integration facilities with our e-learning portal, and we are in train of configuring the necessary permission mappings from the dedicated information systems into the authentication server in order to complete the integration facilities.

III. THE WEB SERVICES

We further describe the web services provided by the e-learning portal. A prior portal version is already available at [23]

A. E-learning functionalities

E-learning systems [14] may be viewed as advanced tools which assist teachers in creating a cooperative, multidisciplinary and explorative learning environment and students in accessing these learning facilities and developing learning interactions within this environment. The implementation of e-learning facilities strongly contributed to the development of the student and goal centered learning model [1]. E-learning facilities are usually provided by means of web services.

The web-based e-learning facilities provided by the described portal are the SharePoint (see [17]) built in ones, adapted to our specific needs, and include:

- ◊ content management (see Figure 3) and sharing,
- ◊ schedule management and sharing,

◊ communication facilities (e-mail – OWA type, discussion lists, etc.) ,
◊ evaluation tools and feed-back facilities;
◊ task management, blog and RSS tools,
◊ survey tools, as well as other functionalities.

The system is also open to adding new web-parts, services or components (for example, the evaluation ones are in train of being developed).

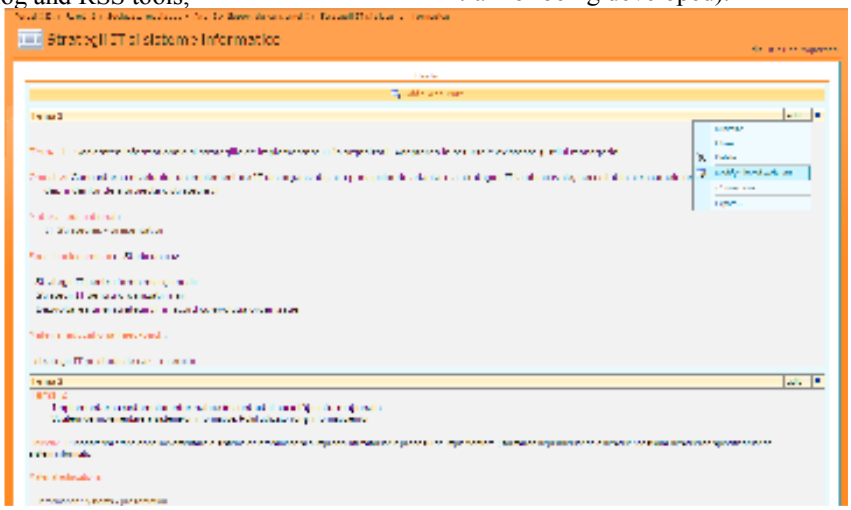


Figure 3: Managing an educational resource (in design permissions)

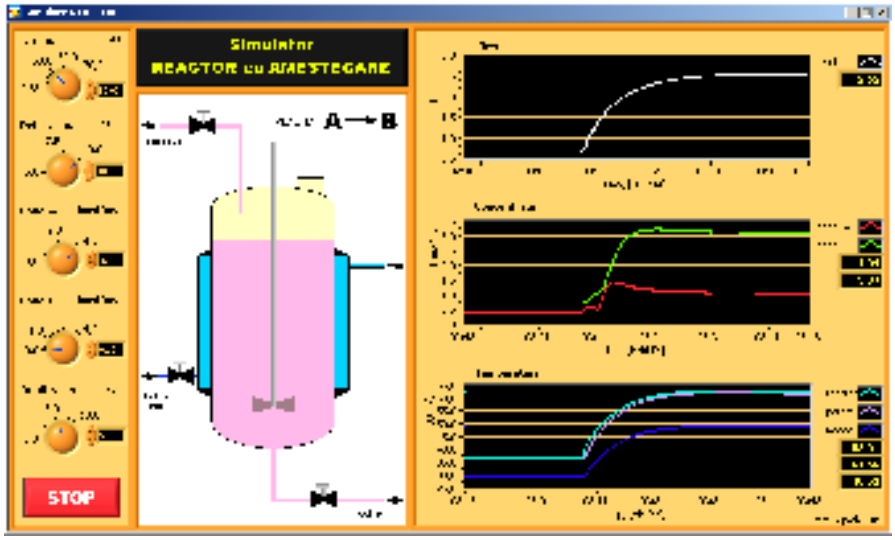


Figure 4 Example of simulation process

B. Virtual lab facilities

Virtual lab services enable modelling of processes that may be tedious to be accessed in real conditions, or are required to be accessed remotely. We decided to include such facilities in our e-learning portal in order to support learning in experimental sciences and sharing of such (typical) experimental knowledge by electronic means. Application fields are related to: process engineering, environmental engineering, physics, chemistry, biology, etc.

The virtual labs facilities are in train of being implemented on the portal and include:

◊ On-line virtual experiments and on-line labs;
◊ Case studies based on mathematical modelling and simulation;
◊ Recorded video sequences and on-line video streaming;
◊ Material posting, to be further processed with dedicated clients

An example of simulation process is presented in figure 5

C. Dedicated services provided by the information systems

AcademicInfo [20] is an integrated information system dedicated to managing educational information, with dedicated processing facilities for secretariats, specific access facilities for students and teachers and relevant synthesis regarding the educational process. The system models educational processes at BBU level, ensuring course selection from all faculties' curricula in study agreements, models in a flexible manner various types of educational activities at all study levels (BA, MA, PhD, continuous

 ISBN: 978-1-61208-100-7

education, specific curricula), ensures multilingual support in processing and reporting.

The dedicated web services that are provided include:

- ◊ *For students*: student curricula and grade access, fee management, student documents and requests, on-line course evaluation;
- ◊ *For teachers*: curricula and grade management (for the activities that are conducted), access to results of student evaluations;
- ◊ *For academic management*: specific syntheses, access to results of student evaluations at faculty or university levels

ManageAsist system is the integrated software system for administrative management that has been developed for our university. The system can be viewed as an ERP system; within its design and implementation, we integrated systematic efficiency principles in software design – see [6].

ManageAsist's principles and facilities are adapted for high education institutions; the system contains the following modules: Document management, Assets, Warehouse, Cashier, Finance, Accountancy, Grants, Human Resources and Acquisitions, and decision assistance facilities. Their implementation has pursued systematic and efficient principles [2]. Each module contains management reports for the corresponding compartment. Relevant synthesis from each compartment will be integrated, together with global management tools into a decision support module.

In [4] we address the advantages of pursuing advanced design principles in the implementation stages of the system, and in designing a flexible framework for efficiently integrating the system's modules. We also deal with means of managing hierarchical data structures, together with efficiency issues in respect with processing them. Each module includes levels [6] for specific document processing, operational facilities and reporting, level that provides management assistance information for the corresponding compartment.

The web services [21] include access to grant financial information and management of acquisition request, including specific reporting facilities for management levels.

Our **Research Management System** [22] is a web based system that we have developed and implemented within Babes-Bolyai University's (BBU) in order to manage research activities. The system offers – via web interfaces – accessible and user-friendly means of collecting specific information, and automatically performing quantitative analyses, syntheses and evaluations based on the collected information. The system may be viewed as a tool for quantitative research evaluation, its more general aim being to ensure proficient management of the research activity within BBU and supporting the design of competitive strategies in the field by means of this dedicated software system.

The system provides specific web-based facilities for:

- ◊ *Academic and research staff:* activity collection and reporting;
- ◊ *Unit / department management*: specific syntheses
- ◊ *Faculty / university management*: specific syntheses

The design and implementation principles of the Research Management software system, its architecture features and its impact in research activity management for the members of the academic & research staff, but especially for research management levels: chairs, institutes, departments, faculties, university are described in [3].

IV. WEB PORTAL EVALUATION

The web portal that we have implemented in order to provide e-learning and web integration facilities is based on a SharePoint solution, which has proven to be very convenient in flexible administration and integration purposes.

A. Evaluating E-learning Portal Functionalities

Regarding the system feed-back, we developed dedicated questionnaires for administrators, students and teachers [6], in order to obtain a general evaluation regarding existing facilities, platform functionalities and to ensure future developments.

The questionnaire has been created and interpreted using the survey functionality built-un in the platform (Share Point Portal); we underline in this respect the flexibility of the platform's tools.

We further discuss the results obtained consequent to monitoring the administrators' survey [8], since administration facilities are relevant for the portal capabilities.

Administrators were requested to evaluate, on a 1 – 5 scale (1=very weak, 2=weak, 3=moderate, 4=good, 5=very good), the following platform characteristics [6]:

- ◊ *administration functionalities* - the average weighted grade was 4.14;
- ◊ *communication functionalities* - the average weighted grade was 3.86;
- ◊ *functionalities for administering educational content* - the average weighted grade was 4;
- ◊ *functionalities for developing educational content* - the average weighted grade was 3.43;
- ◊ *functionalities for platform development* - the average weighted grade was 4.43;
- ◊ *platform adaptability / flexibility characteristics* - the average weighted grade was 3.57;
- ◊ *reporting facilities* - the average weighted grade was 4.

We can notice that all characteristics are positively rated, most of them being qualified above 'good' (weighted grades >=4).

We further discuss some of the most relevant responses in respect with the platform characteristics: *administration functionalities and functionalities for developing educational content* are well rated: 29% very good, 57% good, 14% moderate – see figure 5, 6; *functionalities for platform development* are very well rated: 43% very good, 57% good – see figure 7.

 ISBN: 978-1-61208-100-7

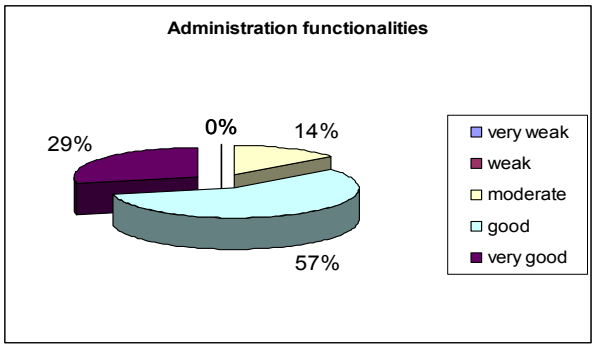


Figure 5: Administration functionalities

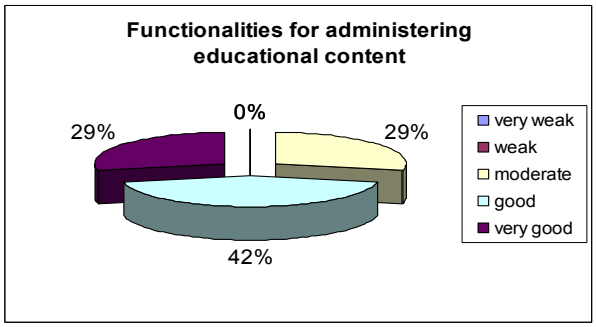


Figure 6: Administering the educational content

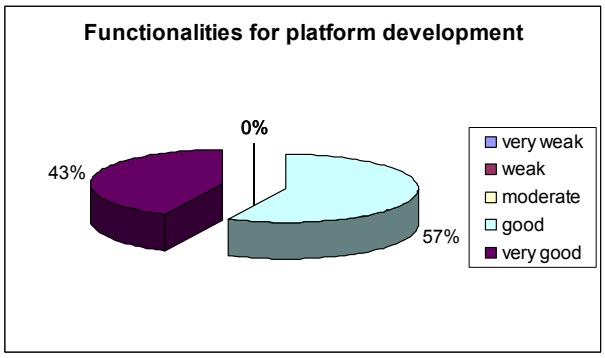


Figure 7: Functionalities for platform development

We may conclude that the adaptability and flexibility characteristics of the platform that were mainly aimed are actually implemented and we have a very good feed-back in this respect.

We shall continue monitoring the system in order to ensure its most appropriate use and development; in this respect, we are confident that our prerequisites regarding adaptability specifications in system upgrades will also prove to be very useful in the future.

B. Authentication Characteristics

Our system provides single sign-on [28] facilities using a MS ILM - Identity Lifecycle Management Server [25] and ISA Server; the authentication is based on Active Directory facilities.

While OpenAuth protocol [26] grants access whithout sharing passwords, the architecture we describe uses the ILM facilities for synchronizing authentication information (User, password), according to [25]. This credential exchange is similar to the one used by OpenID protocol [27], but is performed by means of the ILM built in facilities [25]. The authentication mechanism also implements the MS Domain Trust policy [24].

We consider that the architecture based on ILM server [25] has good implementation advantages, since it already provides built-in web authentication facilities.

C. System Overview and Perspectives

E-learning implementations should pursue the same principles and stages as for other dedicated software systems [2] - the user involvement within the stages of system requirements, verification and implementation are of outmost importance for a successful implementation. Though e-learning facilities are fairly standardized, it is important to take into account future upgrades of the implemented system.

The implementation of the present e-learning system and the undergoing integrated portal within "Babes-Bolyai" University of Cluj-Napoca [5], [6], Romania systematically applied the above described principles. The flexibility system requirements and de-centralized system administration that were pursued are expected to prove their efficiency in the future developments.

V. CONCLUSIONS AND FUTURE WORK

The paper focuses on web service integration and tools for system integration as advanced tools for providing such integrated services. Our case study is performed on an academic institution, the universities' case being quite complex, since their activity covers a wide range of areas: education and learning, research, administration.

We describe an efficient integration solution for providing web services into a global portal, including web-based facilities offered by dedicated information systems. The solution is based on MS technology and provides means of integrating various information systems by implementing a single authentication server and mapping specific facilities from the dedicated information systems, using different database management systems, into the portal, for each user category. This architecture is based on a global integrated database and a permission mapping scheme for ensuring appropriate access into the dedicated information systems. We are in train of defining the necessary permission mappings in order to fulfill the implementation.

The system framework integrates various web services that are provided within the portal: learning facilities, virtual labs and dedicated information facilities.

The advantages of the proposed solution rely in providing a uniform web framework for: database synchronization of various information systems databases and web access to e-learning and information collaboration & sharing tools and dedicated system facilities. The proposed framework enables data and service integration that may be further exchanged within federated web services.

This web service integration solution has a good extensibility degree and may be applied in various cases.

 ISBN: 978-1-61208-100-7

ACKNOWLEDGMENT

The present work is supported by the EU funded grant, within the European Fund for Regional Development, "CCE 124/323/31.08.2009 SMIS 4424 - Sistem electronic aplicativ integrat de educaţie al Universităţii Babes-Bolyai" – Integrated applied electronic system for education of Babes-Bolyai University - BBU, contracted by BBU with the Romanian Ministry of Communication and Information Society, Organismul Intermediar pentru Promovarea Societăţii Informaţionale (the Intermediary Structure for Promoting the Information Society), during 31-08-2009 – 31-08-2011

We thank to the whole development team in our IT department for their contribution to developing ManageAsist, AcademicInfo, Research management information systems and to administering the e-learning portal: Florentina Tufiş, Călin Miu, Simona Nemeş, Dan Pop, Monica Bojan, Carmen Pavel, Ana Iuhos, Ana Bara. We are in train of registering the intellectual property rights of the information systems for the whole implementation team.

REFERENCES

[1] M. Allen, Guide to e-Learning. Wiley, 2002

[2] A. Andreica, IT Strategies In Increasing Business Competitiveness, Studia Europaea, LI, 3, pp.139-148, 2006

[3] A. B. Andreica and P. S. Agachi, "Design and Implementation of An Integrated Software System for Managing Research Activities in Universities", 7th RoEduNet International Conference - Networking for Research and Education, UT Press, Ed: E. Cebuc, pp. 90-95, 2008

[4] A. B. Andreica, D. Stuparu, and F. Ghetie, "Design and Implementation of an Erp System for Universities", Proceedings of IADIS Information Systems 2009, IADIS Press, Eds: M. Nunes, P. Isaias, P. Powell, pp. 315-322, 2009

[5] A. B. Andreica, "Design and Architecture of an Integrated E-learning Environment. Case Study on Babes-Bolyai University, Cluj-Napoca, Romania", Research, Reflections and Innovations in Integrating ICT in Education, vol 1, Formatex, Badajoz, Spain, Editor: A. Mendez Vilas, A. Solano Martin, J. Mesa Gonzalez, J. A. Mesa Gonzalez, pp. 507-514, 2009

[6] A. B. Andreica, F. Covaci, D Stuparu, and G. Pop, "An E-Learning Web Portal with System Integration Facilities", Web Information Systems and Technologies 2010, Valencia, Spain , Proceedings of 6th International Conference WEBIST , vol 1, INSTICC, Editor: Joachuim Filipe, Jose Cordeiro , pp. 131-136, 2010

[7] M.Berry and G.Linoff, Mastering Data Mining, John Wiley & Sons, 2000

[8] Client/Server and the N-Tier Model of Distributed Computing, Micromax Information Services Ltd., 1999

[9] E. Gamma, R. Helm, R. Johnson, and J. Vlissides, Design Patterns, Teora, 2002

[10] M. T. Goodrich, R. Tamassia, and D. Yao, "Notarized Federated Identity Management for Web Services" http://www.cs.brown.edu/cgc/stms/papers/notarizedFIM.pdf , accessed July 2010

[11] Hasselbring, W. "Information System Integration". Communications of the ACM, 43 (6). pp. 32-36, 2000

[12] Hasselbring, Wilhelm and Pedersen, Susanne, "Metamodelling of Domain-Specific Standards for Semantic Interoperability". Lecture Notes in Computer Science, 3782, pp. 557 – 559, 2005

[13] K. Hoganson and M. Guimaraes, "N-Tier Client/Server Course", Consortium for Computing Sciences in College Conference, Dunwoody, Georgia, 2003

[14] W. Horton and K. Horton, E-learning Tools and Technologies: A consumer's guide for trainers, teachers, educators, and instructional designers, Wiley, 2003

[15] D. Stuparu, A Andreica, and I. Mantu, "Comparing Access Techniques on Databases in Distributed Application Frameworks", Proc. of Collaborative Support Systems in Business and Education, BBU, Cluj-Napoca, pp. 1-10, 2005

[16] R. Webster and F. Sudweeks (2006) Teaching for e-Learning in the Knowledge Society: Promoting Conceptual Change, in Academics' Approaches to Teaching. Current Developments in Technology-Assisted Education http://www.formatex.org/micte2006/pdf/631-635.pdf , accessed June 2010

[17] MS Learning Gateway and SharePoint Portal http://www.microsoft.com/education/solutions/higheredportals.aspx , accessed June 2010

[18] PostgreSQL Team , "High Availability, Load Balancing, and Replication" http://www.postgresql.org/docs/8.3/static/high-availability.html , accessed June 2010

[19] Business Explorer for Web Services, http://www.alphaworks.ibm.com/tech/be4ws , accessed July 2010

[20] BBU AcademicInfo System http://academicinfo.ubbcluj.ro/Info , accessed July 2010

[21] BBU ManageAsist System http://manageasist.ubbcluj.ro , accessed July 2010

[22] BBU Research Management System http://infocercetare.ubbcluj.ro , accessed July 2010

[23] BBU E-learning portal https://portal.portalid.ubbcluj.ro , accessed July 2010

[24] Federated Identity Patterns in a Service-Oriented World http://msdn.microsoft.com/en-us/architecture/cc836393.aspx , accessed July 2010

[25] Identity Lifecycle Management Server http://www.microsoft.com/windowsserver2003/technologies/idm/ilm.mspx , accessed July 2010

[26] OAuth protocol - http://oauth.net/ , accessed July 2010

[27] OpenID protocol - http://openid.net/ , accessed July 2010

[28] SingleSignOn http://www.authenticationworld.com/ , accessed July 2010

[29] WS-Federation - Web Services Federation Language, Dec 2006, BEA Systems, IBM Corporation, Layer 7 Technologies, http://www.ibm.com/developerworks/library/specification/ws-fed/ accesses July 2010

From Heterogeneous Sensor Sources to Location-Based Information

Tracking and support of service technicians in an industrial environment

Mareike Kritzler
Institute for Geoinformatics
University of Münster
Münster, Germany
kritzler@uni-muenster.de

Andreas Müller
Advanced Technologies and Standards
Siemens AG, Industry Sector
Nürnberg, Germany
Andreas_w.Mueller@siemens.com

***Abstract*—This paper describes the transformation process from spatio-temporal coordinates to location-based support. This transformation is described as a workflow starting with heterogeneous sensor sources, which provide spatio-temporal data. As a next step, data fusion operations provide a precise and accurate location. This location is subsequently enriched with context information. The main contribution is to establish a dynamic link between the spatio-temporal aspects of a smart industrial indoor environment and its descriptive semantic information model in order to enable location-based support of service technicians with mobile devices.**

Keywords - Tracking, Sensors, LBS, Georeferencing.

I. INTRODUCTION

Smart indoor environments are used in multi-disciplinary research areas and different contexts like Smart Homes, Classrooms or Industrial Environments. They are equipped with different technologies, sensors and actors. The purpose of Smart Environments is to support human beings in their daily life in real time [1].

A. Problem

Smart environments are equipped with different sensor sources. Location and tracking technologies collect spatio-temporal data in the form of three-dimensional coordinates. Single sensor sources are not precise and accurate enough but the combination of various technologies and different levels of instrumentation allows an exact indoor positioning. In this context three major problems arise:

- Without the fusion of all obtained spatio-temporal data, the exact spatial position of tracked items cannot be determined. The resulting Cartesian coordinates describe a position at a certain timestamp based on a known reference system in the environment.
- Without the enrichment of fused positions with context information, we would lack information about the current spatial context, which is required for different subsequent tasks.
- Without the deployment of applications (to support users) for mobile devices, the context coordinate cannot be used to obtain location-based information in a smart environment.

B. Motivation

To provide location-based support for human beings in a Smart Environment, the transformation from fused spatio-temporal coordinates to context information can deliver necessary location-based information. A workflow from coordinates to information is needed to be applicable to Smart Environments in different contexts. Location-based information is necessary in all kinds of applications.

For example, smart industrial work environments require service technicians to be constantly provided with correct instructions for performing various maintenance tasks. Service technicians use mobile devices for in-situ context-sensitive information provision. Especially in dynamic environments the applicability of work instructions is strongly dependent on the technician's spatial position. However, location-based information can only be provided if individual positions are known exactly. The consideration of the associated contexts allows to close the gap towards a semantically meaningful processing.

The presented work is motivated by the need for spatial support for service technicians in an industrial environment to efficiently and safely accomplish their maintenance work.

C. Background

Information required by service technicians is always dependent on prior phases in the plant lifecycle (see Figure 1).

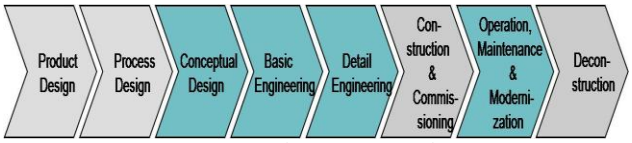


Figure 1. Plant lifecycle (adapted from [10])

During the phases *Conceptual Design*, *Basic Engineering*, and *Detail Engineering*, plant drafts, designs, and documentation of the plant are created. These are essential information sources for maintenance tasks. Among others, these include topological, structural and functional information, stored in different formats such as visual representations, e.g., 3D computer-aided design (CAD)

 ISBN: 978-1-61208-100-7

models or floor plans, or text-based semantic information models. Thus, spatial knowledge is created and used for the respective plant prior to the *Operation, Maintenance and Modernization* phase. However, each phase relies on individual specialized information representations, resulting in a heterogeneous information landscape. Product/Plant Lifecycle Management (PLM) systems attempt to cope with this heterogeneity by employing a more unified information model over the entire plant lifecycle, making task-specific information available for all phases on request. Still, when performing maintenance tasks, a service technician needs to use floor plans, maps or other navigational means in order to reach and identify a component to be serviced and to subsequently access the related information from a PLM system. By connecting these capabilities and information warehouses with localization and spatio-temporal information processing, service support systems can benefit from unified access to all necessary PLM information based on the technician's current position, thus being transformed into location-based services (LBS) for maintenance tasks.

The proposed problems were addressed in a joint project that took place in cooperation between the Institute for Geoinformatics and the Siemens AG. At the Siemens location in Nürnberg Moorenbrunn an industrial research facility, called the *Smart Automation Center* (SmA), is operated (see Figure 2). The SmA, which is equipped with different kinds of sensors, serves as a lab environment for the development of various conceptions and technologies in the field of manufacturing automation. There, among others, conceptions for multi-modal support of service technicians are developed.

Figure 2. Picture of the SmartAutomation Center in Nürnberg Moorenbrunn

D. Outline

Section II gives a short description of further smart environments. In section III the setup of the smart environment with the sensors and use cases is shown. Section IV describes the architecture and the components, which are used to get from heterogeneous sensor sources to location-based information. In the last section V the paper is concluded and an outlook is provided.

II. RELATED WORK

Smart Environments are developed and established for various purposes and with different contexts. Homes become smart with intelligent surfaces, wall displays or by monitoring the movement of their inhabitants [2]. This paper focuses on the smart industrial environment available at the Siemens laboratory. In the following, further examples for smart research laboratories are listed:

A. *SmartFactory* [3]

The SmartFactory is a manufacturer-independent European demonstration factory for industrial applications of modern information technologies. The factory aims at the development of innovative technologies for industrial plants. Furthermore, applications for different industrial branches are developed. The work done in the SmartFactory is divided into five parts:

- Localization services
- Virtual factory
- Control systems
- Mobile devices
- Basic technologies

B. *Living Lab Innovative Retail Laboratory* [4] (IRL)

The IRL is a research laboratory, which focuses on topics related to intelligent shopping support systems. The assistance systems are tested concerning their suitability for daily use. Different forms of interactions with consumers like speaking products as well as intelligent shopping carts are developed. Furthermore, indoor positioning and navigation are part of the research.

C. *Bremen Ambient Assisted Living Lab* [5] (BAALL)

BAALL is a 60 m² big apartment, which is invisibly equipped with different technologies. The apartment is disability-friendly and in accordance with the requirements of elderly people. The idea behind the project is that elderly persons can stay as long as possible in their own apartments, which are able to assist and support the inhabitants when they need help. Focus lies on the mobility assistance and on environmental support. For mobility assistance, an intelligent walker has been developed [6]. Furthermore, a controller for an automatic wheelchair was developed for smart driver assistance [7].

The interoperability and standardization of the inbuilt components are an important requirement for the development of assistance systems.

III. SETUP

The SmA allows the addressing of different challenges along the plant lifecycle. In the following the purpose of the SmA, the applied technologies and a use case, which is taking place in the SmA are described.

 ISBN: 978-1-61208-100-7

A. *Smart Automation Center*

With its different functional modules, the SmA realizes a complete exemplary product lifecycle by filling, inspecting, commissioning, and recycling bottles with various products. For the entire plant, highly detailed 3D CAD models are available (see Figure 3), as well as PLM information models for selected components.

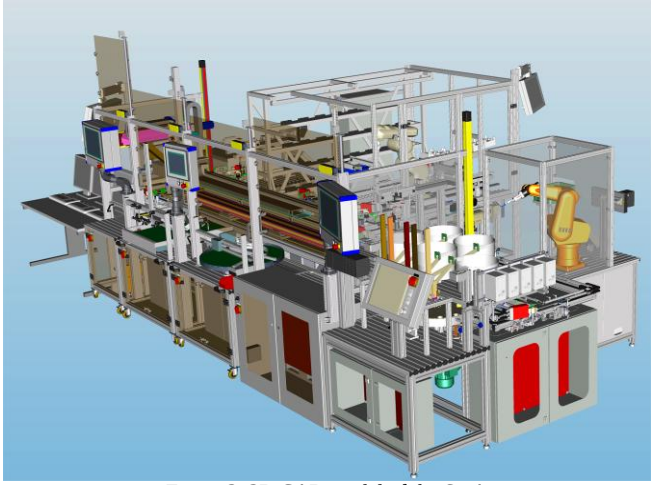


Figure 3. 3D CAD model of the SmA

Among other activities, ideas for the improvement of multi-modal support for service technicians are developed in the SmA. These include ideas how to:

- obtain proper / correct instructions, so-called task flows, in dynamic working environments
- efficiently place and use digital notes on components as a means of collaboration support
- obtain necessary component-related information, dynamically adapted to individual situational needs.

The SmA is equipped with various kinds of sensors, which are discussed in the next section.

B. *Tracking and localization technologies in the SmA*

Different tracking and localization technologies are integrated in the SmA to obtain spatio-temporal data (see Figure 4). In this scenario, Ultra Wide Band (UWB) (Ubisense) and RFID technology are in use. Installed keystroke sensors, which have a fixed location are also used for location determination. Additionally, localization via Wireless Local Area Network (WiFi) Fingerprinting is possible as well.

- UWB: This generic term characterizes radio systems with huge bandwidths [8]. The Ubisense [9] system (Series 700), which uses UWB radio signals, is established for multi-user tracking in the SmA. The mobile location tags transmit UWB radio signals to four receivers installed in the SmA. This allows three-dimensional indoor localization. Empirical experiments in the SmA show a precision of approx. 20 cm on the horizontal plane and approx. 30 cm along the vertical axis. Unfortunately, in some areas of the SmA, positioning is strongly impaired because of shielding due to metallic materials (like in the high rack storage).
- RFID: The RFID system, which is in use in the SmA is installed as a terminal approach. This means that the positions of the RFID readers are dynamic while the positions of the RFID tags remain static. The RFID readers are attached to mobile devices and change their locations in time with the user. The RFID tags are placed in the SmA at different components and have fixed positions (a three-dimensional coordinate). The position has to be taken a priori and stored in a database. The timestamp of the reading event together with the identifier of the RFID-reader are used for tracking.
- Keystroke sensors: All mechanical sensors, which can be used by human beings are summed up as keystroke sensors, e.g., the emergency stop button, the control panel of the bottle picker or the touch panels, which control components of the SmA. All of these sensors have a fixed position. Furthermore, their usage can be logged. The position has to be taken a priori and stored in a database. The keystroke sensors can be used as a source for precise positioning in the SmA because the exact location is known as well as the time stamp when it was in use. However, the localization is anonymous and in a multi-user scenario the coordinates can only be assigned to a single user if other tracking and localization technologies are in use at the same time.

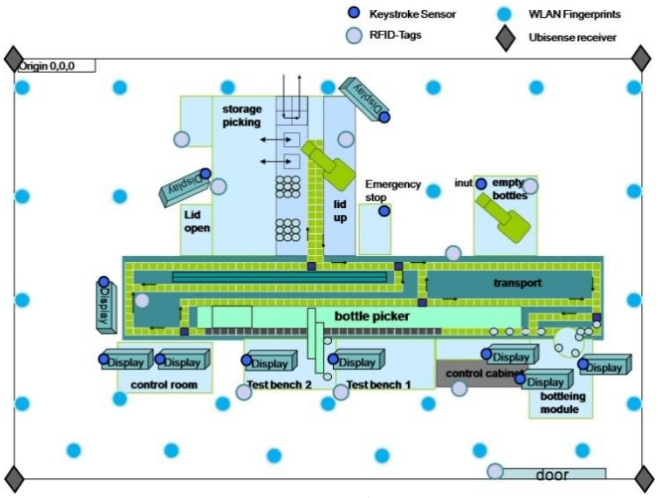


Figure 4. Schematic overhead shot map of the SmA with integrated sensor sources for location determination

- WiFi: Data collection via WiFi fingerprinting uses the received signal strength (RSS) of access points, which are installed in the building hosting the SmA. The fingerprints have to be taken a priori with a mobile device for the whole SmA. The precision of the localization depends on the amount of taken

 ISBN: 978-1-61208-100-7

fingerprints. Furthermore, the strength of WiFi signals is variable and can lead to imprecision.

C. Use cases

Service technicians are supposed to have in-situ access to all required information and directly use all required information on both the maintenance task to be accomplished and the component to be serviced. Hence, several fundamental use cases have to be considered in terms of such mobile support for service technicians (see Figure 5).

Although the SmA serves as a relatively small lab environment, the use cases considered here are equally valid for full-size industrial plants and other smart environments.

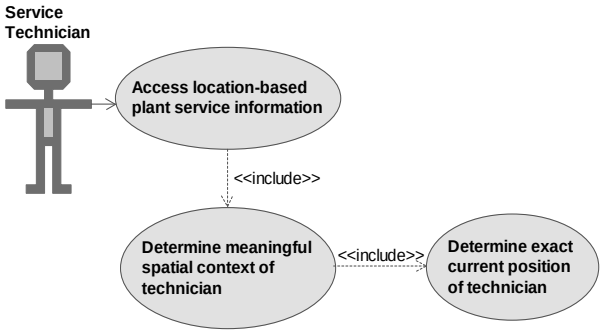


Figure 5. Main use cases considered in the scenario displayed in UML notation

Firstly, the technician's exact current spatial position needs to be determined. However, not all localization technologies are equally well-suited for an entire plant. Instead there are different areas within the plant where some localization technologies typically show better performance than others. For instance, high rack storage areas might block UWB localization. In such cases, making plant components directly identifiable by equipping them with e.g., RFID tags would be useful. However, this would result in high effort for the plant operator, especially when carried out after plant construction. From this point of view, different localization technologies have to be considered in a combined approach in order to compute the actual current position.

Afterwards, the semantically meaningful spatial context of the technician's position has to be determined. This is necessary because of the nature of issues in the field of service support. For instance, service technicians might need to ask for the components that a specific to-be-serviced component in front of them is connected with, e.g., electrically ("Which components are powered by this power supply?") or even in terms of the entire production process ("What are the effects of shutting down this conveyor belt?"). The service phase is part of the plant lifecycle. In this context usually semantic models are employed for the representation and management of this type of service-related plant or component information. Such semantic models are the basis of PLM systems. However, due to their origin, these models typically do not rely on spatial information but instead use unique names (URIs) for the identification of the described components and their relations. If at all, spatial information is merely provided as an optional attribute and thus cannot be used as identification criteria. Furthermore, the required domain logic for handling spatial issues is typically not part of PLM systems or other semantic information management systems. Therefore a contextual link between the purely spatial world and the world of semantic plant information management has to be introduced.

Finally, the determined contextual meaning of the position needs to be used for location-based access to service-related information. Since the service technician moves within a semantically described plant, his movements and position in the vicinity of components can be used to provide him with location-based information from a plant information management system.

IV. ARCHITECTURE

This work consists of three different parts. First, the spatio-temporal data has to be collected and fused. Second, the Cartesian coordinates have to be enriched with context information. Third, the coordinate has to be used as an input for applications on the mobile device (see Figure 6).

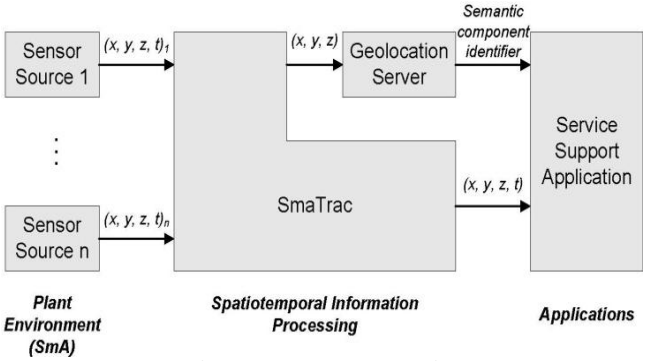


Figure 6. Flow of the Cartesian coordinates for mobile support

A. SmaTrac – Handling of tracking data

Smart Environments are equipped with different kinds of sensor sources. The SmaTrac framework was developed to integrate and process spatio-temporal data obtained by heterogeneous sensor sources from moving items (e.g., like service technicians). This framework, which was implemented in Java does not rely on, nor is it limited to, one tracking or positioning technology but instead considers the combination of several technologies to calculate the best positioning result. SmaTrac is scale-independent (in space and in time) and considers different sizes and characteristics of the tracked items as well as different tracking environments or contexts. The framework consists of four parts:

 ISBN: 978-1-61208-100-7

- Data collection:
  The framework has to be able to use data obtained by heterogeneous sensor sources and has to provide common storage for the data.
  The implementation provides a universal data schema for a relational database. This schema is able to store both data, which has to be taken a priori and spatio-temporal data, which is taken in real time. The data scheme stores localization events, which consist of timestamps and IDs for the sensor sources. The sensor source ID is the key to the coordinates. Furthermore, the schema allows the storage of metadata concerning the tracking or localization technologies, which is important for further processing.
- Data processing:
  The processing has to enable the combination of data obtained from different technologies.
  The spatio-temporal data is processed in a core component. The core is able to use different sensor fusion approaches, but in the first implementation the position is calculated by the arithmetic mean. For more advanced fusion approaches, the precision and accuracy are taken into account when new coordinates are calculated from coordinates, which were obtained by different sensor sources.
- Data providing:
  The framework has to be able to fetch data out of a relational database as well as to provide it to external applications.
  The framework provides different internal and external interfaces. Internal interfaces enable the retrieval of stored data from the database and provide the fused data to different applications. External interfaces are used to provide the processed data (the output of the framework) to external applications as their data input. The data is provided in Extensible Markup Language (XML) format to make the reuse of the processed data with already implemented software possible.
- Data usage:
  The framework needs to use calculated coordinates for further domain-specific applications.
  One application, which can be used for different contexts and tracked items is the visualization application. The use of Java 3D and the Virtual Reality Modeling Language (VRML) makes the display of all kinds of three-dimensional context models of different environments possible.
  Other applications analyze large amounts of data, for example via machine learning techniques, to find patterns of tracked items.

B. Geolocation Server – Giving tracking data a meaning

Most tracking or localization technologies use a coordinate-based approach to provide locations. Available indoor localization and tracking systems offer coordinates in a predefined reference system. SmaTrac uses and calculates Cartesian coordinates. Cartesian coordinates describe locations by a set of numerical values. Positions are determined sufficiently by that approach but the circumjacent context is neither described nor known. The Geolocation Server overcomes this lack of contextual awareness. It introduces a service layer, which enriches Cartesian coordinates with context information. The usage of well-known standards allows its introduction into any service chain, and therefore the enhancement of traditional tracking with context-based information. The Geolocation Server is mainly divided into two functional parts:

- Data processing:
  Context information has to be assigned to specific coordinates (by annotation of coordinates with object references of their surrounding context) to realize a LBS, which provides context information.
  In this implementation, the Geolocation Server uses a Web Feature Service (WFS) and Filter Encoding (FE). Existing CAD files (see Figure 3) were converted into 2.5D shapefiles (these are models of the environment) and annotated with Unique Resource Identifiers (URI) pointing toward additional information stored in ontologies. The usage of FE permits to query for a set of coordinates. The retrieved context of the coordinate is transmitted, encoded in the Geographic Markup Language (GML) and serves the URI along with the geometry of nearby objects.
- Data providing:
  The identified entities of the circumjacent context of the coordinates have to be communicated via a machine readable interface. The interface has to provide the logical entities' geometry and a set of attributes, which can be used to obtain further information like electrical or mechanical connections.
  The implementation uses the URIs to gain additional needed information of any kind. This approach allows one to use this LBS – the Geolocation Server – as a machine-readable middleware layer for other applications assisting users at their task.
  It is possible to obtain information about the surroundings of the coordinates by blending existing blueprints and floor plans with the tracked coordinates in a common reference system. Annotations in the blueprints point toward additional resources, which can hold functional descriptions or topological information.

C. Application – Providing location-based information

On this level, a variety of applications are possible. This could include the display of electrical or mechanical connections of a component, the provision of maintenance

 ISBN: 978-1-61208-100-7

instructions, the adaptive display of floor maps, or the handling of digital notes attached to components.
Two different applications have been created:

- The SCADA system SIMATIC WinCC [11], which is installed in the SmA for plant control, represents plant areas and components by means of operable image elements. This control system was extended by an ontology describing structural and functional aspects of SmA components. Subsequently, the image elements were enriched by mappings to features identifiable by the Geolocation Server and to individuals in the ontology. In this scenario, the service technician can learn about the exact location of a specific SmA component and its vicinity along with its structural and functional aspects by selecting the corresponding image element onscreen. By using extensions of WinCC for mobile devices and transferring the respective images, it is also possible to e.g., highlight certain image elements on the mobile devices according to the technician's current position.
- As part of the German Federal Ministry for Research and Education (BMBF)-funded project AVILUS [12], semantic enhancements of digital graffiti – so-called "Virtual Post-It" – have been researched. By means of Virtual Post-It, service technicians can attach semantic descriptions of the actions performed during service tasks directly to plant components. These Virtual Post-Its can subsequently be used by the technicians for collaboration purposes as well as for dynamic adaptation of location-based service information situationally supplied by PLM systems. Here, the setup provides essential localization information, since both the technician's position and the locations of affected components need to be determined constantly.

V. CONCLUSION AND OUTLOOK

This paper describes the architecture of a dynamic link between different sensor sources and a location-based application in the case of indoor tracking of service technicians in an industrial smart environment. Cartesian three-dimensional coordinates obtained by heterogeneous sensor sources are first fused to get the best (that means the precise and accurate position) indoor location. The new coordinate is then used in a georeferencing process to gain access to location-based and context-related information, which is provided by a mobile device.

The main contribution of this work is the architecture of a transformer from Cartesian coordinates to location-based information. The architecture is independent of both the used localization and tracking technologies and the smart environment. Thus, semantic information models describing a smart environment become accessible on grounds of a user's spatial position.

In addition to it, this paper shows the results of a productive cooperation between theoretical university research and industrial application, which are in use in industry.

In future, the development of Virtual Post-It will continue. Furthermore, extending the setup with prediction of specific tasks and likely maintenance workflows is being considered.

ACKNOWLEDGMENT

We thank the members of the Siemens AG, Industry Sector, Advanced Technologies and Standards for their support and we thank the AVILUS team for providing us with data. We also thank Antonio Krüger for his research advising.

REFERENCES

[1] P. Mikulecký, T. Lišková, P. Cech, and V. Bureš. "Book Series on Ambient Intelligence and Smart Environments". Published by IOS Press. Volume 1: Ambient Intelligence Perspectives. 2009.

[2] A. S. Taylor, R. Harper, L. Swan, S. Izadi, A. Sellen, and M. Perry. "Homes that make us smart". Journal Personal and Ubiquitous Computing, Springer, London, Volume 11, Number 5, pp 383-393, June, 2007.

[3] Homepage of SmartFactory: http://www.smartfactory-kl.de/. Date of access: 2010-05-23

[4] Homepage of IRL: http://www.dfki.de/web/living-labs-de/irl. Date of access: 2010-05-23

[5] Homepage of BAALL: http://baall.informatik.uni-bremen.de/en/index.php/Main_Page. Date of access: 2010-05-23

[6] T. Röfer, T. Laue, and B. Gersdorf. "iWalker - An Intelligent Walker providing Services for the Elderly". In Technically Assisted Rehabilitation 2009.

[7] T. Röfer, C. Mandel, and T. Laue. "Controlling an Automated Wheelchair via Joystick/Head-Joystick Supported by Smart Driving Assistance". In Proceedings of the 2009 IEEE 11th International Conference on Rehabilitation Robotics, pp. 743–748, 2009.

[8] K. Siwiak. "Ultra-wide band radio: introducing a new technology". Vehicular Technology Conference, 2001. VTC 2001 Spring. IEEE VTS 53rd. Vol 2. Pp 1088-1093.

[9] Homepage of Ubisense System: http://www.ubisense.net/en/products. Date of access: 2010-05-23

[10] VDI-Richtlinie 4499 Blatt 2: "Digitale Fabrik – Digitaler Fabrikbetrieb". December 2009.

[11] Homepage of WinCC SCADA: http://www.automation.siemens.com/mcms/human-machine-interface/de/visualisierungssoftware/scada-wincc/Seiten/Default.aspx. Date of access: 2010-05-23

[12] Homepage of AVILUS: http://www.avilus.de. Date of access: 2010-05-23

 ISBN: 978-1-61208-100-7

Modeling Unified Interaction for Communication Service Integration

Juwel Rana
CSEE Department
Luleå University of Technology
SE-971 87, Luleå
Email: juwel.rana@ltu.se

Johan Kristiansson
Ericsson Research
SE-971 28, Luleå
Email: johan.j.kristiansson@ericsson.com

Kare Synnes
CSEE Department
Luleå University of Technology
SE-971 87, Luleå
Email: kare.synnes@ltu.se

***Abstract*—Social network inspired communication services has made tremendous success, allowing users to communicate and share user generated contents in an efficient way. At the same time, Tele-com services are becoming more open, which makes it possible to develop improved social networking services. One of the problem that needs to be addressed for developing such services is how to fetch useful social information and make it available for the services running in the Cloud or personalized devices. This paper presents a generalized On-line interaction model that collects useful information from well known social networking services, and transforms the information into unified interaction patterns, which can be utilized for social data propagation or for discovering communication patterns. Ultimately, this allows the applications to incorporate social data for enabling smarter functions. The proposed interaction model is useful for presenting information about callers or propagating presence information to the classical address book, prioritizing information, inviting user for forming micro-communities.**

Keywords*-*interaction patterns; social networks; communication services; communication pattern discovery

I. INTRODUCTION

Today, Internet and computers has become essential part of every-day life. More and more users are using Web based social networking and communication services to interact and sharing information about their life [1] [2] and [3]. This trend leads towards the success of today's social networking services (e.g., Facebook, Twitter, LinkedIn, MySpace, and so on). For example, today Facebook has more the 500 million active users, and this number is increasing rapidly everyday.

At the same time as communication services has become an essential element of everyday life, the mobile handset industry is experiencing a paradigm shift towards open and more powerful mobile platforms [4]. These new open platforms are mainly driven by Internet companies such as Google and Apple that want to provide a rich Web experience anytime and anywhere, while reaching even more users. Ultimately, by providing a one-click-away Web experience users can better control their digital life and easier upload information at the point of inspiration, making today's Web services even stronger.

DISCLAIMER: The work has been carried out as part of an academic research project and does not necessarily represent Ericsson views and positions.

While the technological revolution toward open and powerful mobile platforms opens up for new types of pervasive services, it also opens up for the possibility to enrich classical mobile applications with social information [4] [5]. For example, classical address book in mobile phones may updated automatically with information obtained from social networks, image hosting services, Web forums, etc. Another example could be communication services that automatically determine where the users are located by analyzing status updates obtained from sensors and social networks, and then set up dynamic groups [6]. Implementation of such customized mobile applications, has been restricted for a long time by the device manufacturers and operators, but now it is possible to the developers to innovate new applications for the mobile devices.

Considering the huge amount of social data that is available on the web and the multitude of On-line communities adopted by the users, it is still a difficult task to extract useful social information and integrate it with the applications [7]. For example, some users might be using Twitter while other users might be using Orcut or Facebook, which makes it difficult to provide an unified solution that fits all users. Social data can also be differently formatted, and contain different type of information, or including conflicting information [8]. Therefore, there is a need for a software component that can automatically extract information from various social networking services and then summarize it in a machine interpretable format.

The ultimate goal of this paper is to enable communication services to be completely integrated with third party applications (i.e., mobile phone software or applications). As a step towards this goal, the paper addresses the problem of how to collect data from different communication services and get valuable information about users and platforms. Another problem addressed in the paper is how to transform the collected data into a machine interpretable format in an unified manner that can be consumed by the application developers. This problem can be described by the following scenario.

Peter has different contacts in different pervasive services. Some of them are friends, some other are colleagues, and some other may have similar interests and so on. For example, he has different friends in Facebook, Twitter, or in mobile phone's contact application. Peter uses a service to associate his social contacts. Due to tremendous information flow from all his

 ISBN: 978-1-61208-100-7

pervasive services, he wants to have assistance to prioritize his contacts and filter information/content overload based on social strength and context. He expects to have a service which keep track of all his interactions and automatically prioritize contacts based on social strength. The service collects Peter's social interactions based on his preferences and contexts and infers that information for social ranking, dynamic grouping or recommending contacts.

To face the challenges expressed above, we investigated the following research questions:

- How can on-line interaction be modeled in a uniform way to enable identification of users' individual communication pattern?

The paper is organized as follow: Section 2 discusses related work, Section 3 provides an On-line interaction model, Section 4 presents an early evaluation and Section 5 provides a discussion together with future work.

II. Related Works

A. Communication Services in Context of Social Networking

In communication services, social context opens new challenges and possibilities to design mobile applications with more intelligent functioning [9]. Although, adding social context increases security and privacy risks, but it also increases social interaction and collaboration in the Web [10] [11]. Andrew T. et.al, proposed "CenceMe" system, which is able to propagate user's presence status using the mobile sensors automatically to social networks [12]. Comparing with such system, the proposed approach is managing user presence information by aggregating presence information from different social networks and exposes this information to the social graph in an unified manner. Communication in CenceMe system is single-user basis (i.e., owner of the mobile device), while communication in our system is multi-user basis as the system is being able to collect status information for the contact lists.

B. Dynamic Group Discovery

The proposed work is not focusing dynamic group discovery, yet the solution can be associated in discovering dynamic group. Dynamic group can be formed based on social data in the Web contents as well as data generated by pervasive services [13] [6] and [14]. Zin Li et al., proposed social interest discovery mechanism based on user-generated tags [14] [15]. In that approach, user tags are considered as main data for discovering groups. To simplify such approach, proposed solution provides an unified data model for representing On-line interactions associating user's current context. Therefore, by using the data model, the dynamic group discovery can be simplified in large scale. In the proposed solution, collected data is aggregated and represented as MXML based data set to which provides a machine readable view of the data set for the third party applications.

C. Social Data Collection

Social Web API's are not enough to collect social data for analysis from social networking services due to lack of standardization, data formats, and user-understandable data model and access policies. It needs some form of format to continue analysis smoothly [16]. Although, there is a proposal for open social specification but the commercial social network owner has very little adherence to that specification. To analyzing social data, there are also dependencies on social network owners [17]. Consequently, collecting social data in a simple and efficient way is a research challenge.

D. Social Data Aggregation

Use of semantic Web in social networking perspective is very promising [18] [19]. Sharing social data in the Web benefits the users for connecting, communicating and managing relationship automated and efficient ways. FOAF ontology is well-recognized to represent relationship of people in the Web [20]. But, the specification is not generalized enough for automatic mapping (from user perspective) of different kinds of Social Networks to discover relationship. In the current applications, FOAF is used to specify people relationships in terms of community. But there is need of a mapping which may map or collect social data from different social networks in semantically meaningful way. Therefore, the proposed service also contains a FOAF engine which is being able to interpret generic social data in FOAF data set and map among different social data sources.

The main contribution in this paper are (1) a generalized and simplified model for on-line interactions to generate most of the interaction using on-line communication tools, (2) possible communication patterns discovered using process mining tools, and (3) a study that evaluates social strength from multiple perspectives. Social strength is discussed in [21].

III. Modeling On-line Interaction

On-line communication is mostly driven by different form of interactions (e.g., Email, SMS, Face-booking, Twittering).These interactions are instrumented by different kind of technological support and forms actionable tools for communication. In the paper, interaction models for on-line interactions provides the way of transforming different form of interactions in a unified format. Unified form of interaction is necessary for processing and analyzing the social data. Figure 1 depicts three main stages of modeling and generating on-line interactions. Generation of on-line interaction is important to support ASG framework by providing tremendous interactions logs. Consequently, it provides a uniform way of representing on-line interactions. For the analysis of communication history for generating social strength, a uniform representation of interaction is very important, because it provides a systematic way of creating data-sets of interaction logs in a unified manner (i.e., interactions from different sources are unified in a single data-source). The first stage is an interaction life cycle where each interaction is initiated by a user of

 ISBN: 978-1-61208-100-7

the communication service towards a contact or a group of contacts of the same or different services. The interaction pattern model is also designed to represent flow of activities in interactions. Format of interaction logs are discussed for capturing interactions from different sources for monitoring and analyzing communication history. Details of each of these steps are given below.

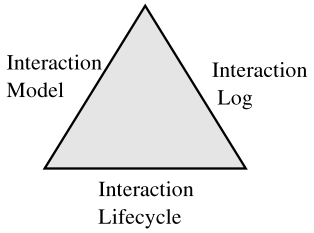


Fig. 1. On-line Interaction Building Blocks

A. Interaction Life-cycle

Figure 2 depicts the interaction life-cycle. The model simply generates one-to-one interaction, one-to-many interactions or bi-directional interactions which form conversation. To segregate conversations from interaction logs, co-relation among interactions need to be discovered. Due to simplicity, interaction is considered as a unique communication unit by ignoring conversations. This simplifies the counting of the usage of communication tools from communication history by providing a straight forward way. The interaction is initiated by the user of the service through a service client (e.g., Facebook's iPhone application). The service might be able to capture location data (with sensor's associated with application carrier) and time for instance to propagate the content (e.g., picture) via communication platform (e.g., Facebook) using a particular tool (e.g., Facebook/Photo-share). In addition, other examples are SMS, MMS, phone call, audio file sharing application, video sharing application, commenting, social tagging, tweets, or re-tweet, etc.

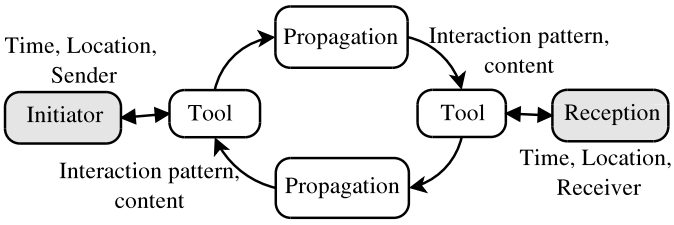


Fig. 2. On-line Interaction Life-cycle

B. Interaction Patterns

Interaction patterns provides the main steps to accomplish an interaction (e.g., initiation, tool selection, propagation, reception, etc) and may differ from platform to platform and on selection of tools. Some interactions are unidirectional which never return reply or response. On the other hand, some interactions implicitly need confirmation for inviting colleagues to participate in meeting. For example, public tweet and @receiver communication tools of Twitter application are quite different. Public tweets are initiated for a group of users while @receiver is for a particular user and it might form as a conversation at the end[22]. And eventually, it becomes more different in comparison with different tools of different communication services for instance, in Facebook and Twitter. Thus, to the aim of discovering communication patterns, we propose simplified on-line interaction patterns which comply with interaction life-cycle, vice-versa able to represent most of the communication tools by this. Figure

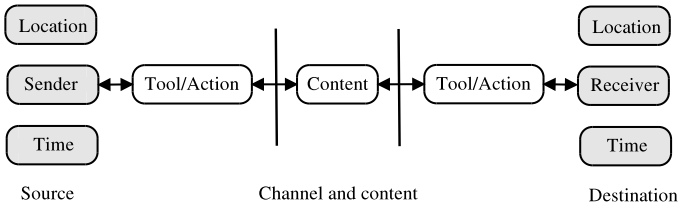


Fig. 3. Interaction pattern.

3 show communication pattern models which can be easily interpreted to form Twitter message and Facebook message. The pattern could be also used for interpreting phone calls, SMS and MMS. Interaction patterns tell the users about their communication habit. Therefore, it is important to identify communication patterns discovered from interaction logs for the recommending better communication tools and services to the specific users.

C. Interaction Log

Interaction logs contain communication history of the individuals. Table 1 shows the simple view of the interaction logs. Here we mention the basic properties/features for forming interactions. The table contains only interactions. For example, if a user replies to a corresponding message, the system considers the follow up replies as a unique interaction. Here, the sender contains information of message initiator, time and location. Action is name of the tools (SMS), platform can be Facebook or Twitter, and receiver contains the information of receiver, time and location. From Table 1, it is easier to measure frequency of interactions, platforms which helps to measure social strength.

D. Overview of the System Architecture

The proposed interaction model can be applied in the traditional aggregator services for simplifying analysis of social data. In general, aggregator services are responsible for collecting interactions from different communication services. To analyze these interactions, the aggregator services have to rely on some unified format. The proposed model is used to representation all the interactions in unified format. Figure 4 depicts overall system architecture where the proposed interaction model can be easily deployed. In the figure, different communication services are connected with the social data

 ISBN: 978-1-61208-100-7

TABLE I
SAMPLE INTERACTION LOGS

Sender	Tool	Platform	Tool	Receiver
Kare:Time:Lulea	SMS	Telecom operator	Reply	Peter:Time:Stockholm
Johan:Time:Lulea	Phone call	Telecom operator	Receive	Peter:Time:Stockholm
Josef:Time:Lulea	Tweet	Twitter Web Application	Reply	Peter:Time:Sttorckholm
Juwel:Time:Stockholm	SMS	Telecom operator	Reply	Peter:Time:Stockholm

aggregator service. The aggregator service interprets all the collected interactions in the unified format. The social network miner uses the unified datasets to analyze interaction for mining communication patterns or supporting other third party services. Finally, client applications can be designed to get various services from pervasive service miner.

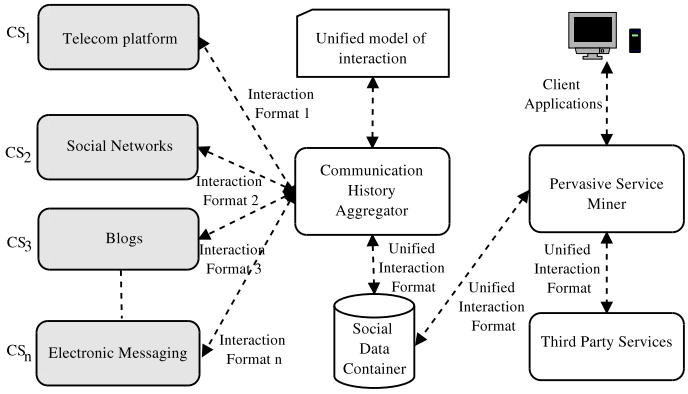


Fig. 4. System Architecture

IV. EVALUATION

A. Experimental Data Collection

An experiment was set-up to evaluate the interaction model. In the experiment, an interaction simulator is implemented to generate data-sets randomly. The simulator follows Weibull distribution to generate realistic interactions. Although there are Web-APIs (e.g., Facebook API) available for accessing real interaction data, but most of the cases, they are protected by strict privacy policy and user-specific authentication key [23]. Therefore, to run the experiment with fair amount of social data, a large number of user's interactions in different services are needed. However, these amount data could be managed by the communication services owner via publicly accessible interface. In practice, both of these options are not feasible to build the data-set considering the fact of privacy policy. Therefore, the pre-generated interaction data-set is used to run the experiment. For communication pattern mining purposes, these interactions are transformed to MXML form to make it machine readable [24]. Moreover, all these interactions are generated on the basis of proposed interaction model. Table 1, represents sample data which forms frequency log.

B. Implementation

Figure 5 depicts the ASG Powered Twitter Client for the Android Operating Systems mobile devices. The application is implemented for accessing the interaction data. The application shows how different kind of groups can be formed with social strength of relationship (high, medium or low strength), preferences and contexts. Therefore, information overload can be controlled by offering presence information in different groups based on social strength. In the current prototype, it is possible to generate dynamic group by aggregating presence information from different communication services, where the interaction model helps to represent the presence information in an unified view.

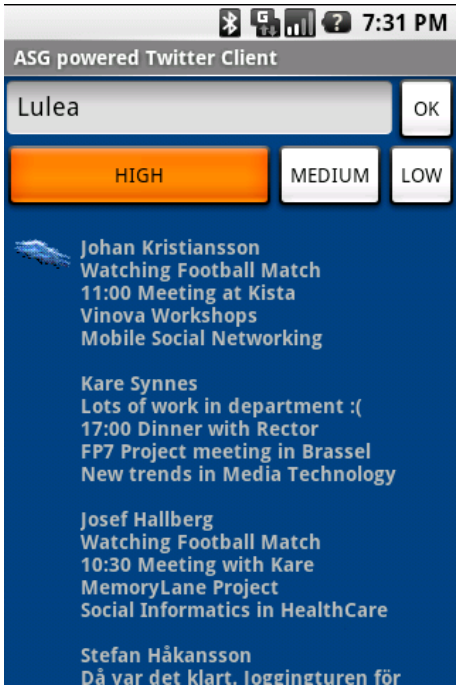


Fig. 5. Prototype of ASG Powered Twitter Client

C. Experiments using PROM process mining tool

The unified interaction model simplifies social data processing and analysis. For instance, using a process mining tool called PROM [24], the proposed model provides good result for discovering communication patterns. In PROM, each interaction in the interaction log is considered as a process instance in MXML form. The process is considered as all of the possible communication path. The experiment is run to discover communication pattern of the user from the process instance.

In PROM, MXML formed interaction logs are used to discover communication pattern. An example of MXML

based interaction logs is illustrated in Figure 6, which contains three events. Each event represents a process instance. In the first instance, the user Juwel sent a Facebook message from the location Gotenberg at time 2009-1-1-T0:0:10 to his contact Johan. Johan received the message at location Lund in time-period 2009-1-1-30.

The process work-flow miner module of the PROM tool is used to discover user communication patterns. The tool is capable to illustrate the discovered communication patterns analyzing interaction logs. This technique helps to automatically initiate new interaction towards the user in his/her preferable communication tool discovered using PROM. Referring back to Peter scenario mentioned in Section 1, where he wants to know his communication patterns towards all his contacts, now he is able to view a set of communication patterns using his previous communication history.

```
<ProcessInstance>
  <AuditTrailEntry>
    <WorkflowModelElement>Initiate
    </WorkflowModelElement>
    <EventType >Facebook\Message
    </EventType>
    <Timestamp>1900-1-1T:0:0:10
    </Timestamp>
    <Originator>Juwel</Originator>
    <Location>Gotenberg</Location>
  </AuditTrailEntry>
  <AuditTrailEntry>
    <WorkflowModelElement>Propagate
    </WorkflowModelElement>
    <EventType >complete
    </EventType>
    <Timestamp>1900-1-1T:0:0:20
    </Timestamp>
    <Originator>Facebook</Originator>
  </AuditTrailEntry>
  <AuditTrailEntry>
    <WorkflowModelElement>Receive
    </WorkflowModelElement>
    <EventType >Facebook/ACT</EventType>
    <Timestamp>1900-1-1T:0:0:30
    </Timestamp>
    <Originator>Johan</Originator>
    <Location>Lund</Location>
  </AuditTrailEntry>
</ProcessInstance>
```

Fig. 6. Process Instances in MXML format

D. Results

One of the main contributions of this paper is to provide a new simplified approach of Web data aggregation, representation and analysis for discovering communication patterns or measuring social strength. However, collecting Web data can be easily tackled by using vendor specific Web-APIs while the real challenge still remain in merging such data for forming an aggregated social data-set, which can be used for mining and analysis purposes. To be able to accomplish this challenge, this paper contributed a generalized interaction model to capture all social data in a unified format. To test the interaction model, different type of interactions from different communication sources are interpreted in MXML format. The interaction model is able to discover user-specific communication patterns (tested in PROM tools). To reach that result, different test operations were done on different MXML files with different number of interaction instances and found that, all identical patterns are discovered. Therefore, the on-line interaction model performs significantly well in generalizing on-line interactions and discovering identical communication patterns. Figure 7 elicits a discovered communication pattern using PROM process mining tool. Here, the user uses Facebook Messaging, Photo-sharing and Photo-commenting tools for communication purposes.

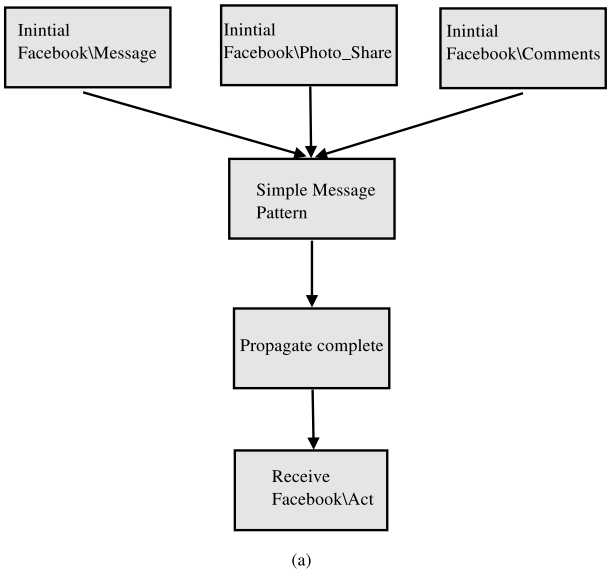


Fig. 7. Discovered Interaction Patterns

V. DISCUSSION

Analysis of on-line social interaction is one of the key elements of measuring social strength [1] [2] and [8]. Therefore, a generalized model is very important to represent on-line interaction. The proposed mode is able to aggregating interactions generated by different communication and social networking services and represents those in a unified format for further analysis and data mining purposes. For instance, the model is quite effective discovering communication patterns. In the previous section, the MXML form of interactions is evaluated for discovering communication patterns. However,

 ISBN: 978-1-61208-100-7

considering only on-line interactions limit the reliability as large portion of daily interactions (driven by face-to-face) in the physical world which are not detected by the system. Therefore, there need to be some tools to fetch physical world communication. We consider this as a challenging problem, which remain as future work. Another limitation of the work is that the experiment is run based on emulated social data with the assumption that social data are fetched from the cloud (i.e., contact lists and interaction logs). However, the proposed interaction model is able to interpret most of the interactions in the social media, for example in Facebook, Twitter and even interactions using Mobile phones.

VI. Conclusions

Due to immersive growth of Web-based communication services, there is an emergent need of integrating communication services by user-specific social data (i.e., social strength) to predict user's preferences and to filter information-overloads (i.e., micro-blogs, news feed, etc). The proposed on-line interaction model provides a simpler way of aggregating on-line interactions from large number of communication services in a unified manner. Initial prototype infers that the model is useful to work in the real-life, although there are some challenges (e.g., capturing real-world interactions, preserving user privacy) need to be addressed before the practical use of such model.

VII. Acknowledgments

This work was funded by the Research Area Multimedia Technologies of Ericsson Research, where Stefan Hkansson has been vital for the direction and execution of the research. The work was also supported by the Centre for Distance-spanning Technology (CDT) at Lule University of Technology and by the European regional (Mal-2) project Satin-II, funded by the Swedish Agency for Economic and Regional Growth, the County Administrative Board of Norrbotten, Norrbotten County Council, and the City of Lulea.

References

[1] A. Ankolekar, G. Szabo, Y. Luon, B. A. Huberman, D. Wilkinson, and F. Wu, "Friendlee: a mobile application for your social life," in *MobileHCI '09: Proceedings of the 11th International Conference on Human-Computer Interaction with Mobile Devices and Services*. New York, NY, USA: ACM, 2009, pp. 1–4.

[2] R. Grob, M. Kuhn, R. Wattenhofer, and M. Wirz, "Cluestr: mobile social networking for enhanced group communication," in *GROUP '09: Proceedings of the ACM 2009 international conference on Supporting group work*. New York, NY, USA: ACM, 2009, pp. 81–90.

[3] J. Rana, J. Kristiansson, H. Josef, and K. Synnes, "Challenges for mobile social networking applications," in *EuropeComm 2009: Proceedings of the First International ICST Conference Communications Infrastructure. Systems and Applications in Europe*. Communications Infrastructure. Systems and Applications in Europe: Springer Berlin Heidelberg, 2009, pp. 275–285.

[4] S. Poslad, *Ubiquitous Computing: Smart Devices, Environments and Interactions*, 1st ed. Wiley, May 2009. [Online]. Available: http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/0470035609

[5] N. Banerjee, D. Chakraborty, K. Dasgupta, S. Mittal, S. Nagar, and Saguna, "R-u-in? - exploiting rich presence and converged communications for next-generation activity-oriented social networking," in *MDM '09*. Washington, DC, USA: IEEE Computer Society, 2009, pp. 222–231.

[6] J. Hallberg, M. B. Norberg, J. Kristiansson, K. Synnes, and C. Nugent, "Creating dynamic groups using context-awareness," in *MUM '07: Proceedings of the 6th international conference on Mobile and ubiquitous multimedia*. New York, NY, USA: ACM, 2007, pp. 42–49.

[7] J. Rana, J. Kristiansson, J. Hallberg, and K. Synnes, "An architecture for mobile social networking applications," in *CICSYN '09: Proceedings of the 2009 First International Conference on Computational Intelligence, Communication Systems and Networks*. Washington, DC, USA: IEEE Computer Society, 2009, pp. 241–246.

[8] B. Aleman-Meza, M. Nagarajan, L. Ding, A. Sheth, I. B. Arpinar, A. Joshi, and T. Finin, "Scalable semantic analytics on social networks for addressing the problem of conflict of interest detection," *ACM Trans. Web*, vol. 2, no. 1, pp. 1–29, 2008.

[9] J. Häkkilä and J. Mäntyjärvi, "Developing design guidelines for context-aware mobile applications," in *Mobility '06: Proceedings of the 3rd international conference on Mobile technology, applications & systems*. New York, NY, USA: ACM, 2006, p. 24.

[10] G. Broll, E. Rukzio, M. Paolucci, M. Wagner, A. Schmidt, and H. Hussmann, "Perci: Pervasive service interaction with the internet of things," *IEEE Internet Computing*, vol. 13, pp. 74–81, 2009.

[11] M. Samulowitz, F. Michahelles, and C. Linnhoff-Popien, "Adaptive interaction for enabling pervasive services," in *MobiDe '01: Proceedings of the 2nd ACM international workshop on Data engineering for wireless and mobile access*. New York, NY, USA: ACM, 2001, pp. 20–26.

[12] A. T. Campbell, S. B. Eisenman, K. Fodor, N. D. Lane, H. Lu, E. Miluzzo, M. Musolesi, R. A. Peterson, and X. Zheng, "Cenceme: Injecting sensing presence into social network applications using mobile phones (demo abstract)," 2009.

[13] J. Seo and W. B. Croft, "Blog site search using resource selection," in *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management*. New York, NY, USA: ACM, 2008, pp. 1053–1062.

[14] X. Li, L. Guo, and Y. E. Zhao, "Tag-based social interest discovery," in *WWW '08: Proceeding of the 17th international conference on World Wide Web*. New York, NY, USA: ACM, 2008, pp. 675–684.

[15] F. Fuchs, D. Berndl, and G. Treu, "Towards entity-centric wide-area context discovery," *Mobile Data Management, IEEE International Conference on*, vol. 0, pp. 388–392, 2007.

[16] J. Heer and D. Boyd, "Vizster: visualizing online social networks," in *Information Visualization, 2005. INFOVIS 2005. IEEE Symposium on*, Oct. 2005, pp. 32–39.

[17] B. A. Huberman, D. M. Romero, and F. Wu, "Social networks that matter: Twitter under the microscope," *ArXiv e-prints*, December 2008. [Online]. Available: http://arxiv.org/abs/0812.1045

[18] P. Mika, "Social networks and the semantic web," in *WI '04: Proceedings of the 2004 IEEE/WIC/ACM International Conference on Web Intelligence*. Washington, DC, USA: IEEE Computer Society, 2004, pp. 285–291.

[19] U. Bojārs, J. G. Breslin, V. Peristeras, G. Tummarello, and S. Decker, "Interlinking the social web with semantics," *IEEE Intelligent Systems*, vol. 23, no. 3, pp. 29–40, 2008.

[20] L. Ding, L. Zhou, T. Finin, and A. Joshi, "How the semantic web is being used: An analysis of foaf documents," in *In Proceedings of the 38th International Conference on System Sciences*, 2005.

[21] J. Rana, J. Kristiansson, and K. Synnes, "Enriching and simplifying communication by social prioritization," *Social Network Analysis and Mining, International Conference on Advances in*, vol. 0, pp. 336–340, 2010.

[22] M. Cheong and V. Lee, "Integrating web-based intelligence retrieval and decision-making from the twitter trends knowledge base," in *SWSM '09: Proceeding of the 2nd ACM workshop on Social web search and mining*. New York, NY, USA: ACM, 2009, pp. 1–8.

[23] E.-A. Baatarjav, R. Dantu, and S. Phithakkitnukoon, "Privacy management for facebook," in *ICISS '08: Proceedings of the 4th International Conference on Information Systems Security*. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 273–286.

[24] A. Rozinat, M. T. Wynn, W. M. P. van der Aalst, A. H. M. ter Hofstede, and C. J. Fidge, "Workflow simulation for operational decision support," *Data Knowl. Eng.*, vol. 68, no. 9, pp. 834–850, 2009.

 ISBN: 978-1-61208-100-7

Using Context-aware Workflows for Failure Management in a Smart Factory

Matthias Wieland, Frank Leymann, Michael Schäfer
Institute of Architecture of Application Systems
Universität Stuttgart, Germany
{firstname.lastname}@informatik.uni-stuttgart.de

Dominik Lucke, Carmen Constantinescu, EngelbertWestkämper
Institute of Industrial Engineering and Management (IFF)
Universität Stuttgart, Germany
{firstname.lastname}@iff.uni-stuttgart.de

Abstract—**In factories many processes are executed in parallel. The manufacturing processes are managed by Manufacturing Execution Systems. In the case of machine failures these systems provide only rudimentary or no support to the workers or shop-floor managers. As a consequence the failures have to be fixed as fast as possible for being able to continue manufacturing processes. For such cases context-aware workflows can be used to support the workers and to coordinate the work that has to be done for repairing purposes. In the Nexus Project we introduced the concept of context-aware workflows and context integration processes to be able to implement all kinds of processes going on in a smart environment. As a case study we modeled a failure management process as a workflow and executed it in a factory. Furthermore, we show the concepts behind this kind of workflows: the context integration processes and the context-aware human tasks. Finally, end user applications for the interaction of the workers with the workflow are presented. For that we developed an application concept providing a mobile solution for workers and a web-based solution for an office environment. The main contribution of this paper is to show how to implement such a failure management process as a context-aware workflow.**

Keywords-context-awareness; context models; workflow systems; human tasks; production environments; humane system design; mobile applications; ubiquitous computing

I. INTRODUCTION

Workflow technology as automation method gained major influence in many enterprises and within the software industry. It provides methods and corresponding products to support modeling, execution, and management of business processes that have been carried out manually and through a diversity of non-integrated systems before. The integration of these non-integrated systems is possible by employing workflow management systems. With modeling their processes as workflows (instead of hard-coding them into their systems), the enterprises were enabled to change their processes more easily; they became more flexible and adaptive, to survive within a more and more dynamic market [1]. However, there are major application domains where processes are not supported by workflow technology yet like the domain of manufacturing environments. Here, the processes are planned and executed using Manufacturing Execution Systems (MES) or Enterprise Resource Planning systems (ERP). A difficulty of these technical processes is that they are crossing the boundary to the physical world.

A commonly used context definition by Dey [2] states that context is "any information that can be used to characterize the situation of entities". In this definition entities are real world objects. So, a fundamental difference between traditional business processes and manufacturing processes is that within manufacturing processes context data has to be used that is captured based on these real world objects. Since miniature sensor technology and wireless communication becomes ubiquitous, it is possible to capture and observe more and more of these real world events. This leads to the idea of a Smart Factory, which allows the (automatic) collection and distribution of information, knowledge, and tasks to all work places based on physical events [3].

The Smart Factory approach developed at the Universität Stuttgart, defines a Smart Factory as a factory that is context-aware and assists people and machines in execution of their tasks by using context [4]. Here, context-aware means that the system can take context information as the location and status of a factory object, like the position of a tool and its working state (in work, damaged, etc.) into consideration.

Failure management in manufacturing is a very important topic. Only a rudimentary support of failure management is available in the standard manufacturing management systems, e.g., a notification of a failure via SMS or email is possible. This is caused by the high complexity and the diversity of failures. To improve failure management we show in this paper our approach of supporting failure management by coordinating and supporting the repair process with context-aware workflows. Furthermore, we show how humans can be integrated into context-aware workflows since human work is needed in the repair process. Based on this real-world scenario of failure management we explain the concepts we used to implement a prototype with a set of context specific end user interfaces.

The paper is structured as follows. Section 2 describes related work and fundamentals. Section 3 describes the basis for the system which is the *extended context data model*. In Section 4 the main concepts of *context-aware workflows* the context integration processes and the *context-aware human tasks* are introduced. Section 5 presents the *failure management process*. Finally, in Section 6 the prototype implementation is presented by explaining the *architecture* for execution of the context-aware workflow and the *end user applications* for human interaction with the context-aware human tasks. As well, *measurements* of the prototype implementation are presented. Section 7 concludes the paper and points out future work.

 ISBN: 978-1-61208-100-7

II. Related Work and Fundamentals

The Workflow Management Coalition defines workflows as "the computerized facilitation or automation of a business process, in whole or part" [5]. That means the execution of workflows is fully computerized. In contrast, business processes describe the progress of work done in an enterprise in a form that is not directly executable by workflow systems. However the idea of ubiquitous and pervasive computing and the idea of context-aware and mobile computing open up new possibilities for using business processes [6]. The benefit of integrating context-aware computing into workflows is to extend the application area of workflows from business processes to manufacturing processes. This is a big advantage because now a company can support the production with the same technology already used in the back office or administration [7]. In [8], we have already discussed all the main technical issues about context-aware workflows. In this paper however, we want to enhance the concepts for allowing the integration of humans into context-aware workflows. This allows building ubiquitous work environments using the concept of context-aware human tasks.

Business processes are normally modeled using graphical notations such as the Business Process Modeling Notation (BPMN) [9] or the event-driven process chains (EPCs) [10]. None of them contain explicit elements for modeling context-aware processes. For the automated execution with a computer, the processes must be expressed as workflows using a workflow execution language like YAWL [11] or WS-BPEL [12]. Since WS-BPEL (or BPEL for short) became the standard in this area, it is used as basis for the development and execution of our context-aware tasks.

Note that BPEL already supports integration of humans by the BPEL4People extension [13] or vendor specific solutions. For this purpose, a task management system distributes work tasks to human participants. There are no systems available that allow the modeling of context-aware tasks in a standard BPEL environment. However, some research approaches exist in this area: xBPEL [14] is a BPEL extension for modeling mobile participants in workflows. The PerCollab system executes xBPEL and allows integration of people into BPEL workflows without constraining the users to their desktop PC. The WHAM System [15] supports mobile workforce and applications in workflow environments based on IBM products. In contrast to these solutions our system builds on top of the already used WS-BPEL4People standard and can be used as addition on top of a running workflow system in parallel to normal business-workflows and conventional human tasks.

III. Extended Context Data Model

The Nexus approach [16] federates various spatial context models, which are stored in the so-called context servers. For the integration of context information from different context models a standardized context schema is needed. In the Nexus project we decided to build an object-oriented data model for that purpose. The basic objects are represented in the *Nexus Standard Classes*. Based on this standard schema extensions can be defined for different application domains [16]. This has the advantage that the context data can be shared between different applications even from different domains. For the failure management in the Smart Factory we have defined such an extended context schema [17], [18] of which an excerpt is shown in Figure 1.

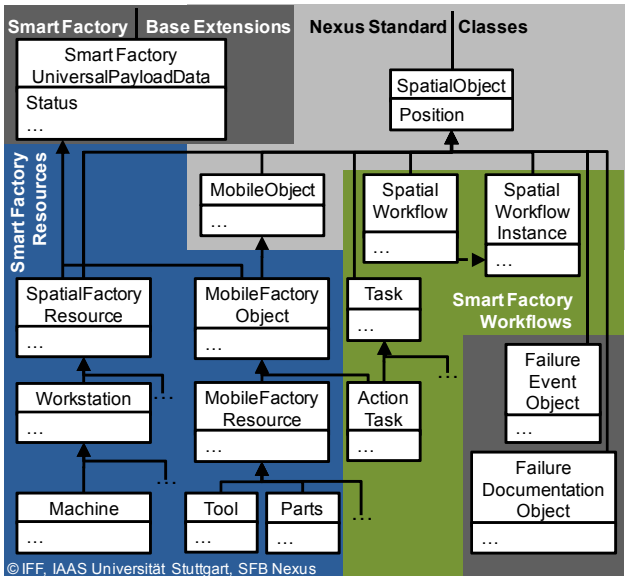


Figure 1 Excerpt of the context data model for integration of workflow and production context based on the Nexus standard classes

The Smart Factory Data Model consists of several packages. Following packages are important in the failure management scenario: *SmartFactory Resources* for representing production resources such as machines and tools. *Smart Factory Workflows* for the representation of processes and tasks context in the Smart Factory. The package *SmartFactory Base Extensions* contains a set of virtual objects representing documentation or detected failures. Further packages not in the focus of this paper refer to production orders, products, sensors and actors. In order to ensure that objects of the Smart Factory have attributes describing location (position attribute) and condition (status attribute), they inherit these attributes from the Nexus Standard and Smart Factory Base Extension classes. The classes from *Smart Factory Base Extension* package contain additionally attributes for target locations (location specification), execution dates of tasks or document references to rarely needed data such as manuals.

Like Figure 1 shows in detail the package *Smart Factory Resources* represents the real-world objects of a factory, such as buildings, production segments, machinery and equipment (tools, fixtures and testing equipment), storage resources and transportation means. The package *Smart Factory Workflows* contains an object for holding the context of a workflow model (Spatial Workflow), e.g., the failure management workflow with an area where it can be applied. In case a failure occurs an instance of a workflow (Spatial Workflow Instance) is created at the location of the handled failure event. The instance also creates human tasks (Task, Action Task) for the manual work that has to be done to repair the

 ISBN: 978-1-61208-100-7

failure. The concepts and functionality behind context-aware workflows is described in the following section.

IV. CONCEPTS OF CONTEXT-AWARE WORKFLOWS

In this section the needed concepts for realization of context-aware workflows with human interaction are shortly presented. For further details of the basic concepts other papers are already published that present them in more detail [8], [6], and [19]. Hence, in this section only the most important concepts are explained for understanding the realization of failure management workflows.

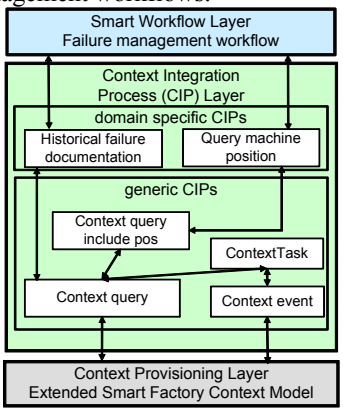


Figure 2 Context Integration Process realizing context-aware workflows

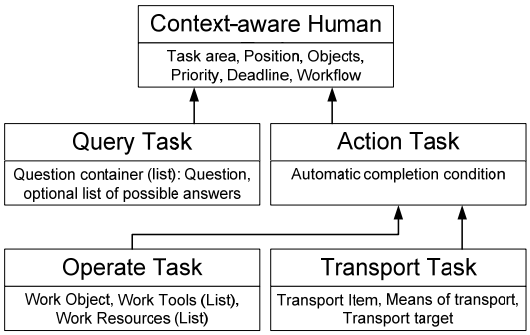


Figure 3 Modeled types of context-aware tasks

A. Workflow Concepts

We called the special kind of Workflows that are able to control smart environments "Smart Workflows" [8]. They form the top most layer of the context-aware workflow concept (see Figure 2) and represent the application logic. As most workflows they are orchestrating the control flow between automated programs as workflow activities. The smart workflows use Context Integration Processes (CIP) to access the context models in a domain specific application optimized way. This allows that the applications can be modeled by domain experts without technical knowledge about the context management system. Inside the Context Integration Process Layer the CIPs form a hierarchy of workflows calling each other as sub processes. The generic CIPs are used by every smart workflow and are the gateway to the context management system. Furthermore, the context task CIPs provide additional functionality to create and manage human tasks in a workflow system and providing the task context in the context model.

B. Context-aware Human Tasks

The tasks are the way to integrate human work into the automated workflows. This is needed in order to complete manual work with workflows. We have defined different types of tasks for different kinds of work. The most important ones are shown in Figure 3. The tasks are classes in the context model and have different attributes, which are inherited from the super classes. As a consequence the management of the tasks is split up between the workflow system and the context management system. The workflow system manages the state and execution of the tasks using a task list and a human task service. The context management system manages the context as for example the location of a task or the area where it is visible. By intersecting both the workflow data such as roles of people that are able to execute the task and context data like the location of people, the amount of potential tasks per user can be reduced. Using that concept only tasks are shown on the work list that are e.g., in a radius of 100m near a user of the role worker. This helps users to select appropriate tasks faster than normally, because the list contains fewer tasks and is context-aware. Furthermore, the context-aware tasks have another advantage. They are connected to real world objects, which are the reason for the task. This information is important for the user, so he does not have to search for the object but can locate it in a map. Furthermore, this has also advantages for automation: An action task can be automatically completed by monitoring the real word object. A transport task is completed when the object reaches its target location. An operate task is completed automatically when the object of work switches to the state OK again. This automatic completion saves time, because a user does not have to go to a terminal to enter that he has finished the task. He only has to interact with the real world objects. Only for the question task a user has to input data into the computer for answering a question about the real world object he has examined. For example a question could be: "Please conduct a sight check of tool 132". The answer can be the remaining usage time of the tool. By that context can be acquired or verified where no sensor is available or the sensor is not precise enough.

For implementing all the task types each is modeled as a CIP that has application logic for creating the task in the workflow system and in parallel insert it into the context model. After the task is completed it is deleted out of the context model. Most important, while the task is executed the contexts of its connected real world objects are monitored and on occurrence of the completion condition the CIP completes the task using the privileges of the Context management system.

V. FAILURE MANAGEMENT PROCESS

In a factory many processes are executed in parallel, e.g., different manufacturing and maintenance processes. The processes are executed normally with Manufacturing Execution Systems (MES).

 ISBN: 978-1-61208-100-7

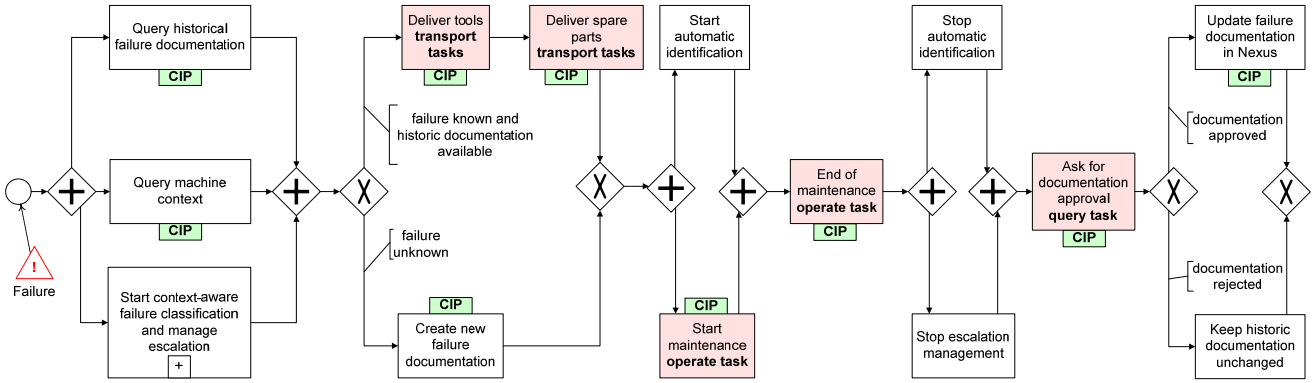


Figure 4 Context-aware workflow for handling a failure in the Smart Factory

In case of machine failures only rudimentary or no support is provided to the workers or shop-floor managers by these systems. In situation the workers are usually on their own to fix the failure as fast as possible for being able to continue manufacturing processes. For such cases context-aware workflows can be used to support the workers and coordinate the work that has to be done for repairing the failure efficiently. In the Nexus Project we introduced the concept of context-aware workflows and context integration processes to be able to implement all kinds of processes going on in a smart environment. As a case study we modeled this failure management process as a workflow and executed it in a real factory. The main contribution of this paper is to explain how to implement such a process as context-aware workflow.

Figure 4 shows a simplified version of the process for handling a failure. The complete BPEL code is accessible on our web-site [20]. First the workflow calls Context-Integration Processes (CIP) as sub-workflows to gather all needed context information about the failure: the context of the machine (location) and historical documentation about similar failures to analyze how and with what tools and parts the failure was repaired the last lime. In parallel the failure is classified in its seriousness and effects on the complete production. Then a branch is made whether a solution to the failure is already known or not. If a solution is available in historic documentations the needed tools and spare parts for the maintenance have to be transported to the machine with the failure. This is done using *transport tasks*. When the resources reach the target location the transport tasks are completed automatically by a context event. After that the maintenance is started by creating an *operate task*. In parallel to the task execution a documentation system monitors automatically, which tools and spare parts are actually used by the worker. Technical, the identification of tools and spare parts is implemented by using RFID tags. Due to harsh environmental conditions (e.g. metal influence, dirt and water) a object specific selection and mounting of the RFID tags for a reliable identification is required. In the prototype we did this for important tools (e.g. different types of screwdrivers, ladders, transport boxes, …). The *operate task* monitors the machine switching from failure state to OK state and completes then automatically. At the end of the workflow the automatic gathered documentation is presented in a *query task* to the worker for approval. Now the worker can decide whether he wants to edit and afterwards publish the automatically recorded documentation in the context model for improvement of future failure handling or if he wants the documentation to be deleted for privacy reasons.

VI. IMPLEMENTATION

We have created a prototypical implementation of the described concepts and modeled the failure management workflow. For that we first designed the extended Smart Factory context data model. Then we developed the task models and deployed all to the context servers. After this we modeled the context-aware workflows with BPEL (CIPs and failure management workflow) for coordinating the work in the Smart Factory. We also developed an application concept with different applications (mobile and web-based) for accessing the work tasks in distinct environments (on the move and in office). With this tools available context data stored in any context server can be visualized in order to get an overview of the factory.

A. Architecture

The needed architecture for executing the context-aware workflows consists of the following important parts (shown in Figure 5): The context management system implemented by the Nexus Federation. It manages the context model in a distributed scalable system. The second part is the workflow system that consists of the Workflow Management System (in our implementation it is an Oracle BPEL engine), the human task management service and an application server for hosting web-applications. The applications are used as user interface for the employees to conduct the human tasks and also for managing the executed workflows. For the management the standard Oracle BPEL dashboard can be used. But the human task application has to be extended to take the context of the tasks into account. Therefore, we developed a new web-client that shows all tasks and connected real world objects on a map (Figure 8). By that the context of the tasks is visible. Furthermore, a mobile application for accessing the tasks on a mobile device is required (Figures 6, 7). This results from the environment of the smart workflows. Normal workflows are used in business environments where all

 ISBN: 978-1-61208-100-7

users have a desktop computer and are not walking around all the time. In a smart environment, such as a smart factory, the users are mobile at all times. As a consequence they have to carry their task list with them. Furthermore, the mobile device knows the current location of the user and can filter the tasks by proximity. We implemented both applications and tested them in a Smart Factory on their usefulness and performance (see Section D).

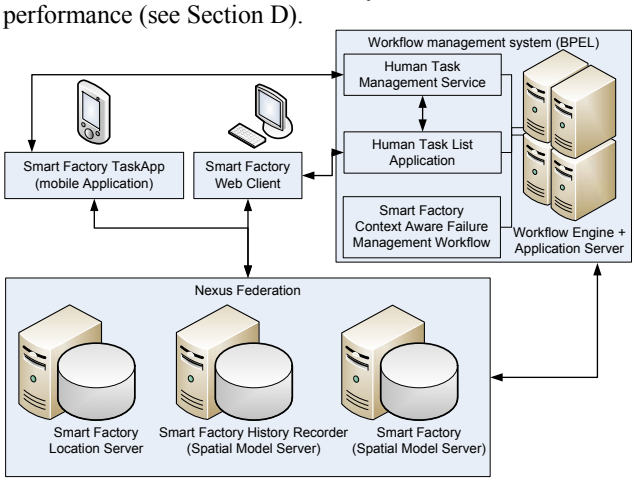


Figure 5 Architecture for execution of context-aware workflows

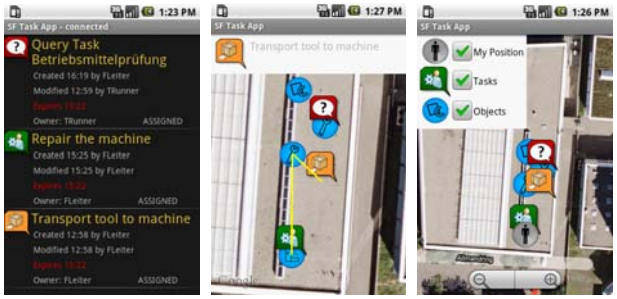


Figure 6 Mobile task application showing a work list, work map with selected transport task and the map filter

B. *Mobile Client TaskApp*

Figures 6 and 7 show screenshots of the mobile application on a Google android phone. This mobile application can be used anywhere on a mobile phone supporting android 1.6 or higher. The tasks are presented in two ways, firstly as usual in a task list (Fig. 6 left side) and secondly what is new (on a mobile device based on BPEL4People) as a task map (Fig. 6 right side). This means the context of the tasks is used to display them in a map in addition with the objects that are related to the task. Furthermore, the position of the mobile phone is shown as user location for a better orientation of the user. For example the transport task has a dependency to the tool that has to be transported (see Fig. 6 middle) and a connection to the target machines where the target of the transport is located.

Figure 7 shows the details view of the different task types. Here all the connected real world objects are listed and can be selected to be shown on a map. Furthermore, the task can be claimed and completed here by one fingertip. This calls the workflow system and changes the state of the task. It is also shown if the task is already executed and who is working on it. Also the task can be shown on the work map for finding the shortest way. The colors represent the priority of the tasks (1-5: green over orange to red).

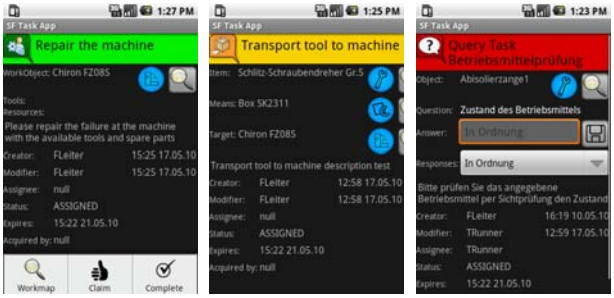


Figure 7 Mobile task application showing task details

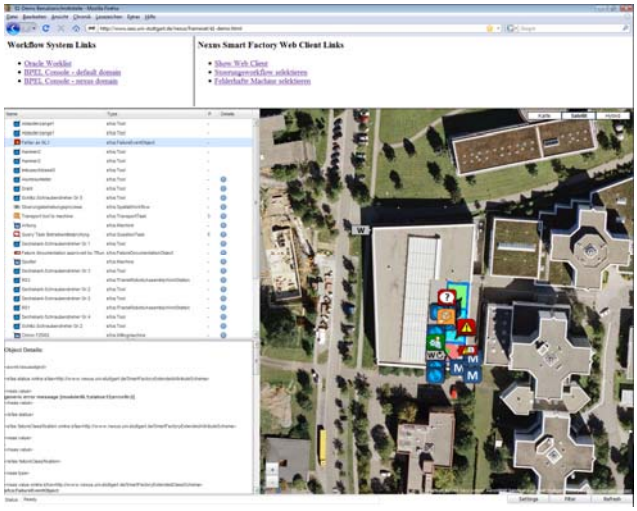


Figure 8 Web client showing context objects in the Smart Factory

C. *Web Client integration Portal*

The Web Portal Application is needed in parallel to the mobile client for presentation of the whole factory context in an office scenario for analysis. Here a much bigger screen size is available and the overview of the whole context can be achieved. Furthermore, nothing has to be installed on the client computers. Figure 8 shows a screenshot of the web client showing all the objects in the Smart Factory on a map. An operate task is shown in orange. Again, the color represents the priority of a task. It is important to note that no workers are tracked for privacy reasons and hence are not visualized in the map. The symbol for a person represents a human task, which can be executed by different workers. Also the re-localization of the task is only done indirectly e.g., on location change of the real world object connected to the task. A further advantage of the web client is that it provides links to the workflow system for presenting workflow models for starting the workflow and workflow instances for observing their execution state. Furthermore, links to the workflow task management system for working on the tasks.

 ISBN: 978-1-61208-100-7

And last but not least links to factory information systems with additional information about the resource, e.g., manuals or 3D representations. For focusing on one aspect a filter for object types is provided. For instance only failure events and machines can be shown as an overview of the current state of the machines in the factory. Or only tasks can be shown if the web-client is used as task map.

D. Runtime Measurements

We conducted measurements of both the web-client and the mobile client to find out the limitations of our solution. Table 1 shows the time (in seconds) for processing the data 1, 10, 50 and 100 tasks on the mobile client running on two different mobile Android phones and on the web client. Processing means to query the data from all sources and integrate it. The visualization is done in parallel and is not measured here. The result is that the system is able to handle at least up to 100 tasks in parallel. The refresh time on the mobile clients however is very high because of the limited processing power on the mobile devices. But for the user of the applications it is not noticeable that the updates take long, because the update is executed in a background thread. The fast execution of the web-client results from the fact, that only the data from the context model is used in the web client and no workflow data has to be integrated.

TABLE I. RESULTS OF THE RUNTIME MEASUREMENTS

Type of Client and Device Brand	Amount of Tasks			
	1	*10*	*50*	*100*
Web-Client - Firefox	0,08s	0,11s	0,36s	0,45s
Mobile-Client-HTC G1	1,63s	8,35s	37,82s	82,27s
Mobile-Client-Motorola Milestone	1,50s	5,10s	28,58s	53,49s

VII. CONCLUSION AND FUTURE WORK

In this paper we showed how to support the failure management in production environments by implementing this process as a context-aware workflow. The workflow uses context-aware human tasks for integration of the human workers executing the process.

Future work is to find a method how to allow the worker to work with the mobile applications using speech recognition to keep his hands free. In addition, performance improvements will be done on the mobile client for achieving a faster update of the task list for being able to handle more tasks in parallel.

ACKNOWLEDGMENT

This interdisciplinary research is funded by the Collaborative Research Center *Nexus: Spatial World Models for Mobile Context-Aware Applications* (grant SFB 627).

REFERENCES

[1] F. Leymann and D. Roller, *EnglischProduction Workflow - Concepts and Techniques*. PTR Prentice Hall, Januar 2000.

[2] A. K. Dey, "Understanding and Using Context," *Personal and Ubiquitous Computing*, vol. 5, no. 1, pp. 4–7, 2001.

[3] E. Westkämper, L. Jendoubi, M. Eissele, and T. Ertl, "Smart Factory - Bridging the Gap Between Digital Planning and Reality," *Manufacturing Systems*, vol. 35, no. 4, pp. 307–314, 2006.

[4] D. Lucke, C. Constaninescu, and E. Westkämper, "Smart Factory. A Step towards the Next Generation of Manufacturing," *CIRP Manufacturing Systems*, vol. 42, 2008.

[5] WfMC, "Workflow Management Coalition." [Online]. Available: http://www.wfmc.org/

[6] M. Wieland, O. Kopp, D. Nicklas, and F. Leymann, "Towards Context-Aware Workflows," in *CAiSE´07 Proceedings of the Workshops and Doctoral Consortium Vol.2, Trondheim, Norway, June 11-15th, 2007*, B. Pernici and J. A. Gulla, Eds. Tapir Acasemic Press, June 2007, Workshop Paper, pp. 577–591.

[7] M. Wieland, F. Leymann, L. Jendoubi, D. Nicklas, and F. Dürr, "Task-orientierte Anwendungen in einer Smart Factory," in *Mobile Informationssysteme - Potentiale, Hindernisse, Einsatz. Proceedings MMS´06*, ser. Lecture Notes in Informatics (LNI), T. Kirste, B. König-Ries, K. Pousttchi, and K. Turowski, Eds., vol. P-76. Bonn: Gesellschaft für Informatik, February 2006, Conference Paper, pp. 139–143.

[8] M. Wieland, P. Kaczmarczyk, and D. Nicklas, "Context Integration for Smart Workflows," in *Proceedings of the Sixth Annual IEEE International Conference on Pervasive Computing and Communications*. Hong Kong: IEEE computer society, March 2008, Conference Paper, pp. 239–242.

[9] OMG, "Business Process Modeling Notation, V1.1," http://www.omg.org/spec/BPMN/1.1/PDF, 2008.

[10] A.-W. Scheer, O. Thomas, and O. Adam, "Process Modeling using Event-Driven Process Chains," in *Process-Aware Information Systems*, M. Dumas, W. Aalst, and A. Hofstede, Eds. Wiley & Sons, 2005, pp. 119–146.

[11] W. M. P. van der Aalst and A. H. M. ter Hofstede, "YAWL: Yet Another Workflow Language," *Information Systems*, vol. 30, no. 4, pp. 245–275, 2005.

[12] *Web Services Business Process Execution Language Version 2.0*, http://docs.oasis-open.org/wsbpel/2.0/, OASIS Std.

[13] *WS-BPEL Extension for People*, http://www.oasis-open.org/committees/bpel4people/charter.php, OASIS Std.

[14] D. Chakraborty and H. Lei, "Pervasive Enablement of Business Processes," in *Proc. of the Second IEEE Intl. Conf. on Pervasive Computing and Communications (PerCom 2004), 14-17 March 2004, Orlando, FL, USA*, 2004, pp. 87–100.

[15] J. Jing, K. E. Huff, B. Hurwitz, H. Sinha, B. Robinson, and M. Feblowitz, "WHAM: Supporting Mobile Workforce and Applications in Workflow Environments," in *RIDE*, 2000, pp. 31–38.

[16] M. Großmann, M. Bauer, N. Hönle, U.-P. Käppeler, D. Nicklas, and T. Schwarz, "Efficiently Managing Context Information for Large-Scale Scenarios," in *Proc. of the Third IEEE Intl. Conf. on Pervasive Computing and Communications*, 2005, pp. 331 – 340.

[17] D. Lucke, C. Constantinescu, and E. Westkämper, "Context Data Model, the Backbone of a Smart Factory," in *Sustainable Development of Manufacturing Systems: Proceedings of the 42nd CIRP Conference on Manufacturing Systems*, Grenoble, France, June 3-5 2009.

[18] ——, "Fabrikdatenmodell für kontextbezogene Anwendungen: Ein Datenmodell für kontextbezogene Fabrikanwendungen in der Smart Factory," *wt Werkstattstechnik online*, vol. 99, no. 3, pp. 106–110, 2009.

[19] M. Wieland, D. Nicklas, and F. Leymann, "Managing Technical Processes Using Smart Workflows," in *Towards a Service-Based Internet, First European Conference, ServiceWave 2008, Madrid, Spain, December 10-13, 2008. Proceedings*, ser. Lecture Notes in Computer Science, P. Maehoenen, K. Pohl, and T. Priol, Eds., vol. 5377. Springer Verlag, December 2008, Conference Paper, pp. 287–298.

[20] "Failure management workflow," http://www.iaas.uni-stuttgart.de/nexus/demo/FailureManagement.bpel.

(all links have been followed on July 6th, 2010)

 ISBN: 978-1-61208-100-7

Tangible Applications for Regular Objects: An End-User Model for Pervasive Computing at Home

Spyros Lalis[1], Jarosław Domaszewicz[2], Aleksander Pruszkowski[2], Tomasz Paczesny[2], Mikko Ala-Louko[3], Markus Taumberger[3], Giorgis Georgakoudis[1], Kostas Lekkas[1]

[1] CERETETH & University of Thessaly Volos, Greece {lalis, ggeorgak, kolekkas}@inf.uth.gr

[2] Institute of Telecommunications Warsaw University of Technology Warsaw, Poland {domaszew, apruszko, t.paczesny}@tele.pw.edu.pl

[3] VTT Technical Research Centre of Finland Oulu, Finland {mikko.ala-louko, markus.taumberger}@vtt.fi

***Abstract*—This paper describes an end-user model for a domestic pervasive computing platform formed by regular home objects. The platform does not rely on pre-planned infrastructure; instead, it exploits objects that are already available in the home and exposes their joint sensing, actuating and computing capabilities to home automation applications. We advocate an incremental process of the platform formation and introduce tangible, object-like artifacts for representing important platform functions. One of those artifacts, the application pill, is a tiny object with a minimal user interface, used to carry the application, as well as to start and stop its execution and provide hints about its operational status. We also emphasize streamlining the user's interaction with the platform. The user engages any UI-capable object of his choice to configure applications, while applications issue notifications and alerts exploiting whichever available objects can be used for that purpose. Finally, the paper briefly describes an actual implementation of the presented end-user model.**

Keywords—***Sensor and actuator networks, ubiquitous and pervasive computing, smart homes, system and application management, user interaction, tangible interfaces.***

I. INTRODUCTION

The continuous technological developments in the area of embedded computing and networking make it possible to digitally augment regular home objects with computing, sensing, actuation and communication capabilities, making them not only smart but also capable of cooperation with each other. In the near future, the household is likely to be populated with a host of such objects, ranging from usual appliances like a refrigerator, an electric kettle, or a TV, to infrastructural elements like doors, windows, and lamps, down to small devices such as temperature sensors, smoke detectors, and motion sensors.

Significant potential for advanced functionality can be created by transforming a collection of digitally-augmented regular objects into an open pervasive computing platform that allows home automation applications to exploit the different sensing and actuation capabilities of participating objects in a combined way. For instance, one application could employ temperature sensors and smoke detectors to infer the presence of fire. Another application could save on the electricity bill by controlling the operation of lights and appliances based on the user's activity and demand-response offers of the electric utility. Yet another application could double check that a window is not left open unintentionally while the thermostat setpoint for the heater located in the same room is above the outside temperature.

A multi-object computing platform, as described above, can be implemented by letting the nodes embedded in objects expose the local sensing and actuating capabilities in a suitable way, as well as communicate with each other to provide other middleware-level services to the applications. However, the underlying software and hardware is only a part of the challenge. An equally important aspect is to consider how the end-user perceives and interacts with such a platform. By no means should such a platform be yet another user-attention hungry technology, introducing complex or awkward processes of installation, configuration and administration. This is absolutely crucial if one wishes for it to be embraced by the general public.

This paper describes an end-user model for multi-object computing platforms based on regular digitally-augmented home objects. In the spirit of ubiquitous computing [1], our work is based on the premise that the platform should require the end-user to expend as little mental energy (and be bothered with explicit manual input and intervention) as possible. The main contributions of the end-user model are as follows. First, we advocate a low-profile, incremental process of platform formation. Second, we introduce tangible, object-like artifacts, such as the *community key* and the *application pill*, to represent important platform entities and functions. Third, we streamline the conventional user interaction with the platform, for the cases that cannot be handled using these special objects. The paper also describes a concrete implementation which is being pursued in the POBICOS project [2]. Notably, the presented end-user model is largely platform-independent and could be realized using different combinations of networking, hardware and software technologies.

The rest of the paper is organized as follows. Section II gives an indicative scenario and lists the main elements of the envisioned multi-object pervasive computing platform.

 ISBN: 978-1-61208-100-7

Section III introduces the end-user model, with focus on the special tangible artifacts and the aspect of user interaction. Finally, Section IV outlines the implementation in the POBICOS project, and Section V discusses related work.

II. VISION

Our vision of how pervasive computing could be accomplished in the home based on regular objects is illustrated via the following scenario:

Maria and Peter decide to buy some new appliances. While browsing the stores they notice that some items have a "community-enabled" sticker. A salesperson explains that this is a new technology which makes it possible for regular objects to cooperate. The couple decides to buy a kitchen stove and a TV. They are also given a special community-enabled "key" object for free. At home, after reading the (surprisingly short) manual, they bring the key object close to the TV and press a button to register it with the platform. The process is repeated for the stove. As nothing fancy happens, the couple quickly forgets about this technology.

Weeks later Peter buys a cook book. He notices that it comes with a small community-enabled object labeled as "the new home safety application pill by CoolApps Ltd". He registers the pill object following the usual process, and then pushes a button on the pill to start the application.

One day Peter is baking a cake. He goes to the balcony to get some fresh air and stays there for a while. Suddenly, he hears a rather unusual alarm tone coming from the bedroom. As Peter enters the house, he sees a message on the TV screen informing him about a problem with the stove. He rushes back to the kitchen and is relieved to see that the stove turned itself off just before his cake was about to turn into coal. Peter recalls that some time ago Maria bought a new community-enabled enabled alarm clock for their bedroom and, fortunately, registered it with the platform.

Peter recalls that, according to the manual, the home safety application comes with some pre-set parameters that can be modified to customize its behavior. Peter uses the TV to browse these settings, and decides to change the default policy for alerts to enable the engagement of voice messages.

This simple scenario captures, to a large extent, several key elements of our vision. These are described in more detail in the following.

A. *Unplanned, incremental formation from regular objects*

The user forms the multi-object computing platform in an incremental way, by adding objects to it. This can be done at any point in time and without thinking about the objects' digital augmentation, a particular platform configuration, or a specific application. The user buys objects in order to employ them according to their natural functionality (a lamp is bought and placed at a particular location to light that area), not because they can contribute to the platform. Most often, the user is not even aware of the capabilities the object may provide to the platform. Contrary to a system that is engineered for a specific purpose, there is no a priori specified arrangement or reliance on infrastructure.

B. *Open, multi-application platform*

The user can add new and remove existing applications at any point in time. Multiple applications may co-exist and run concurrently, subject to the resource constraints of the objects that make up the platform. Like in conventional systems, applications are typically developed by third parties that are not affiliated with object manufacturers.

C. *Tangible artifacts for straightforward administration*

Special, object-like, physical artifacts are used to embody important platform entities and functions which the user should be aware of and to which the user should have immediate access. For instance the "key" is required to add and remove objects to/from the platform and the "application pill" is used to start/stop the execution of a particular application. Making special entities and functions tangible and representing each of them with a different physical object relieves mental ambiguities (as to which object should be used to perform a function) and simplifies interaction (a dedicated, single-function object can have a tuned interface compared to a general-purpose object that is loaded, perhaps even over-loaded, with several different functions).

D. *Streamlined user interaction*

Ideally, user interaction occurs solely via the tangible artifacts introduced for platform formation and application management. However, in practice, additional interaction is often needed: (i) to let applications notify or alert the user; (ii) to let the user configure applications. The former is a one-way communication towards the user; the expected user reaction is to act in the real world, not to interact with the platform or application. In the latter case, the user, not the platform, is in charge of the interaction, i.e., the user chooses when to engage in the interaction and which object to use to do the setup. Importantly, in both cases, the platform is self-contained, relying on the native interaction capabilities of regular objects that are already available in the home. There is no reliance on computer-like objects such as a PC, a PDA or a mobile phone. While such objects are allowed to participate in the platform, they are not required to support user interaction.

III. END-USER MODEL

Along the lines of Section II, we propose an end-user model that describes, in a more formal and structured way, how the user perceives and interacts with the platform. The model consists of (i) basic terminology, (ii) special objects the user must be aware of, and (iii) a generic interaction pattern for the more conventional aspects of user interaction with the platform. The model is presented in a canonical way, striving for a clean separation of entities, roles and functionalities. Relevant use cases are described with reference to the scenario given in Section II.

A. *Basic terminology*

1) Community-enabled object: a regular object that provides sensing, actuation, and computing capabilities to the platform. Community-enabled objects can be marked,

 ISBN: 978-1-61208-100-7

e.g., with a sticker, so that the user can distinguish them from objects that are not community-enabled.

2) Object community: a collection of community-enabled objects in a home participating in the same platform (Figure 1). An object community is formed by adding and removing community-enabled objects in an explicit yet dynamic fashion. It represents a well-defined scope in terms of security vis-à-vis objects that are not part of the community, as well as in terms of the operational range of applications that run in the community.

B. Tangible artifacts

1) The community key object: The key object is used to add and remove other objects to/from the community (it also generates and transfers security-related keys and credentials in the background). It is the first object that must be acquired and it is mandatory to form an object community and to control its membership. The prototypical user interface for the key object is (Figure 2a): (i) a keypad for entering the name and PIN for an object community; (ii) two buttons, for triggering the addition and removal of objects; (iii) a LED for indicating the status and result of the last action; and (iv) close range communication ability with other objects for exchanging data in a safe manner without requiring a shared secret.

Use case: Initializing the key object

Peter and Maria switch on the community key for the first time. The LED on the key turns red. They enter a name and a PIN of their choice for their object community. The LED turns green, indicating it is ready to be used.

Use case: Adding an object to the community

Maria switches on the key and enters the name and PIN of the object community. The LED on the key turns green. She brings the key close to a newly purchased, community-enabled alarm clock and presses the "add" button. The LED on the key blinks for a few seconds and then turns green. Maria successfully added an object to the community.

Use case: Removing an object from the community

Peter turns on the key and enters the name and PIN. He brings the key close to the alarm clock and presses the "remove" button. The LED on the key blinks for a few seconds and then turns green. Peter successfully removed the alarm clock from the community.

.

2) The application pill object: Each applications is packaged in a distinct community-enabled object, called the application pill. The pill serves as a deployment and control vehicle for the application: it is used to start/stop application execution in the community and provides basic status information about the operation of the application. The user conceptually identifies the pill which the application itself; in other words, for the user, the pill *is* the application.

Figure 1. An indicative object community: bidirectional arrows indicate community-enabled objects, the dashed line around objects indicates the boundary of the community.

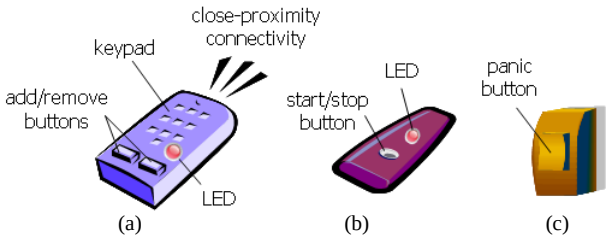


Figure 2. Tangible community artifacts: (a) key object, (b) application pill object, (c) panic button object.

Just like any other object, a pill must be added to the object community via the key before starting the application. The prototypical interface for the pill is (Figure 2b): (i) a push-button or switch to start/stop the application; and (ii) a LED for indicating the status of the application. One can imagine application pills being sold in stores and kiosks or given out for free bundled with products related to the application. At home, an application pill can be placed at location that allows for a casual periodic monitoring of its status LED.

Use case: Starting an application

Peter gets a home safety application pill object. He adds the pill to the community following the usual process. He then pushes the pill button to start the application. The LED on the pill blinks for several seconds and eventually turns green, indicating that the application is running.

 ISBN: 978-1-61208-100-7

Use case: Stopping an application

Peter wishes to stop the home safety application. He picks the application pill and depresses its button. The LED on the pill starts blinking. After a few seconds it turns off, indicating that the application has been stopped.

3) The panic button object: This object is used to forcefully terminate all applications running in the community at the push of a single button (Figure 2c). This could be required in case applications start behaving erratically or if the user feels uneasy about the overall platform behavior. The panic object can be likened to the master power switch in the electricity panel of a house or the reset button of a personal computer. As already mentioned, each application can be stopped by depressing the button of the respective pill. However, searching for and interacting with each individual pill can be quite stressful if the user is in a hurry. What's more important, stopping an application via the pill corresponds to a soft shutdown, under the control of the application program, which is clrealy undesirable when the application is malfunctioning.

Use case: Killing all applications

Maria notices an obscure object behavior without being able to infer what causes the problem. She quickly walks to the hallway and presses the button of the panic object attached on the main electricity panel. Soon, the weird behavior stops and the LEDs of all application pill objects turn off. Being more relaxed, Maria arranges for the technician to drop by the next day in order to get a closer look at the problem.

C. More convetional user interaction

A typical community will include several objects that do not have considerable user interface capabilities. In fact, objects like a window, a lamp or a motion detector do not have any proper user interface at all. On the other hand, objects like a TV or a digital frame can support (very) rich user interaction. Our approach is to rely on objects with advanced UI capabilities for configuring applications, while at the same time letting applications engage even simple to notify or alert the user. The key elements of our user interaction scheme are as follows.

1) Notifications & alerts: Applications may occasionally need to request the user's attention; this is achieved through notifications and alerts. The difference between the two is that notifications convey a verbal message whereas alerts do not carry such information (the user is responsible for finding out the cause of the alert). Notably, the actual form of notifications and alerts depends on the object that provides this function, each object supporting a different, perhaps complementary, flavor. Even simple objects like a lamp or a doorbell can contribute in this respect, especially for alerts, e.g., by blinking and respectively ringing at a certain alarming pattern. Of course, more complex objects with audiovisual capabilities, such as a TV or a radio, can be employed to make notifications via text or voice messages.

Use case: Being alerted/notified by an application

Peter's alarm clock in the bedroom and his wristwatch start beeping intensely (alerts). He walks into the living room and notices a message flashing on the TV screen warning him about a possible hazard with the stove (a notification).

2) Application setup: Once started via the pill, an application will ideally run without any interaction and rarely issue alerts or notifications. However, in the general case, some setup will be required, e.g., to change default thresholds or specify user preferences. For this purpose, each object that has a sufficiently powerful user interface is expected to allow the user to browse the list of applications running in the object community, and inspect or modify their settings. The setup process follows the native interaction style and look-and-feel of the object that provides the application setup function (consider differences between a remote-driven TV and a mobile phone with a touch-screen). Importantly, the setup can be accomplished with any UI-capable object, and the user can freely choose the one that suits him best.

Use case: Configuring an application

Peter decides to inspect the settings of the home safety application using the TV. He presses the "community function" button on the remote and browses the application list shown on the TV screen. He selects the home safety application, reviews its settings and decides to change the default policy for alerts. When the change is confirmed, Peter presses the "community function" button on the remote and the setup window disappears from the screen.

IV. IMPLEMENTATION

The presented platform concept and end-user model is currently being implemented in the POBICOS project [2]. Several ideas and features of POBICOS have their roots in ROVERS [3], which is a predecessor of this work. This section gives an overview of the POBICOS platform and briefly describes the system-level mechanisms used to achieve the end-user functionality described in the previous sections.

A. POBICOS platform overview

The POBICOS platform follows a middleware approach whereby each object supports a standard API. Objects may feature different middleware extensions depending on their sensing, actuating and computing capabilities. The application programming model is based on mobile code units, called micro-agents, which execute on top of a VM environment [4]. Each application typically consists of several cooperating micro-agents that spread in the community to exploit the capabilities of objects (Figure 3).

 ISBN: 978-1-61208-100-7

The POBICOS middleware is implemented on top of TinyOS v2 for the Imote2 from Crossbow using an eZ430-RF2480 ZigBee subsystem from Texas Instruments for the wireless communication between nodes. Regular objects are prototyped using Imotes. A generic adapter box with an Imote and a power level converter (Figure 4a) is used to POBICOS-enable objects and external systems via RS232.

B. Adding and removing objects

The key object, implemented on an Imote, maintains a registry with the addresses of all objects that are part of the community. The registry is updated when an object is added to or removed from the community. Registry updates can be propagated to the community in an asynchronous fashion by several objects (not just the key). To avoid inconsistencies, the key assigns to each update a monotonically increasing version number, enabling objects to detect duplicates and take into account membership changes in the right order.

The close-proximity communication between the key and the object being added/removed is implemented using the short-range mode of the 802.15.4 radio on the Imote (in principle, any near-field communication technology can be used for this purpose). When the add/remove button is pressed, the key establishes a connection with any object that is close-enough to respond, retrieves the object's address and performs the requested interaction (updating the registry as needed). Provided the range is small-enough, this guarantees that the proper object will be addressed but also that no other object can eavesdrop on the conversation.

Objects that are part of the same community encode and decode the messages exchanged between them over ZigBee using a community-wide encryption key. This is generated by the key object based on the name and PIN chosen by the user, and is transmitted to each object as part of the addition process. More details about the security approach and respective key and registry management protocols in POBICOS can be found in [5].

C. Starting and stopping applications

The application pill object is also implemented using an Imote. It contains the entire application code bundle, i.e., the binaries of all micro-agents of the application. The bundle is loaded on the Imote from a PC via the serial port. Pressing the application pill button leads to the instantiation of the micro-agents on the local or remote nodes (under the control of the application program). Depressing the button causes the micro-agents to be removed.

D. Notification and alerts

The POBICOS middleware features special instructions for notifying and alerting the user. Both types of instructions range from a high abstraction level such as "alert using whatever means possible" to more specific levels like "alert visually" or "alert by siren sound". Some objects support the notification and alert instructions in a manner that is compatible with their natural/native functionality. For example, in the current prototype an alarm can be raised by a variety of objects such as a TV (controlled via a POBICOS-enabled set-top-box; Figure 4c), a lamp or a beeper (both controlled via a POBICOS-enabled power plug; Figure 4b).

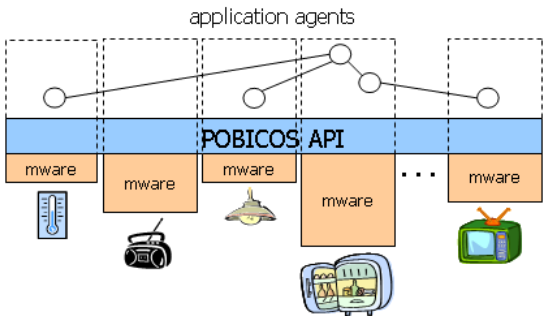


Figure 3. POBICOS platform concept: objects feature different middleware extensions based on their capabilities, application micro-agents are placed on available objects to exploit them.

When a micro-agent invokes an abstract instruction at runtime, it is mapped to a more specific instruction supported by its host. Thus, applications using abstract notification and alert instructions can exploit a wide range of objects, which may provide different specific instructions; this obviously comes at the price of having less control on the way the user will be actually notified/alerted. It is up to the programmer to decide what the meaningful tradeoff is for each occasion.

Last but not least, the POBICOS middleware provides a primitive for instantiating multiple copies of micro-agent on as many objects support the instruction(s) invoked by it. This allows an application to engage several objects at once for the purpose of alerts and notifications, thereby increasing the probability of catching user attention. It is important to note that this does not require any additional effort on behalf of the programmer.

E. Application setup

The setup functionality is implemented based on (i) a distributed protocol for fetching/updating the configuration settings of all currently running applications, and (ii) a user interface front-end for browsing and changing these settings. The first component is part of the middleware core running on the Imote. The second component is optional and needs to be developed separately for each object, depending on its UI capabilities. At this point, a front-end is available for the PC which communicates with the first middleware component on the Imote via the serial port. Proper UI front-ends are under development for a TV set-top-box and a mobile phone.

V. Related Work

In our model the user defines the operational and security scope of the community by adding and removing objects via the key object in a conscious and explicit way. The issue of knowing which devices belong to the same system scope also arises in mobile ad-hoc systems that allow wearable and portable devices to dynamically participate in a personal area network, e.g., the 2WEAR system [6] and the Spartan BodyNet [7]. These systems assume that devices have been a priori assigned a unique id indicating their owner and already

 ISBN: 978-1-61208-100-7

(a) (b) (c)

Figure 4. POBICOS-enabling real objects: (a) the RS232 adapter; (b) the power socket; (c) the TV set-top-box.

hold security keys that can be used to encrypt data and to perform a challenge-response scheme. In general, ad-hoc wireless technologies provide network-level association mechanisms based on a shared secret but do not specify how a device ends-up with this information. ZigBee implements its own security scheme but the transmission of keys from the coordinator to a new device that joins the network occurs via an ordinary open message exchange over wireless.

Significant research has been done on many aspects of user interaction in smart spaces/environments, e.g., [8] [9]. Of particular importance are alternative methods of input and control, e.g., see [10] for controlling devices via hand-based gestures, or [11] for supporting voice-based interaction with appliances. Our end-user model does rely on advanced UI. In fact, it is designed to exploit regular objects that are likely to be part of a household anyway, via modes and modalities the user is already familiar with. Moreover, it does not focus only on UI-capable objects but allows even simple objects to be engaged for notifying/alerting the user.

The vision of ubiquitous computing [1] is for the system to provide the desired functionality without distracting the user. Our model is conceived along these lines, requiring user intervention only for application configuration, which happens under user control; the user decides when to start such an interaction and is free to pick any UI-capable device for this purpose. Notification and alarms are introduced as first-class aspects of domestic computing since they play a key role in raising user awareness.

Tangible interfaces and the importance of having special objects dedicated to special functions have received a lot of attention in the HCI domain, see [12] for an overview. In the spirit of the community key proposed in our model, [7] discusses the use of a lock-shaped object to enable privileged functionality in a wearable system, while [13] proposes a wristwatch as an authentication device for ubiquitous service access. Also, the concept of the application pill has some similarities with the 2WEAR application wallet [6] and the personal server [14]. The former carries the code/state of applications that exploit I/O peripherals found in the personal area network. The latter serves as a personal data drive that can connect to applications running on nearby PCs to access/process this data. The main difference is that each pill is dedicated to a single application and features a minimal UI for controlling and monitoring its execution hence the pill is in fact a tangible representation of the application itself.

ACKNOWLEDGMENT

This work is funded by the 7th Framework Program of the European Community, project POBICOS, FP7-ICT-223984.

REFERENCES

[1] M. Weiser, “The Computer of the 21st Century”, in Scientific American, 256(3), 1991, pp. 78-89.

[2] POBICOS project web site. http://www.ict-pobicos.eu/

[3] J. Domaszewicz , M. Roj, A. Pruszkowski, M. Golanski, and K. Kacperski, “ROVERS: Pervasive Computing Platform for Heterogeneous Sensor-Actuator Networks”, Proc. WoWMoM 2006, pp. 615-620.

[4] A. Pruszkowski, T. Paczesny, and J. Domaszewicz, “From C to VM-targeted Executables: Techniques for Heterogeneous Sensor/Actuator Networks”, Proc. WISES 2010, in press.

[5] P. Tarvainen, M. Ala-Louko, M. Jaakola, I. Uusitalo, S. Lalis, T. Paczesny, M. Taumberger, and P. Savolainen, “Towards a Lightweight Security Solution for User-Friendly Management of Distributed Sensor Networks”, Proc. ruSMART 2009, pp. 97-109.

[6] S. Lalis, A. Savidis, A. Karypidis, J. Gutknecht, and C. Stephanides, “Towards Dynamic and Cooperative Multi-Device Personal Computing”, in The Disappearing Computer, LNCS 4500, 2007, Springer, pp. 182-204.

[7] K. Fishkin, K. Partridge, and S. Chatterjee, “Wireless User Interface Components for Personal Area Networks”, in IEEE Pervasive Computing, 1(4), 2002, pp. 49-55.

[8] B. Brumitt, B. Meyers, J. Krumm, A. Kern, and S. Shafer, “EasyLiving: Technologies for Intelligent Environments”, Proc. HUC 2000, pp. 97-119.

[9] B. Johanson, A. Fox, and T. Winograd, “The Interactive Workspaces Project: Experiences with Ubiquitous Computing Rooms”, in IEEE Pervasive Computing, 1(2), 2002, pp. 67-74.

[10] ASM. Rahman, M. Hossain, J. Parra, and A. El Saddik, “Motion-path based Gesture Interaction with Smart Home Services”, Proc. ACM MM 2009, pp. 761-764.

[11] T. Kostoulas, I. Mporas, T. Ganchev, N. Katsaounos, A. Lazaridis, S. Ntalampiras, and N. Fakotakis, “LOGOS: A Multimodal Dialogue System for Controlling Smart Appliances”, in New Directions in Intelligent Interactive Multimedia, SCI 142, 2008, Springer, pp. 585-594.

[12] K. Fishkin, “A Taxonomy for and Analysis of Tangible Interfaces”, in Personal and Ubiquitous Computing, 8(5), 2004, pp. 347-358.

[13] J. Al-Muhtadi, D. Mickunas, and R. Campbell, “Wearable Security Services”, Proc. ICDCS Workshops 2001, pp. 266-271.

[14] R. Want, T. Pering, G. Danneels, M. Kumar, M. Sundar, and J. Light, “The Personal Server: Changing the Way we Think about Ubiquitous Computing”, Proc. Ubicomp 2002, pp. 194 - 209.

 ISBN: 978-1-61208-100-7

Continuous Gesture Recognition for Resource Constrained Smart Objects

Bojan Milosevic, Elisabetta Farella, Luca Benini
DEIS - Dipartimento di Elettronica, Informatica e Sistemistica
Università di Bologna
Bologna, Italy
{bojan.milosevic, elisabetta.farella, luca.benini}@unibo.it

Abstract — **Tangible User Interfaces (TUIs) feature physical objects that people can manipulate to interact with smart spaces. Smart objects used as TUIs can further improve user experience by recognizing and coupling natural gestures to commands issued to the computing system. Hidden Markov Models (HMM) are a typical approach to recognize gestures sampled from inertial sensors. In this paper we implement a HMM-based continuous gesture recognition algorithm, optimized for low-power, low-cost microcontrollers without floating point unit. The proposed solution is validated on a set of gestures performed with the Smart Micrel Cube (SMCube), which embeds a 3-axis accelerometer and an 8-bit microcontroller. Through the paper we evaluate the implementation issues and describe the solutions adopted for gesture segmentation and for the fixed point HMM forward algorithm. Furthermore, we explore a multiuser scenario where up to 4 people share the same device. Results show that the proposed solution performs comparably to the standard forward algorithm and can be efficiently used for low cost smart objects.**

Keywords — Hidden Markov Models; Tangible Interfaces; Smart Objects; Gesture Recognition; Fixed Point.

I. INTRODUCTION

Tangible User Interfaces (TUIs) introduce physical, tangible objects that augment the real physical world by coupling digital information to everyday objects. The system interprets these devices as part of the interaction language. TUIs become the representatives of the user navigating in the environment and enable the exploitation of digital information directly with his/her hands. People, manipulating those devices, inspired by their physical affordance, can have a more direct access to functions mapped to different objects.

The effectiveness of a TUI can be enhanced if we use sensor augmented devices, which can provide a bridge between the physical and the digital world. Such *smart objects* may be able to recognize user gestures and improve human experience within interactive spaces. Furthermore, the opportunity to execute on-board a gesture recognition algorithm, without the need to end data streams from the local sensor to a central base station, results in *extended battery life*, improved system *scalability* and easier handling of *mobile TUIs*.

In this work we present an algorithm for segmentation and gesture recognition implemented on-board of a smart object, the *Smart Micrel Cube* [4], which embeds an 8 bit microcontroller, a digital accelerometer and a Bluetooth transceiver. This device can be used as the tangible interface of an interactive tabletop setup (like in the TANGerINE Project [4]) or as a mobile context-aware interface towards a smart, environment [6]. The algorithm detects the beginning and the end of motion segments and uses Hidden Markov Models to recognize the executed gesture. Unlike our implementation, gesture segmentation from a continuous stream of inertial data often relies on user collaboration (e.g., pushing a button wile executing a gesture [11]) or integrates information from various types of sensors (e.g., ultrasonic [19], microphones [21]). HMMs have been broadly applied to gesture recognition [11], [14], [15] but implementation on low performance devices are limited to high resource mobile-devices and 32 bit microcontrollers [2]. We focused on a resource constrained platform and addressed implementation issues for a 8 bit fixed point microcontroller.

The rest of the paper is organized as follow: Section II reports on related works and the sub-sequent Section III describes the system and the recognition procedure. Following, we characterize our implementation in Section IV; discuss experimental analysis and results in Section V and we conclude our paper in Section VI.

II. RELATED WORK

The use of TUIs has been proposed in many scenarios where users manipulate digital elements. This have been proved to be useful especially in applications for entertainment and education [16], exploration of virtual environments [9], media content creation and manipulation [10], [18]. The entertainment market is rapidly embracing tangible and gestural interfaces in several new scenarios, as for game-console controllers, such as the Wii, or for mobile devices and smart phones.

Smart objects with gesture recognition capabilities can enhance the expressiveness of TUIs. The MusicCube, for example, is a tangible interface used to play digital music like an MP3 player [5]. The cube is able to understand the face pointing upward and a set of simple gestures. This ability, together with a set of controls and buttons, is used to choose the desired playlist and to control music volume.

Gestures executed with natural hand and arm movements are variable in their spatial and temporal execution, requiring classifiers suited for temporal pattern recognition. Typical approaches include Dynamic Time Warping (DTW) [13],

 ISBN: 978-1-61208-100-7

Neural Networks [3], and Hidden Markov Models (HMMs). HMMs are often used in activity recognition since they tend to perform well with a wide range of sensor modalities and with temporal variations in gesture duration. They are also used successfully in other problem domains, such as speech recognition, for which they were initially developed [17]. Several variants of HMMs have been proposed to recognize inertial gestures: in [11] 5-state ergodic discrete HMMs are evaluated with the Viterbi algorithm to classify gestures performed with a handheld sensor device in several tasks (interaction with a TV, a presentation or a CAD environment). The work of Mantyla et al. [15] uses 7-states Left-to-Right models and the forward algorithm to classify actions performed with a mobile phone equipped with an accelerometer. Both implementations have similar performance and rely on a PC to execute all computations. In our work we are using low-power hardware without a floating point unit, so we implemented a fixed-point variant of the forward algorithm, presented in a previous work [22].

Using HMMs to classify gestures from a continuous stream of data brings another issue to solve: the recognition procedure needs to discriminate actually executed gestures from all the other arbitrary movements. Hoffman et al. [8] use a sensorized glove to recognize hand gestures: to segment the data stream they compute the velocity profile of the sampled accelerations and apply a threshold to identify the motion segments. In [7] a Gaussian model of the stationary state is used with a sliding window approach to find pauses in movements, which identify the beginning and the end of a gesture. Amft et al. [1] presented an algorithm to recognize arm activity during meal intake, with accelerometers placed on the arm and the wrist of the user. To segment gestures they use the Sliding Window and Bottom-up (SWAB) algorithm [12] and the angle of the lower arm as the segmentation feature. While those works have focused to develop recognition solutions, none of them deals with computation or memory limited devices. We found a similar solution implemented on a wristwatch device, using a 32 bit ARM microcontroller [2], but there are no works targeting low-cost, low-power 8 bit microcontrollers, such is the Atmel ATmega168 used in this work.

III. SYSTEM OVERVIEW

The smart object used in this work is a cube shaped artifact, the Smart Micrel Cube (SMCube) illustrated in Fig. 1. It embeds a low-cost, low-power 8-bit microcontroller (Atmel ATmega168), a Bluegiga WT12 Bluetooth transceiver, which supports Serial Port Profile (SPP) and a MEMS tri-axial accelerometer (STM LIS3LV02DQ) with a programmable full scale of 2g or 6g and digital output. The cube is powered through a 1000 mA/h, 4.2 V Li-ion battery. With this battery the cube reaches up to 10 hours of autonomy during normal operation.

The processing flow is illustrated in Fig. 2. Accelerations on the three axes are sampled at a rate of 31.75 Hz within the range of ±2g. The accelerometer represents the sampled data with a 16 bit integer value, and reaches a resolution of 1 mg.

In the pre-processing stage, sampled data are filtered with an averaging filter to eliminate high frequency noise. This filter computes the average value of the last 4 samples: this window

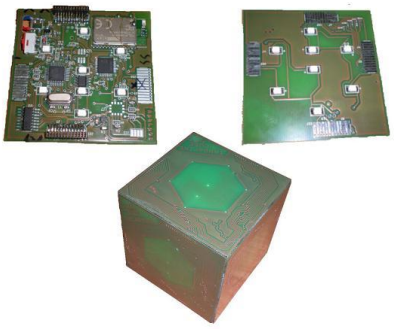


Figure 1. Smart Micrel Cube: on the top left the inner surface of the master face, with all the main components and on the top right the inner surface of the other faces

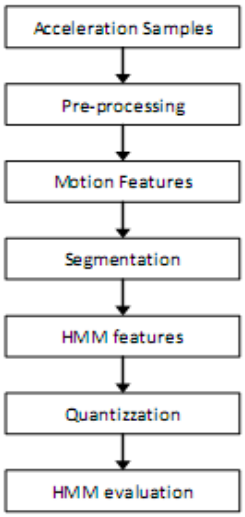


Figure 2. System processing blocks

length was chosen to use simple shift operations instead of divisions. An offset filter removes the stationary gravity acceleration, measured during a calibration phase which consist in sampling data when the device is placed still on a table.

The next two stages implement a motion detection algorithm, used for segmentation. For this purpose two features are used: delta, which consists in the difference between the actual sample and the mean of the last 4 samples, and the variance. A Finite State Machine (FSM) uses these features to find out if the device is in one of the 4 possible states: still on a table, still in user's hand, in movement, shaken. We asked users to hold still in a hand the device right before and after executing a gesture. In this way we segment as gestures only motion segments which start and end with this particular condition and have a limited duration.

Each acceleration vector of a gesture, $\{a_x, a_y, a_z\}$ is converted to equivalent spherical coordinates $\{\varphi, \theta, r\}$, as represented in Fig. 3. From those vectors, magnitude information r is discarded and the angles φ and θ are used to identify the direction of the movement performed, represented

 ISBN: 978-1-61208-100-7

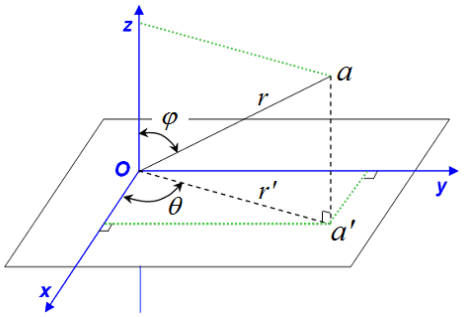


Figure 3. Spherical coordinates

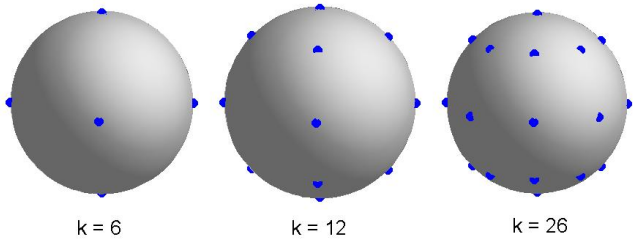


Figure 4. Codebook vectors for direction quatization.

as a point on a unitary sphere. In order to cluster this data for the discrete HMMs, a quantization algorithm is applied, with *k* vectors or codes in the codebook. In this case, codebook vectors are uniformly distributed points on the unitary sphere, as illustrated in Fig. 4. Since *k* must be determined empirically we decided to conduct tests to find a codebook size delivering a satisfying trade-off between results and algorithm complexity.

The last block of the processing chain is the HMM classifier. During the training phase we collected several gestures to build and validate models for each gesture. The training was implemented on a PC with Matlab and the HMMtoolbox, which uses the well-known Baum-Welch algorithm. We chose to use discrete HMMs, with 7-state Left-to-Right models for all gestures, according to the results of a preliminary exploration. For the on-line recognition a modified fixed-point version of the forward algorithm was implemented in both Matlab and C on the AVR platform.

IV. IMPLEMENTATION

The goal of this work is to implement the whole gesture recognition algorithm on board of the SMCube. The main tasks performed are the segmentation of gestures from a continuous data stream and the HMM-based gesture recognition.

A. Segmentation

The segmentation algorithm was implemented ad-hoc, and even if it was developed and optimized in this particular setup, it can be used in similar scenarios.

It is not possible to recognize gestures with only one accelerometer, if they are part of a larger continuous movement. To overcome this problem we added a limitation in gesture execution: users must hold still the device for few instants before and after a gesture. In this way, it is fundamental for the algorithm to identify when the device is still in user's hand and when a movement starts and ends. To evaluate the state of the device we compute the variance of the filtered signal and use a FSM. The variance uses sliding windows of 4 samples; it is calculated for each axis and then summed to have a total information of the intensity of the movement.

When the cube is placed still on a surface, the variance values observed are near zero. If a user holds the device in a hand, we measure a low and uniform variance, always within a limited interval. Since movements bring to higher values, it is possible to classify those conditions with empirically determined threshold values, combined with a few sample delays to avoid spurious transitions.

In this way, we segment every movement that is encapsulated between two states when the device is still in user's hand. Since all the used gestures have limited duration, we added a condition on the minimum and maximum time for the movements to be segmented as gestures. This helps to avoid many unwanted movements being identified as gestures.

Despite those conditions, this algorithm still identifies a lot of random movements as gestures, leading to many false positive results from the classifier, as shown later in Section V. To improve the performance and the usability of the device, we introduced a new operation mode, called *Assisted Segmentation*. In this mode the user performs a shake gesture to enable and disable the segmentation and recognition of all the other gestures. The shake gesture is recognized using the segmentation FSM and has a high accuracy: during our tests we obtained a correct classification ratio of 100% within 80 executions of the gesture and only one false positive.

By default, gesture recognition is disabled, and the user can move the device in any way (e.g. walking with the device or using the device as a pointer on an interactive table). When needed, the user performs a shake to "wake up" the smart object, activating the recognition algorithm and then executes the wanted gestures to interact with the system. During this time the user can pay attention to the movements performed, to avoid the recognition of random movements as gestures. Another shake disables the interaction capabilities of the device, and the user can move it freely.

This user-assisted segmentation technique increased the overall performance of the device, reducing drastically the number of false positive recognitions.

B. HMM Gesture Recognition

The main feature used for gesture recognition is the direction of the movement, represented by the direction of the acceleration vector, sampled at each frame. This information is obtained converting the 3D acceleration vector $\{a_x, a_y, a_z\}$ in spherical coordinates, and using only the two angles $\{\varphi, \theta\}$.

To efficiently compute the two angles of the acceleration vector we implemented an algorithm based on the CORDIC algorithm [20]. Using the notation in Fig. 3, this algorithm first estimates the phase θ and the magnitude r' of the complex number $(a_x + ia_y)$, then again estimates the angle φ, using r' and a_z. All computations are done with integer values, giving us a resolution of 1 degree and a maximum error of 2 degrees, which is acceptable since we are dealing with human motions and don't need higher accuracy.

Discrete HMMs are less computationally demanding than those operating on continuous observations, so they are the best choice in our case, since we are focusing on a limited resource implementation. As input to the discrete models we need to use

 ISBN: 978-1-61208-100-7

a discrete feature symbol to represent the directional information. The two angles calculated, that identify the arbitrary 3D orientation of a unitary vector, are quantized to the nearest vector of the codebook by a simple minimum distance classifier. In this way, the stream of two angles is converted in a stream of codebook indices, which is a suitable input to discrete HMMs. The number of vectors in the codebook was empirically determined and a codebook with k = 26 vectors uniformly distributed on the spherical surface resulted to be the best trade-off between quantization accuracy and processing complexity.

The off-line HMM training phase builds a model for each of the gestures to recognize, using sample instances of the gestures. We used the Baum-Welch algorithm, and initialized the training models with several random probability distributions. Among the resulting HMMs, those with the lowest training error were chosen.

To improve the model behavior, when dealing with input gestures that are slightly different from those used during training, we modified the symbol observation probability distribution (i.e. the observation matrix B in the discrete case). A model with a uniform observation matrix B_0 recognizes every gesture with a same low probability. We interpolated the trained models with the uniform one, by weighting the observation matrixes with a factor ε as given by the equation

$$B' = \varepsilon B + (1 - \varepsilon) B_0. \quad (1)$$

The optimal ε factor was empirically obtained and lies in the range [0.7 – 0.9].

The on-line recognition algorithm evaluates the executed gesture with all of the trained models, and selects the model with the highest probability. For this purpose we used a fixed point version of the forward algorithm, as introduced in [22]. This implementation deals with the lack of a division unit in the low power microcontroller embedded in the device, and proposes a different scaling procedure that uses shifts and a logarithmic representation of the probabilities. Our previous work compared the performance of this implementation against a standard floating point algorithm. The results showed that a 16 bit fixed point algorithm has the best trade-off between classification rate and computational complexity.

V. ANALYSIS AND RESULTS

A. Experimental Setup

For the validation of our algorithm we used a set of 7 gestures, illustrated in Fig. 5. All gestures are formed by natural movements, start and end with the user holding the cube in the same position and are executed on the vertical plane in front of the user, holding the device every time in the same orientation.

We collected gestures executed by four people, all male students with an age of 26 years. To build and validate the HMMs each user executed 80 instances of every gesture, during different days. Those gestures were continuously executed, with a few seconds of interval between two consecutive instances, and segmented with our algorithm.

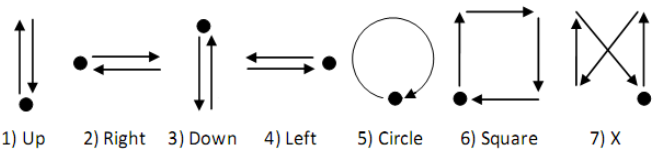


Figure 5. Gestures used for algorithm validation. The dots indicate the start and end position.

Gestures from three users were used for both modeling and validation, while gestures from the fourth user were used only to test the models. From each user we also collected several continuous streams, containing gestures and random movements, to simulate an actual usage of the device and test the overall recognition algorithm. The whole dataset was collected with the SMCube and sent via Bluetooth to a PC. No feedback from the device or the PC was given to the users during the execution of the gestures.

To easily test the performance, the algorithm was implemented in Matlab, taking care to simulate the computational constrains of the 8 bit microcontroller and using only integer computations with controlled variable size.

B. Test and Simulations

In the first place we used the collected dataset to train a set of HMMs for each user, using the floating point notation with double precision. Each model have been trained using 15 reference instances, 15 loops for the Baum-Welch training algorithm, and 10 initial random models. The floating point models were then converted in fixed point, represented only by 16 bit integers. Each user's models were validated with his/her own gestures, not used in the training phase, and with gestures from other users.

In Fig. 6 we present classification results in function of the interpolation factor ε. The circled points refer to the case when models trained by one user are validated with his own gestures; the triangular points when the models trained on a user are validated on the other user's gestures, and the squared points indicate when a global model is used on all users.

The classification performance is measured with the *Correct Classification Ratio* (CCR), defined as the ratio between the number of correctly classified instances and the total number of instances.

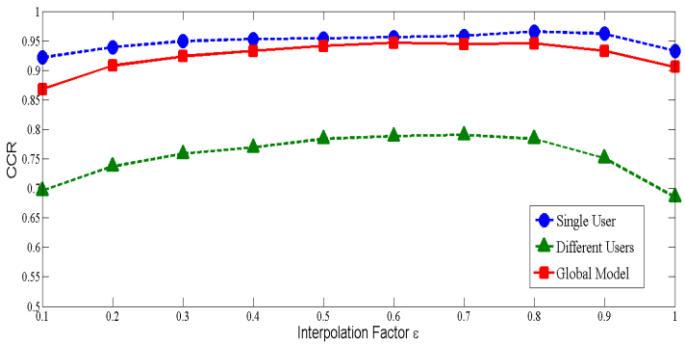


Figure 6. Average CCR for various ε values: the first line is in a single user scenario, the second is for the multi-user scenario and the third is for the global model

 ISBN: 978-1-61208-100-7

The results show how the classifier performs well in a single user scenario with recognition rates up to 99.7%. The algorithm has some limitations in a multi-user scenario, when recognizing gestures from a user with models trained by another user. We found that interpolating the trained models with a uniform one gave some advantages, and the best case is with an interpolation factor $\varepsilon = 0.8$. Table I shows the classification rates for the various users in this case; Tab. II shows the classification matrix in the best case and Tab. III in the worst case.

To overcome the limitations in the multi-user scenario, we put together all the gesture instances, regardless to the user who executed them, and build a global model for each gesture. These models were trained using 15 randomly chosen gestures and validated on 200 gestures from four users. The results of this case are presented in Fig. 6 with the squared points, where we can observe how in this case we have a performance comparable to the single user scenario, despite we are classifying gestures from all the users. We can also notice that the fixed point implementation has performances comparable to the floating point one, and our algorithm is suitable for low-performance smart objects. Table IV shows the correct classification matrix for the best case, with $\varepsilon = 0.6$.

With this global model we evaluated the overall algorithm performance, using the collected continuous streams of data which contain gestures and random movements. In this way we could test the segmentation and recognition algorithms together. Table V presents the results of this analysis. The automatic segmentation algorithm has good performance in recognition executed gestures, but gives also a lot of false positive results, identified by the insertions. The performance depends on what the user is doing and how the device is moved: long and continuous movements are easily rejected, but short movements, similar to gestures, trigger the recognition algorithm leading to a false positive. Deletions indicate how many times the algorithm misses a gesture, and this happens only if the gesture is executed too quickly or too slowly. Minimum and maximum duration times are derived from the collected dataset, and deletions may happen only in extremely short or long gestures.

To improve the device usability we proposed the assisted segmentation algorithm, which lets the user disable the recognition of gestures when not needed. Recognition rates for this algorithm are the same of the automatic one, since it uses the same HMMs, but in this case we have almost no insertions.

TABLE I. CLASSIFICATION RATES IN MULTI-USER SCENARIO

Training Set	Validation Set		
	User 1	*User 2*	*User 3*
User 1	99%	66.7%	85.9%
User 2	94%	92.8%	85.5%
User 3	93.3%	71.9%	99.7%

TABLE II. CCR BEST CASE: USER 3 USES HIS OWN MODEL

Performed gestures	Classified as						
	Up	*Right*	*Down*	*Left*	*Circle*	*Square*	*X*
Up	62	0	0	0	0	3	0
Right	0	65	0	0	0	0	0
Down	0	0	65	0	0	0	0
Left	0	0	0	65	0	0	0
Circle	0	0	0	0	65	0	0
Square	0	0	0	0	0	65	0
X	0	0	0	0	0	0	65

TABLE III. CCR WORST CASE: USER 2 USES MODEL TRAINED BY USER 3

Performed gestures	Classified as						
	Up	*Right*	*Down*	*Left*	*Circle*	*Square*	*X*
Up	33	0	0	0	1	28	3
Right	0	25	0	30	6	1	3
Down	12	1	43	1	0	7	1
Left	0	3	0	60	1	1	0
Circle	0	0	0	5	42	15	3
Square	0	0	0	0	13	50	2
X	3	0	15	3	2	3	39

TABLE IV. CCR FOR THE GLOBAL MODEL

Performed gestures	Classified as						
	Up	*Right*	*Down*	*Left*	*Circle*	*Square*	*X*
Up	194	0	3	0	2	1	0
Right	0	187	1	11	0	1	0
Down	1	0	199	0	0	0	0
Left	0	1	0	199	0	0	0
Circle	0	0	0	4	177	19	0
Square	0	0	0	0	13	187	0
X	3	0	12	0	1	2	182

TABLE V. GLOBAL MODEL CONTINUOUS RECOGNITION PERFORMANCES

	Auto Segmentation	*Assisted Segmentation*
Executed Gestures	83	78
Correctly Classified	71	62
Insertions	45	2
Deletions	6	5

C. Processing Perfomance Results

All the tasks needed for the gesture recognition algorithm were implemented on a PC in Matlab and on the ATmega168 microcontroller in C, using the AVR-GCC compiler and a 8 MHz clock. The Table VI presents the computational costs needed to perform the main operations at each frame. The Matlab implementation uses only fixed point operations on 16 bit integers, to simulate the embedded version and can be easily ported on other microcontrollers.

Each gesture model requires 3 matrices of 16 bit variables and with the implementation choices (7-state models with a 26 vector codebook) we need 462 Bytes to store each model. The microcontroller used has only 1 Kbyte of RAM, so we stored the models in the 16 KB of FLASH memory, used as program space. The entire application uses up to 12480 bytes of FLASH and 360 bytes of RAM memory.

We found a similar fixed-point implementation of HMMs in [2], which uses a 32-bit ARM7 microcontroller running at

 ISBN: 978-1-61208-100-7

65MHz. They recognize only 3 gestures and have recognition rates comparable to ours, but a shorter execution time (2.7 ms).

TABLE VI. COMPUTATIONAL COSTS

	ATmega168 (ms)	*PC-Matlab (ms)*
Preprocessing	0.03	0.37
Segmentation	0.20	0.40
Feature extraction	0.17	0.19
HMM (1 gesture)	0.73	0.83
HMM (7 gestures)	5.00	5.91
Total	6.13	7.70

VI. CONCLUSIONS AND FUTURE WORK

In this paper we presented an on-line segmentation and gesture recognition algorithm, implemented on a low-cost, low-power 8 bit microcontroller.

The automatic segmentation algorithm effectively finds executed gestures, but introduces also false positives. To address this issues, we introduced an *Assisted Segmentation* mode, which disables the gesture spotting algorithm when not needed, through the execution of a shake gesture.

We optimized a fixed point implementation of the HMM algorithm, which can be implemented on such low-performance microcontrollers, maintaining the same results as in a floating point case. The recognition algorithm can be used in a multi-user scenario, employing a global model trained with gestures from various users. It can be improved if the device is used only by one user, training the algorithm with only his/her own gestures.

In a future work we will explore the best ways to provide a feedback to the user, to see if the user can be trained to adapt his behavior and gestures in order to maximize the performance of the device.

VII. ACKNOWLEDGEMENT

Part of this work has been supported by SOFIA project funded under the European Artemis programme SP3 Smart environments and scalable digital service (Grant agreement: 100017) (www.sofia-project.eu).

REFERENCES

[1] O. Amft, H. Junker, and G. Troster, "*Detection of eating and drinking arm gestures using inertial body-worn sensors*", in Proceedings of the 9th IEEE international Symposium on Wearable Computer (ISWC), pp.160-163, 2005.

[2] R. Amstutz, O. Amft, B. French, A. Smailagic, D. Siewiorek, and G. Troster, "*Performance Analysis of an HMM-Based Gesture Recognition Using a Wristwatch Device*", In Proceedings of the 2009 international Conference on Computational Science and Engineering – Vol. 02, pp.303-309, 2009.

[3] G. Bailador, D. Roggen, G. Tröster, and G. Trivino, "*Real time gesture recognition using continuous time recurrent neural networks*", in 2nd Int. Conf. on Body Area Networks (BodyNets), Article n°15, 2007.

[4] S.Baraldi, L.Benini, O.Cafini, A.Del Bimbo, E.Farella, L.Landucci, A.Pieracci, and N.Torpei, "*Introducing TANGerINE: A Tangible Interactive Natural Environment*", in proceedings of ACM MultiMedia 2007, pp.831-834, 2007

[5] M. Bruns Alonso, and V. Keyson, "*MusicCube: a physical experience with digital music*", Personal Ubiquitous Comput., Vol.10, Issue 2-3, pp.163-165, 2006.

[6] A. Cayci, J. B. Gomes, A. Zanda, E. Menasalvas and S. Eibe, "*Situation-Aware Data Mining Service for Ubiquitous Environments*", Third International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies (UBICOMM), pp.135-140, 2009.

[7] G. S. Chambers, S. Venkatesh, G. A. West, and H. H. Bui, "*Segmentation of Intentional Human Gestures for Sports Video Annotation*", in Proceedings of the 10th international Multimedia Modelling Conference, pp.124-130, 2004.

[8] F. G. Hofmann, P. Heyer, and G. Hommel, "*Velocity Profile Based Recognition of Dynamic Gestures with Discrete Hidden Markov Models*". in Proceedings of the international Gesture Workshop on Gesture and Sign Language in Human-Computer Interaction, pp.81-95, 1997.

[9] C.-R. Huang, C.-S. Chen, and P.-C. Chung, "*Tangible photorealistic virtual museum*", IEEE Comput. Graph. Appl., Vol.25(1), pp. 15-17, 2005.

[10] S. Jordà, M. Kaltenbrunner, G. Geiger, and M. Alonso, "*The reacTable: a tangible tabletop musical instrument and collaborative workbench*", in Proceedings of the International Conference on Computer Graphics and Interactive Techniques, Article n°91, 2006.

[11] J. Kela, P. Korpipää, J. Mäntyjärvi, S. Kallio, G. Savino, L. Jozzo, and D. Marca, "*Accelerometer-based gesture control for a design environment*". Personal Ubiquitous Comput., Vol.10 (5), pp. 285-299, 2006.

[12] E. Keogh, S. Chu, D. Hart, and M. Pazzani, "*An online algorithm for segmenting time series*", in Proceedings of hte 2001 IEEE International Conference on Data Mining, pp.289-296, 2001.

[13] L. Kim, H. Cho, S. H. Park, and M. Han, "*A tangible user interface with multimodal feedback*", in Proceedings of the 12th international conference on Human-computer interaction, pp. 94-103, 2007.

[14] C. Lee and Y. Xu, "*Online, Interactive Learning of Gestures for Human/Robot Interfaces*", 1996 IEEE International Conference on Robotics and Automation, pp. 2982-2987, 1996.

[15] V.-M. Mantyla, J. Mantyjarvi, T. Seppanen, and E. Tuulari, "*Hand gesture recognition of a mobile device user*", IEEE International Conference on MultiMedia and Expo, Vol: 1 (c), pp. 281-284, 2000.

[16] C. O'Malley and D. Stanton Fraser, "*Literature Review in Learning with Tangible Technologies*", Technical report, url: http://www.telearn.org/open-archive/browse?resource=298_v1, last access on 25/05/2010.

[17] L. R. Rabiner, "*A tutorial on hidden Markov models and selected applications in speech recognition*", in Proceedings of the IEEE, Vol.77(2), pp.267-296, 1989.

[18] B. Schiettecatte and J. Vanderdonckt, "*AudioCubes: a distributed cube tangible interface based on interaction range for sound design*", in Proceedings of the 2nd international Conference on Tangible and Embedded interaction (TEI), pp.3-10, 2008.

[19] T. Stiefmeier, G. Ogris, H. Junker, P. Lukowicz, and G. Tröster, "*Combining Motion Sensors and Ultrasonic Hands Tracking for Continuous Activity Recognition in a Maintenance Scenario"*, 10th IEEE International Symposium on Wearable Computers (ISWC), pp. 97-104, 2006.

[20] J. Volder, "*The CORDIC computing technique*", IRE Trans. Electron. Comput., pp. 257-261, 1959.

[21] J. A. Ward, P. Lukowicz, G. Troster, and T. E. Starner, "*Activity Recognition of Assembly Tasks Using Body-Worn Microphones and Accelerometers*". IEEE Trans. Pattern Anal. Mach. Intell. Vol.28, pp. 1553-1567, 2006.

[22] P. Zappi, B. Milosevic, E. Farella and L. Benini, "*Hidden Markov Model based gesture recognition on low-cost, low-power Tangible User Interfaces*", Entertainment Computing, Volume 1 (2), 75-84, 2009.

 ISBN: 978-1-61208-100-7

SIREN: Mediated Informal Communication for Serendipity

Nikolaos Batalas, Hester Bruikman, Dominika Turzynska, Vanesa Vakili, Natalia Voynarovskaya, Panos Markopoulos

Department of Industrial Design,
Eindhoven University of Technology
Den Dolech 2, 5612AZ Eindhoven, The Netherlands
Email: {n.batalas, h.s.d.bruikman, d.turzynska, v.vakili, n.voynarovskaya, p.markopoulos}@tue.nl

***Abstract*—The process of education and innovation often involves individuals, whose expertise lies in diverse fields. Informal communication amongst them can prove to be invaluable towards their collaboration, but rarely does it extend beyond one's already established social circle. This paper proposes SIREN, a system, which makes use of sensor nodes to detect encounters between colleagues in their workplace, as they undertake their daily tasks and spreads information virally from one to another. The aim is to introduce a channel of informal communication that is not disruptive to their routine. We developed a system prototype and conducted a field test to detemine whether the premise of encounter-based information sharing offers any added value over information sharing without discrimenation, amongst users in the same work setting. The results indicate that SIREN can help break down barriers and promote subsequent direct communication between users.**

***Keywords*-human-centered computing; wireless sensor networks; serendipity**

I. INTRODUCTION

Large organizations operating in campus settings, such as universities, tend to bring together people from a variety of backgrounds, to contribute their expertise to the challenges that education and innovation present. Characteristic of the affordances allowed by this environment is Humboldt's vision of teachers and students, not merely engaging in the tutoring and learning processes, but also taking initiatives in the investigation of their own research interests. In his 'Theory of Human Education', Humboldt stresses the importance of the links established between the individual and his/her surroundings for the fulfillment of such a purpose [20].

There is strong support for the importance of informal communication in the establishment of such links. Informality here, refers to the attributes of spontaneity and richness, beyond the impositions of rules and hierarchies and the lack of pre-specification [27]. It is also strongly associated with face-to-face communication. This kind of communication has been found to often constitute the beginnings of scientific collaborations, as reviewed in [24][28]. Additionally, the feeling of community and familiarity that is often implicit in informal communication plays a decisive role in how gratifying the working environment can be, and is considered a necessity for innovation [14]. Kraut, Egido, and Galegher [26] found that through informal communication, collaborative relationships between scientific workers were established.

Corridors and collection points such as elevator lobbies and support services (e.g., printers, water coolers, coffee stations) have been found to be instrumental in promoting informal communication. However, factors such as the dispersion of workers behind physical barriers (doors, stairs, corridors), the lack of acquaintance with others, unaligned schedules and the need to observe hierarchy, tend to discourage informal encounters, or take away from their potential, when they take place. As pointed out by Serrato and Wineman [33], Allen [1] advocates placing support facilities so that they are shared by workers, whose physical separation might otherwise inhibit communication.

Given the importance of informal communication between colleagues, a vast array of technologies are available, which can help distributed coworkers overcome physical barriers and can also facilitate contact and foster collaboration. These range from landlines and mobile phones to video and audio teleconferencing, from voicemail to email and mailing lists, and from instant messaging to wikis and personal blogs. A recent survey [12] indicates a massive dominance of tools that actively transmit data to recipients found in a list of contacts (email, instant messaging) over media that passively allow for potential public access, such as wikis or blogs. Eventually, dissemination of information through a network of contacts, is bounded by the extent of one's contact list. Moreover, these tools operate under the assumption that people work solely at their desks, completely disregarding a user's physical environment.

To address these shortcomings, we examine in this work the numerous chance encounters that take place in the physical setting of an organization, in this case a University. We consider chance encounters to be the unintended meetings of people, who are either familiar or unfamiliar with each other. They take place not only between established collaborators or socially involved colleagues, but also between plain acquaintances, non-strangers or even total strangers. Chance encounters are valuable in themselves when they cause people to strike conversations, but this happens only rarely. Still, due to their episodic nature and the fact that they are

 ISBN: 978-1-61208-100-7

situated within the physical space, they carry high potential for allowing an individual to attain both an implicit awareness and an explicit knowledge of what activities, questions and findings their co-workers are concerned with. Furthermore, they can become the basis for the establishment of channels of informal communication.

To this end, we introduce a platform that attempts to embed the individual into an informal, ad hoc social network where information is shared virally amongst peers, who physically encounter each other in the same work setting, specifically that of a university, with the purpose of inducing serendipity in information exchange for its users. Viral sharing, in the context of physical encounters, refers to the ability of a user to transmit, upon an encounter with another user, information, which has previously been picked up from a different person not involved in the current encounter. Serendipity refers to the happy accident. or otherwise, the beneficial outcome that is potentially latent in these accidental encounters.

In the text that follows we first look at related work that has been conducted in the past. We then lay out our concept and present the platform, an application for it and a field evaluation of the system, juxtaposed with a similar application lacking the premise of physical encounters and the viral dimension.

II. Related Work

A multitude of systems for information sharing or creating awareness have been proposed, with the intention to support informal communication amongst physically dispersed co-workers. Here we choose to discuss those, which we consider to be representative of the tendencies in development. They fall into two main categories :

a media spaces that make use of video and audio to compensate for distance, by creating a virtual shared space, either through videoconferencing or virtual environments. Such applications encourage serendipity by attempting to facilitate informal, accidental encounters between people who would not have that chance in the physical space (tele-proximity).

b applications that have actual physical proximity at the basis of their operation. Given that individuals find themselves within close range of each other, they try to detect these events and provide awareness.

The systems in the first category, when compared against actual physical proximity, allow for some of the latter's most inherent properties to surface. They help illustrate, both by their shortcomings in relation to physical proximity, and their successful substitutions of it, the desirable properties that are inherent in physical interaction and indicate how some could be substituted. In a similar manner, systems in the second category also carry lessons to learn.

Cruiser [30], in 1988 and VideoWindow [16], in 1990, attempted to create artificial proximity in order to support informal communication. Citing co-presence, low personal cost and the concentration of a population of suitable partners as some of the characteristics of physical proximity, the systems tried to leverage audio and video in a unique system. The resulting virtual workplace was supposed to recreate those characteristics and increase the number of potential interactions between coworkers at different locations. It was found that it could not fully account for the merits of actual physical co-presence.

EuroPARC's RAVE system [17] also aimed to support both synchronous collaboration and semi-synchronous awareness amongst physically separated colleagues, by means of audio and video. RAVE considers the general awareness it facilitates to be the underlying foundation that can lead to serendipitous communication and even to focused collaboration.

The Forum Contact Space [23] was another such system, which intended to provide colleagues with a Collaborative Virtual Environment, where chance encounters could take place. To produce these, the concept of Symbolic Acting is employed, where real activity on a desktop computer is mapped onto a state in a virtual world. The authors indicate that online unintended interactions helped lead to real world interactions and develop a feeling of community.

On creating awareness when physical proximity occurs, one of the first studies, in 1992, involved the ActiveBadge [36], a technology first developed for the purpose of providing a central service with information about the location of individuals within a building. The badges could be detected by special sensors placed in areas of interest and inferences could be made not only about a person's whereabouts and how to accurately reach them at that point in time, but also about the people they were with at that given moment.

As early as 1996, Bly and Belloti describe a change of focus in research and design of Computer Supported Collaborative Work (CSCW), from substituting for mobility in informal communication, to supporting its role [2]. However, their proposals were limited to allowing users to be away from their desks without experiencing the negative aspects of wandering around the corridor (e.g., missing important messages and phonecalls). Although their proposals acknowledge the importance of informal communication in corridors and common areas, their systems only at a rudimentary level allowed this and did not facilitate these activities. In a later paper the authors also provide an extensive overview of difficulties in media space designs [11].

MemeTags [5], in 1998, based on ThinkingTags [6] and GroupWear [4], looked into offering conference attendees a tool to promote interaction between them through the exchange of preselected short messages (memes). The exchange would happen by means of electronic badges with an LCD display and infrared communications. Upon facing each other, they would exchange memes and prompt the user to indicate agreement or disagreement. Thus, conference participants could build a shared understanding and lay the ground for future collaboration. Large displays (community mirrors) dispersed through the conference would also let people have an overview of the most agreeable memes and the general activities within the system.

HummingBird [21] introduced the term 'Interpersonal Awareness Devices (IPADS)'. It describes those devices,

 ISBN: 978-1-61208-100-7

which aim to help people initiate contact, rather than sustain the actual communication, and do not rely on any additional infrastructure besides their own kind. HummingBird alerts people by aural and visual means when they are in the vicinity of each other. HummingBird showed more potential when used in unfamiliar situations, such as trips and conferences, and was mostly ignored in the office setting.

ProxyLady [9] uses PDAs equipped with radio transceivers, with the objective of fostering face-to-face communication. The user associates information items with people (candidates for interaction). When a candidate is near, the PDA notifies the user and brings up the associated information item. The implications of this association is that exchanges only happen between users who have already been in some form of contact with each other. Contrary to Hummingbird, which seeks to maintain general awareness in (mobile) groups, ProxyLady aims to increase the frequency and quality of opportunistic, informal communication.

Hocman [15], in 2004, focused on the particular case of motorcyclists. It offered users own control over what they shared in the system, a feature that GroupWear and Hummingbird were lacking. It provided motorcyclists equipped with PDAs while driving, the capability of sharing html pages, audio files and images with other bikers who were also in possession of a PDA and had been in their vicinity. A sound clip was played when another rider was close. A field test found that the bikers appreciated hearing the sound clip, inspecting logs, and browsing contact information. However, they did not believe that Hocman would rationalize biking. The prototype was also found to support the possibilities for further contact.

Social Serendipity[13], in 2005, used profile matching to alert users that possibly interesting people have been found in their proximity. It relied on bluetooth devices to detect encounters and in later iterations also allows users to share their profiles.

All of the above systems have had positive results to report in certain respects, which justifies the notion that there are desirable properties to the sharing of information on the basis of physical proximity. They are quite different from systems aiming to support co-workers in remote locations. The concept we present, which we call SIREN, combines aspects found in the former systems with unique characteristics, for a particular cast of users. In particular :

- SIREN, much like some of the systems discussed, uses the basic concept of information exchange based on proximity.
- It is different from similar systems in that it makes no effort to alert or interrupt users in any way, at the time when the encounters are sensed.
- It also, for the first time, introduces the concept of transmitting received information to others according to a viral model, without the user's initiative.
- SIREN targets members of a university community, where sharing and combining knowledge from different fields is always desirable.

The following paragraphs outline SIREN and detail the design, implementation and evaluation of a prototype that puts the concept into effect.

III. The SIREN concept

SIREN stands for Serendipitous Information-Relaying ENcounters. Encounters between users are the key premise of the system's operation, and serendipity best illustrates the desired effect for the system.

Our concept builds upon three central notions :

A. Physical encounters

Information is exchanged between two users when they find themselves in each other's close vicinity. Users become the routers of the information they carry around. The network paths, which information is routed along, are woven into the physical space users move around in. These encounters happen naturally and accidentally as people move about in their workplace. They can be however unsuspected or brief and the exchange does not require that they become actual interactions or conversations.

From a Human-Computer Interaction perspective, most of these encounters are not meant to interact with the system. Nonetheless, they are sensed and understood by it as input and make up the fragments of situational context for implicit human-computer interaction, as defined in [32]. The implications of looking at physical encounters in terms of implicit input become apparent later, as we separate the concept into two distinct levels, the platform that acquires this information and the application that makes use of it.

Additionally, a significant number of these encounters are part of an individual's episodic personal history [35]. The pieces of information received through them, compared with the features that define the sharing incident (spatial setting, temporal data or data about the origin of the information piece) can lead to semantic encoding of those encounters [35](p. 398).

B. Viral transmission

A single piece of information can be transmitted from carrier to host, rendering the latter into an infectious agent as well, mimicking the way a virus spreads over a population.

Traditionally, information flows selectively, passed on by its host only to contacts that the host thinks it would be of interest to. In addition to this, the principle of homophily tells us that people with similar characteristics tend to network with each other [37]. McPherson, Smith-Lovin and Cook state that 'Homophily limits people's social worlds in a way that has powerful implications for the information they receive, the attitudes they form and the interactions they experience'. [29].

In contrast to this, the viral transmission model promotes the far-reaching capabilities of sharing information with SIREN. It engages unrelated users, who do not encounter each other, into transitive relations amongst them and their common intermediate contacts(Fig. 1). It also helps provide a by-product awareness of the state of interactions within the system and the

 ISBN: 978-1-61208-100-7

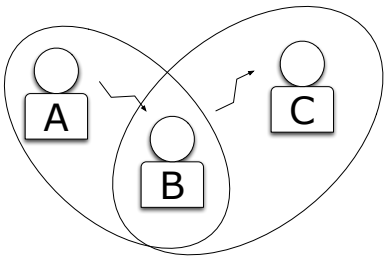


Fig. 1. With viral sharing, A and C enter into a relationship of exchange because of their individual connections to B.

workplace, one which, as described in [34] is not purposefully declared, but unintentionally arises from the plain fact that the history and current state of the system are such.

C. Seeding and viewing of information

User interaction with the system and the information that is transmitted through it need not take place at the level of encounters. The use of an application to load information onto the system and to view information that has been received is still necessary, but only at the user's choosing. The technology itself is not meant to be disruptive to the user's routine and need not demand their attention.

D. Mode of operation and the intended effect

To further elaborate, in our proposed system, users are cast in the role of information carriers and the system attempts to distribute this information from one to another, when they encounter each other as they move around. Once two users are found to be within close range of one another, an encounter is registered and pieces of information they carry are exchanged between them. The exchange takes place effortlessly, not requiring the user's intent or attention. The role of carrier need not be limited to human agents, but can also be assigned to specific objects or locations, such as coffee machines and common rooms, thereby rendering them collectors and hotspots of information. A single piece of information, once picked up by a carrier, joins the pool of pieces of information, which that carrier can transmit, and can be then retransmitted to another one, following a model of viral transmission.

Consequently, a particular piece of information can reach a far greater number of users than traditional personal exchange or contact-list based communication tools (email, instant messaging) could allow for [37]. Moreover, a very diverse network of information flow can be formed, taking advantage not only of the ties between users (strong and weak [19]), but also of elements in their mutual physical surroundings. It is out of the concept's scope to dictate the nature of that information.

The effect that we desire the system to have, in its setting of use, is best illustrated by the concept of serendipity. Serendipity already holds a prominent place in the discussion for creativity and innovation and numerous findings have been attributed to it in science and technology [31]. Nowadays [18] it describes the unsought-for finding that has value. It is also associated with the human faculty of making the finding, or seeing the value or benefit in the fortuitous event. Serendipity is thus the beneficial or happy accident and while we cannot account for the human ability to draw insight from incongruousness, we aspire in providing our users with the accidental findings that they can potentially benefit from. Those findings can either be the information that is shared, or come from the knowledge of the (potentially viral) path that it took to reach the user and the assumptions that might be made about its carriers.

IV. Prototyping SIREN

In designing a prototype of the SIREN concept, we set out to target the workers of the Main Building of our University . We define the goal of the prototype to be to 'Support education and innovation using serendipity, for users in the Main Building'.

The Main Building is one of the largest buildings on campus and home to two different departments. Administrative staff, students (undergraduates and graduates) and professors share the building and its neighbouring cafeteria.

A. Problem domain analysis

A gathering of requirements took place at the beginning, which consisted of 3 steps: a focus group, interviews, and an online questionnaire.

1) Focus Group: The focus group was put together for the purpose of testing the concept's validity with the intended users and also to identify concerns they might face when using the system. It consisted of 7 students from mixed educational backgrounds, who were shown a video prototype outlining the core concepts of information sharing based on physical proximity and viral spreading of that information. They were then asked to individually consider a list of aspects that had not been addressed by the video prototype and possible implementation opportunities.

Results of their individual reflections were addressed in a consequent group discussion. Afterwards, they were transcribed to a hierarchy, by concept and by aspects that we asked them to think about. This gave us an overview of current issues and concerns, detailed in Table 1.

Notably, the focus group identified the social potential of informal knowledge sharing. More particularly, it was reported that hinting at information can be more powerful than its exhaustive provision, in the way that it stimulates curiosity and discussion.

2) Interviews: A set of interviews were conducted with students, postgraduates and professors from two different faculties, both placed in the Main Building. The purpose was to find out how informal communication takes place amongst our potential users and what their attitudes towards information and knowledge sharing are.

The total interviewees were 15. They were all asked 6 questions in a semi-structured way and the answers were analyzed using open coding. Besides gathering qualitative data,

 ISBN: 978-1-61208-100-7

TABLE I
FOCUS GROUP RESULTS

Current Issues
Desk proximity determines degree of communication. Not enough information sharing in current practice. People are protective of information regarding their activities or concerns.
Application concerns
Too much unfiltered information may lead to ignoring everything. Reyling on a device in order to share information could make you less social.
Other Comments
Different sharing intent for different colleagues. Incompleteness of information might make people curious and lead to further communication.

the interviews were aimed at identifying common answers that could be used in the larger-scale survey that was to follow.

The interviews indicated that informal communication is both desirable and sought after, through a variety of means. Besides face-to-face communication, posters placed in corridors are used to communicate what people are working on. Additionally, a person's social network also plays a role in conveying information about people that do not necessarily belong to that network.

Another noteworthy finding was the interchangeable role of knowledge sharing as both an objective in itself and a pretext for communication. People do communicate with the purpose of sharing knowledge with each other on a particular topic. However, they also structure wider aspects of socializing around the sharing of knowledge and sometimes this exchange is the guise under which broader social communication takes place.

3) Online Survey: An online questionnaire was filled out by 31 respondents, all academic personnel with the exception of one. While they belonged to the same faculty, they were affiliated with different sub-departments. The survey focused on issues of information sharing, namely what the need and motivation for it could be, what kind of information is shared, in which ways, and how physical barriers affect the process.

The information that respondents mostly seek out about others is what they are working on, especially if it might be relative to their own field of research. Their methods of gathering information run from informal to formal, becoming more formal as physical distance increases between them. They also reported lower awareness about what others are working on, as physical distance of their respective workplaces increases. Although their desire to know decreases with distance, half of the respondents reported that they would still like to know what colleagues who are situated at a distance are working on.

The resulting prototype was implemented in two separate components :

1) A platform that allows the encounters to be detected and recorded
2) An application that puts the information about the encounters to use, in order to make inferences about the viral sharing of information. It also enables the users to author information, view information that they receive and further the exchange with contacts that have already been sensed.

The prototype was evaluated in a field study and was compared to an automatic way of sharing using only the web application. Both quantitative (log-files of the systems, survey) and qualitative (open-ended items from the survey and a brief interview) data was collected and analyzed.

B. Encounter sensing platform

In the current implementation of our concept, we used sensor node devices, also known as motes, and put them into service as smart badges. We used the Crossbow Mica2dot model [8], which employs an Atmel ATmega128L microcontroller and the CC1000 RF Transceiver [7], packed in the form factor of a 25mm in diameter disk. The device also features an antenna of 8.12cm and a holster for its 3V button cell battery. It is small enough for the user to carry around on their person, and meant to be used in this way. It should be noted that wireless connections implemented with these devices are very unstable. Jea and Srivastava [22] investigate some of the characterics of packet transmission with the Mica2Dot motes.

Due to device limitations no actual information, which users might choose to share, is stored on the mote itself. Instead, each mote acts as a beacon, regularly transmitting an id number (every second). Also, each mote samples an internal stack of received ids at a constant rate, checking if reception of beacon messages from other motes has taken place within a predefined time period (every 5 seconds). If that is the case, it registers an encounter along with a timestamp from its own clock. The motes are set to broadcast in low power, thus having a limited range of 3-7 meters. Consequently, the reception of a mote's id by another is considered to be indicative of the fact that their respective carriers have found themselves in the close vicinity of each other and have had an encounter.

Since these beacons cannot communicate with any other networking infrastructure (e.g., wifi nodes), a second class of motes act as intermediates and along with PCs they are connected to, form gateways to a database server. Upon encounter with the user-motes, these gateways retrieve the list of encounters from the users device and upload it to the server. The server stores each encounter, converts the timestamps that where local to each mote into the actual time and makes these data available for use by any application. Applications should then be able to deduce which information shared by one person should be made available to another.

C. Messaging application

The application implemented on top of this platform, was a messaging application. In its simplest form, it is straightforward to implement and users can readily understand how to use it. We chose the application to be web-based rather than run on each desktop. This allowed it to be multi-platform, lessened the weight of technical support requests from the users and allowed easy and central monitoring of each user's

 ISBN: 978-1-61208-100-7

status. The application design also takes Kortuem & Segall's design principles for wearable communities [25] into account.

Users post textual messages, which are kept in a database. Once an encounter has been reported, the messages posted by one user become available for viewing to the user encountered. Additionally, the messages that each of these users have previously 'picked up' in the same way from other users, become available to their current encounter. In this way the viral transmission takes place.

The overview of what exchanges have taken place is available to the users from the web application, with the emphasis on the display of the text messages themselves, rather than visualization of data. Users are able to:

- Set up a profile and upload a picture.
- Write an initial message, that is a message that is being distributed by the device. The messages can be replied to and constitute starting points for discussion threads.
- Reply to a message, which no longer needs the device in order to be spread. This creates a thread based on the initial message. Replies are visible both to the author of the initial message, and to persons that have also reacted to that message.
- Follow a message, which makes the thread initiated by that message available to the users, without requiring their participation.
- Post invitations to other users to meet for a coffee.

Along with each message, information is also provided to the reader about the author of the message, the time of its posting, and in the case of initial messages, the time of reception through the wireless device.

V. Field Test

The purpose of the study was to perform an exploratory, formative evaluation and to collect ecologically valid data. Specifically, we aimed to:

- Evaluate the experience of using the system.
- Investigate whether physical encounters provide value as the premise for serendipity, over the exchange of messages between randomly selected individuals.
- Test the assumption that the viral spreading of messages effectively increases exposure to unexpected information.

A between-subjects design was used. Users situated in the Main Building where invited to participate both through a recruitment campaign, involving posters and mass-mail invitations, and through our social network.

The final sample of subjects was comprised of bachelor's, master's, and PhD students, as well as academic and managerial staff. The fact that individuals from 2 different departments were chosen, their workplaces positioned in different rooms and floors, further ensured the diversity and dispersion of the participant population. Finally, the participants were assigned to one of two conditions.

1) Using a system where sharing was based on physical co-location (encounter based sharing) with the participation of 15 users.
2) Using a very similar system that lacked any premise for information exchange other than users posting that information. Every exchange would take place between individuals selected randomly from the user population (random sharing). This group consisted of 11 users.

To isolate the effect of encounter-based sharing on the experience of the system's usage, the number of messages that would be received by users in the random sharing group was regulated to match the number of messages that were actually being received by users in the encounter-based sharing group. It could therefore be ensured that the level of usage, a well-known factor in the success of any type of social media, would not be a confounding element in the comparisons made.

After a short briefing and consent process all participants received training, where all the functions of the web application were explained. The encounter-based group received additional instructions on how to handle the motes. Participants in the random-sharing were asked to use the web application, and those in the encounter-based group to use the web application in combination with a mote. The field test lasted 5 days. At the beginning of each day participants in the encounter-based group were expected to use a fresh battery for their wireless sensor. They were also expected to carry it with them if they left their office, and make use of the web application at least once a day. Participants in the random sharing group were also asked to use the web application at least once a day.

Because the number of gateways installed was not yet sufficient for the regular retrieval of content from the motes, we frequently toured the corridors with a mote gateway connected to a laptop and sent commands to motes found within range, to upload their data to the laptop. The laptop would then upload the information to the server. This took away from the realtime aspect of interaction between motes, as an encounter that had already happened when a subject would visit the messaging application, would not appear to have been registered, until a collection round had taken place.

At the end of the study, the participants were handed a questionnaire to measure their experience. They were also asked to provide comments in an open interview. Finally, a debriefing followed and participants were given a reward for their assistance.

VI. Results

The prototype was evaluated following the principle of triangulation on the level of multi-measures. Measures that were used are:

- Content of posts and profiles.
- Log-files of interaction with the web application.
- Scores on a post-usage survey and a brief open interview following this.

The evaluation sought to:

- See if the design objectives and purpose were fulfilled.
- Justify the approach of serendipity and encounter-based sharing by showing that the latter related to positive results.
- Seek reasons for why the approach was effective or not.

 ISBN: 978-1-61208-100-7

TABLE II
MEAN SCORES FOR ENCOUNTER-BASED AND RANDOM SHARING GROUPS. BOLD INDICATES SIGNIFICANCE IN DIFFERENCE.

scale	$\bar{E}$	$\bar{R}$	$p - value$
Perceived Usefulness [10]	4.38	3.33	**0.031**
Perceived Ease of use [10]	2.8	5.27	0.228
Perceived Innovation	4.71	3.55	**0.032**
Perceived Education	4.37	3.45	**0.036**
Professional Communication	5.47	5.27	0.330
Personal Communication	5.8	4.36	**0.004**
Perceived Self Worth [3]	5	4.3	**0.036**
Potential for Connecting to others	4.19	3.15	**0.008**

A. Scales Measured

Table II lists the scales measured and the scores obtained from the post-test questionnaire. In addition to the well known scales referenced, the following scales were measured:

- Perceived Innovation
  - *I believe that possible meetings following an invite facilitate innovation.*
  - *I feel the posts exchanged facilitate innovation and idea generation.*
  - *Using the system helps me innovate.*
- Perceived Education
  - *I believe that possible meetings following an invite facilitate education.*
  - *I feel the posts exchanged enhance education.*
  - *Using the system enhances my education.*
- Personal Communication
  - *I feel it is useful to communicate on a personal level with people I normally wouldn't communicate with.*
- Professional Communication
  - *I feel it is useful to communicate on a professional level with people I normally wouldn't communicate with.*

Additionally, we measured the following on a dichotomous scale :

- Attitude toward adoption of system :
  - *Consider that you have been using a prototype. Would you like this concept to be implemented in the Main Building, assuming that it would be used by most people?*
- Experienced Serendipity
  - *I experienced something random or accidental that I am happy about or benefited from because of my participation in this study.*

Scores from the two groups, in Table II, were compared using a one-tail Mann-Whitney U test, to test the hypothesis that the encounter-based system would perform better than the random-sharing one. The encounter-based sharing group reported significantly higher perceptions on almost all counts. The only exception was on Perceived Ease of Use. Additionally, Professional vs Personal Communication indicates that encounter-based sharing provides added value for informal communication over random sharing.

On whether they had experienced serendipity, 6/15 (40%) reported yes in the encounter-based group, and only 1/11 (9%) had a positive response in the random sharing group. However, a Fischer-exact test showed this difference to not be significant, with $p = 0.093$.

On their attitude toward adoption of the system, in both groups, most subjects said that they would use the system, with 60% of subjects in the random sharing group and 92% in the encounter-based group.

B. Qualitative assessments

The open-ended questions in the questionnaire and the interview that followed, as well as the content of the messages that had been exchanged, helped provide a set of observations about the use of encounter-based sharing as opposed to random sharing. Answers to the questions were qualitatively analyzed by means of clustering key descriptions of individual responses. A similar approach was used for the qualitative analysis of the messages that had been exchanged. The following observations were made :

- Messages that had been shared in both conditions were of an informal nature.
- The messages that participants in the encounter-based group wrote seemed to be less elaborate. Users mostly posted short sentences of greeting or small information about current tasks. In this respect, the system was used more like a micro-blogging tool. Replies to these messages were also short and not promoting the perpetuation of the conversation.
- People in the random sharing group appeared to be more social, posting more meaningful messages. However, discussions in the random-sharing group had been sustained between people who already knew each other, and there were points were the discourse did resemble a forum.
- Users in the encounter-based sharing group sent out invitations to meet with others, some for the purpose of testing the feature. However, most of those invitations did lead to some form of direct communication. Users in the random sharing group barely did so.
- Users that had received information because of an encounter reported that they were more eager to view it. They also trusted the information to be relevant to them.
- Viral transmission allowed users in the encounter-based sharing group to view information from people they did not cross paths with.

VII. DISCUSSION

SIREN performed expectedly better than the random sharing condition: Towards the goal of supporting innovation and education in its deployment environment, participants in the field test reported significantly better potential in SIREN than in random sharing (as seen on table II). They also perceived the system as positively useful, in contrast to the random sharing option: Their perception of the latter's usefulness was rated at 3.3, just below neutral on the 7-point Likert scale.

 ISBN: 978-1-61208-100-7

However, users of both systems were favourably inclined towards their adoption, at 60% for random sharing and at 90% for SIREN. It should be noted here that the specific workplace does not offer a forum or other communication media that might allow the personnel to informally exchange messages. There is a possibility that these results reflect the lack of such a system, especially since the random sharing system seemed to be appropriated more like a forum/BBS than the encounter-based system. A reason for this could be that recipients in the random sharing condition, as is the case in a forum, could not attribute the fact that they had received a message to any other meaningful event, other than their participation in the system.

On the contrary, in the case of SIREN, it could be argued that users were much more conscious of their role in the process of the delivery of a message to someone. Also, communication did take place between people who weren't acquainted with each other. This could explain the caution they applied in composing the messages they wrote. However, the fact that the initial information exchanged follows an encounter that did occur, can help people feel more connected and experience fewer barriers towards direct communication. Hence, subsequent physical proximity and face-to-face communication can be promoted. As we have discussed before, physical proximity builds a foundation for informal communication, a positive perception amongst co-workers and collaboration.

Despite the shorter conversations, people in the encounter-based sharing group felt eventually more open to personal communication with people they would not normally communicate with. In addition to this, they reported significantly better potential for connecting to others. Also, self worth, a moderator of intention to share knowledge [3], was also found to be significantly higher with SIREN.

Another thing to note was that users in the random sharing group perceived their system as easier to use. This can be attributed to the fact that SIREN users were given the onus of carrying around their wireless device and making sure it was operational.

Contrary to our expectations, serendipity was not significantly different between the two groups. Our explanation for this is the short timeframe that the field test took place in, so that not enough serendipitous event did take place for the SIREN group. On the other hand, given enough time, every member of both test groups will report serendipity. We expect a follow up study that will last longer and investigate the rate, rather than the simple occurence of serendipitous events, to show better results in this area.

VIII. Conclusion and Future Work

We presented a system called SIREN, which uses the premise of encounters between colleagues in the workplace to facilitate the non-disruptive exchange of information. The non-disruptive exchange is one respect in which it is different from many of similar systems already proposed. The other new element that it introduces, is that it allows for information to virally spread over the population of coworkers as they encounter each other, for the purpose of promoting its distribution. The underlying assumption is that the inherent properties of routine movement in the workplace can be taken advantage of, in order to achieve both a form of implicit input, and to also provide meaningful context to the recipient of a piece of information for the reason why it reached them. Eventually it could promote the formulation of a mental model of the fields of interest of remotely distributed colleagues and the network connections amongst them and prove supportive of informal communication.

To test the assumption that physical encounters provided added value to the exchange of information, a prototype was developed and a field test was performed to explore how SIREN might be received by its potential user population. The field test tried to isolate the effect of information exchange based on proximity, by conducting a comparison to the condition of exchange in a random fashion, without the requirement that encounters take place. The comparison also focused on aspects of usage that are relevant to the setting of a university as a place of innovation and education. Additionally, it investigated serendipity as the effect of the unsought-for discovery that could be brought about, given that the reception of information because of use of the system, could be the unsought-for event.

Despite limitations in the prototype and the short duration of the field test, overall results were positive. Encounter-based information exchange was deemed to be more supportive of informal communication and was also perceived as better for education and innovation. It should be taken into account that the field test proved not sensitive enough to evaluate the difference between the two systems in their facilitation of serendipity. Participating in the study could have resulted in serendipity by itself.

However, the study reaffirms the consensus that informal communication is desired and perceived as useful. This appears to be particularly true in the setting of an educational institution. The evaluation performed indicates that there is unharnessed potential, to explore in future work, in the way people make use of their physical settings and the chance encounters that happen within the workplace, as a means for the transitive mediation of informal communication. Future work in this area, with the use of more elaborate and reliable prototypes, could include investigations of how such a system could be integrated with current social networking applications, as well as the development of new applications that make use of the encounter-sensing platform.

References

[1] T. Allen. *Managing the flow of technology*. MIT press Cambridge, MA, 1977.

[2] S. Bly, S. B. Consulting, and V. Bellotti. Walking Away from the Desktop Computer: Distributed Collaboration and Mobility in a Product Design Team. *Computer*, pages 209–218, 1996.

[3] J. Bock, G.W. And Zmud, R.W. And Kim, Y.G. And Lee. Behavioral intention formation in knowledge sharing: Examining the roles of extrinsic motivators, social-psychological forces, and organizational climate. *Mis Quarterly*, 29(1):87–111, 2005.

 ISBN: 978-1-61208-100-7

[4] R. Borovoy, F. Martin, M. Resnick, and B. Silverman. GroupWear : Nametags that Tell about Relationships. In *Conference on Human Factors in Computing Systems*, number April, pages 329–330. ACM New York, NY, USA, 1998.

[5] R. Borovoy, F. Martin, S. Vemuri, M. Resnick, B. Silverman, and C. Hancock. Meme Tags and Community Mirrors: Moving from Conferences to Collaboration. In *Proceedings of the 1998 ACM conference on Computer supported cooperative work*, pages 159–168. ACM, 1998.

[6] R. Borovoy, M. McDonald, F. Martin, and M. Resnick. Things that blink: Computationally augmented name tags. *IBM Systems Journal*, 35(3):488–495, 1996.

[7] CHIPCON. SmartRF CC1000.

[8] Crossbow Technologies. Mica2dot Wireless Sensor Mote.

[9] P. Dahlberg, F. Ljungberg, and J. Sanneblad. Proxy Lady-Mobile Support for Opportunistic Communication. *Scandinavian Journal of Information Systems*, 14:3–18, 2002.

[10] F. Davis. Perceived Usefulness, Perceived Ease of Use and User Acceptance of Information Technology. *MIS quarterly*, 13(3):319–340, 1989.

[11] P. Dourish, A. Adler, V. Bellotti, and A. Henderson. Your place or mine? Learning from long-term use of Audio-Video communication. *Computer Supported Cooperative Work (CSCW)*, 5(1):33–62, 1996.

[12] S. D'Urso and K. Pierce. Connected to the Organization: A Survey of Communication Technologies in the Modern Organizational Landscape. *Communication Research Reports*, 26(1):75–81, 2009.

[13] N. Eagle and a. Pentland. Social Serendipity: Mobilizing Social Software. *IEEE Pervasive Computing*, 4(2):28–34, Apr. 2005.

[14] H. a. Earle. Building a workplace of choice: Using the work environment to attract and retain top talent. *Journal of Facilities Management*, 2(3):244–257, 2003.

[15] M. Esbjrnsson, O. Juhlin, and M. Stergren. Traffic encounters and Hocman: Associating motorcycle ethnography with design. *Personal and Ubiquitous Computing*, 8(2):92–99, 2004.

[16] R. Fish, R. Kraut, and B. Chalfonte. The VideoWindow system in informal communication. In *Proceedings of the 1990 ACM conference on Computer-supported cooperative work*, number October, pages 1–11. ACM New York, NY, USA, 1990.

[17] W. Gaver, T. Moran, A. MacLean, L. Lövstrand, P. Dourish, K. Carter, and W. Buxton. Realizing a video environment: EuroPARC's RAVE system. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 27–35. ACM New York, NY, USA, 1992.

[18] L. Goodman. Notes on the Etymology of Serendipity and Some Related Philogical Observations. *Modern Language Notes*, 76(5):454–457, 1961.

[19] M. Granovetter. The strength of weak ties. *American journal of sociology*, 78(6):1360, 1973.

[20] G. Hohendorf. Wilhelm von humboldt (1767-1835). *Prospects: the quarterly review of comparative education*, XXIII(3/4):613–23, 1993.

[21] L. E. Holmquist, J. Falk, and J. Wigstrm. Supporting group collaboration with interpersonal awareness devices. *Personal Technologies*, 3(1-2):13–21, Mar. 1999.

[22] D. Jea and M. Srivastava. Channels Characteristics for On-Body Mica2Dot Wireless Sensor Networks. In *IEEE International Conference on Mobile and Ubiquitous Systems (MobiQuitous)*, 2005.

[23] P. Jeffrey and A. McGrath. Sharing serendipity in the workplace. *Proceedings of the third international conference on Collaborative virtual environments - CVE '00*, pages 173–179, 2000.

[24] J. S. Katz. Geographical proximity and scientific collaboration. *Scientometrics*, 31(1):31–43, Sept. 1994.

[25] G. Kortuem and Z. Segall. Wearable communities: augmenting social networks with wearable computers. *IEEE Pervasive Computing*, 2(1):71–78, Jan. 2003.

[26] R. Kraut, C. Egido, and J. Galegher. Patterns of Contact and Communication in Scientific Research Collaboration. In *Proceedings of the 1988 ACM conference on Computer-supported cooperative work*, page 12. ACM, 1988.

[27] R. Kraut, R. Fish, R. Root, and B. Chalfonte. Informal Communication in Organizations: Form, Function and Technology. *Baecker (1993): Readings in Groupware and computer-supported Cooperative Work. Morgan Kaufman*, pages 145–199, 1990.

[28] R. Kraut, S. Fussell, S. Brennan, and J. Siegel. Understanding Effects of Proximity on Collaboration : Implications for Technologies to Support Remote Collaborative Work. *Distributed Work*, pages 137–162, 2002.

[29] M. McPherson, L. Smith-Lovin, and J. Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, 27(1):415–444, 2001.

[30] R. W. Root. Design of a multi-media vehicle for social browsing. *Proceedings of the 1988 ACM conference on Computer-supported cooperative work - CSCW '88*, pages 25–38, 1988.

[31] M. Rosenman. Serendipity and Scientific Discovery. *Journal of Creative Behaviour*, 22:132–138, 1988.

[32] A. Schmidt. Implicit human computer interaction through context. *Personal Technologies*, 4(2-3):191–199, June 2000.

[33] M. Serrato and J. Wineman. Enhancing communication in lab based organizations. In *Space Syntax Symposium*, 1997.

[34] C. Simone and S. Bandini. Integrating awareness in cooperative applications through the reaction-diffusion. *Computer Supported Cooperative Work (CSCW)*, pages 495–530, 2002.

[35] E. Tulving. *Episodic and Semantic Memory*, chapter 10, pages 382–402. Academic Press, Inc, New York, 1972.

[36] R. Want, A. Hopper, V. Falcão, and J. Gibbons. The active badge location system. *ACM Transactions on Information Systems*, 10(1):91–102, Jan. 1992.

[37] F. Wu, B. Huberman, L. Adamic, and J. Tyler. Information flow in social groups. *Physica A: Statistical and Theoretical Physics*, 337(1-2):327–335, 2004.

 ISBN: 978-1-61208-100-7

Dynamic Object Binding for Opportunistic Localisation

Isabelle De Cock
Willy Loockx
Dept. of Applied Engineering
Artesis University College of Antwerp
Antwerp, Belgium
isabelle.decock@student.artesis.be
willy.loockx@artesis.be

Martin Klepal
Centre for Adaptive Wireless Systems
Cork Institute of Technology
Cork, Ireland
martin.klepal@cit.ie

Maarten Weyn
Dept. of Applied Engineering
Artesis University College of Antwerp
Antwerp, Belgium
maarten.weyn@artesis.be

***Abstract*—In this paper, dynamic object binding is proposed to improve the opportunistic localisation system. Object binding will be realised using Bluetooth. Many different techniques are already fused in the opportunistic localisation system, since Bluetooth is integrated in almost all mobile devices, this sensor will be incorporated in the opportunistic localisation fusion algorithm. In order to correctly use Bluetooth for object binding, some important features like operating range, influence of obstacles and scan time are analysed. A new measurement model is added to the particle filter engine to process incoming Bluetooth data. The client devices are continually scanning for other adjacent Bluetooth devices, this information is sent to the server where the client device position is estimated, based on the other adjacent Bluetooth devices, which are located through other means. This way object binding is realised.**

***Index Terms*—object binding; opportunistic localisation; Bluetooth**

I. Introduction

Today, localisation techniques are widely spread and already integrated in many applications such as GPS car systems, Google Earth, etc. Outdoor localisation is mostly accomplished by means of GPS, but usually GPS does not work indoor because there has to be a minimum of four satellites in line of sight, which is usually not the case indoor. There we can use WiFi [1] or GSM [2] or even other techniques such as Bluetooth [3], Zigbee [4], Ultra Wide Band (UWB) [5].

One big challenge is fusing these techniques into a single system. Acquire the sensor data of multiple sensors can be realised because most mobile devices such as Personal Digital Assistants (PDAs) and smart phones are very often equipped with GSM, GPS, WiFi or a combination of these. A system which combines this technologies is called Opportunistic Seamless Localisation System (OLS) [1].

This solution combines the above mentioned technologies together with the information of accelerometers, compass and camera . All these approaches are seamlessly fused by using an adaptive observation model for the particle filter, taking the availability of every technique into account. A particle filter [6] is a sequential Monte Carlo based technique used for position estimation. Since we are working with a real-time system, it is even harder to estimate the correct position therefore heavy and numerous calculations are not recommended. Limiting the number of particle filters is recommended in order to avoid numerous time-consuming calculations. For example, when this system is implemented at an airport where many devices are present, the system might be delayed due to these calculations for all those devices. Obviously, some devices will travel together such as people by bus, so that it is not necessary to calculate all their positions with different particle filters. Instead, we could combine all these objects and bind them in one group, in which case we only have to calculate one position for this group. This is one of the reasons why Bluetooth may be useful.

Bluetooth is a useful technique to detect other adjacent Bluetooth devices. Which would enable the possibility to detect whether people are moving together. Another interesting reason to use Bluetooth may be the possibility to locate unknown people. This could be useful to determine the amount of people in every area.

This paper is structured as follows: at first the scanning method is analysed followed by some real experiments to determine the operational range of Bluetooth devices. Thereafter, Bluetooth signal strength values are discussed. This is then followed by a short introduction about opportunistic seamless localisation and the explanation of the Bluetooth measurement model. Finally, the results are showed and the last section gives the conclusion of this paper.

II. Methods

In this section the use of Bluetooth and the localisation algorithm will be explained.

A. Bluetooth

Bluetooth [7] is a technique developed by Ericsson. This universal radio interface in the 2.45 GHz band makes it possible to connect portable wireless devices with each other. Bluetooth uses frequency hopping to avoid interference with other devices, which also use the license-free 2.45 GHz band.

 ISBN: 978-1-61208-100-7

1) Discovering: There are two ways of discovering [8] devices when using Bluetooth. The first and mostly used method is inquiry-based tracking. In case of inquiry-based tracking, the base station needs to scan for devices and to page all present devices in order to find them. All devices need to be detectable but they need not to be identified in advance. Scanning for devices absorbs a relatively large amount of time because primarily every base station sends a search-packet on all 32 radio channels. Every detectable device that receives this packet will answer. To avoid collision, every device will send his packet with a random delay. This is the reason why an inquiry has to run for at least 10.24 s to be reliable. Many devices are undiscoverable in order to increase the security and privacy of the owner. This is another technical problem that could occur and consequently it is not possible to find these devices by scanning the area.

A second method of tracking is the connection-based tracking. With connection-based tracking, devices are considered to be in a close range when one device has the possibility to connect with another device. All devices have to be paired with each other and this is a major problem when using the Radio Frequency Communications (RFCOMM) layer [9] connections with connection-based tracking. Practically, this requires human input which is time-consuming. Although, some communication services do not require this, it is still necessary that one of both devices knows the other one exists.

In practice, the creation of an Asynchronous Connectionless Link (ACL) [9] and a basic Logical Link Control and Adaptation Protocol (L2CAP) layer [9] connection is universal and authorisation-free. These connections are limited but they are in compliance with the requirements for tracking usage. It is only necessary to know whether a connection is possible and if this is the case, these 2 devices are in the same range. This connection also supports some low-level tasks such as RSSI measurements and L2CAP echo requests.

Both tracking techniques have their own advantages and disadvantages and they are both not ideal. Choosing the correct technique will depend on the situation. When using inquiry-based tracking, it is possible to find every detectable device without the need of knowing the devices in advance. The major disadvantage will be the relatively long scan time. When we choose the other option, connection-based tracking, the time to find the devices will be shorter and there is also the possibility to find undiscoverable devices. The major disadvantage here is the requirement that at least one party knows about the existence of the other one.

Another option could be a combination of both techniques. Combining these two techniques will not decrease the relatively long scan time because we always need to take the longest scan time in account. The advantage of combining both techniques is the possibility to find known 'undiscoverable' devices as well as unknown discoverable devices.

For this project, the first option is chosen because inquiry-based tracking has the possibility to track unknown devices, which will be useful for this project.

2) Range: Bluetooth devices can be divided in three different classes. Generally, class 1 and class 2 are used instead of class 3, which is due to the very short operating range of class 3.

Class	Maximum Power	Operating Range
1	100 mW (20 dBm)	Up to 100 m
2	2.5 mW (4 dBm)	Up to 10 m
3	1 mW (0 dBm)	Up to 1 m

These operating ranges are frequently used to estimate a position since signal strength is not always a good parameter due to effects like reflection, multi-path propagation, ... [10]

The operating range of a Bluetooth device can be defined by the maximum allowable path loss which can be calculated with Equation 1:

$$L_{total} = 20 * \log_{10}(f) + N * \log_{10}(d) + L_f(n) - 28 \quad (1)$$

$$L_{total} = 40 + 20 * \log_{10}(d) \quad (2)$$

where N is the *Distance Power Loss Coefficient*, f is the Frequency (Mhz), d is the distance (meters) between the nodes , L_f is the *Floor Penetration Loss Factor* (dB) and n is the number of floors penetrated.

When working in an open-air environment, Equation 2 which is the simplified version of Equation 1, can be used [11].

As operating ranges will be used to estimate a position, some tests were done in order to decide which maximum range will be utilized. A Dell XPS M1530 laptop has been set up as a base station. The two test devices were a Samsung E250 mobile phone (test device 1) and a Samsung F450 mobile phone (test device 2). All devices, including the base station are devices of class 2. Measurements were started at a distance of one meter away from the base station and afterwards extended by steps of one meter. Every measurement was repeated five times in order to have reliable results.

Fig. 1. Experiment 1

The first experiment, see Figure 1, was done in open space in which the two test devices are in line-of-sight of the base station.

Both test devices could easily bridge a distance of 9 m. Once the distance was increased, test device 1 was not longer detectable. Test device 2 was detectable until we reached a distance of 12 m.

In the next experiment, the influence of obstacles between the

 ISBN: 978-1-61208-100-7

Fig. 2. Experiment 2

base station and the test devices was tested. This test was firstly done with a window between the base station and the test devices. Secondly this test was repeated with a 14 cm thick brick wall instead of a window, see Figure 2.

Theoretically, obstacles comparable to a wall should significantly decrease the Bluetooth signal or even make it impossible to connect with devices behind such obstacles. It is very hard to predict the attenuation caused by an obstacle because every Radio Frequency (RF) signal has multiple ways to reach the other device. Our test with a window started showing problems with detecting test device 1 at a distance of 4 m. Test device 2 remained detectable up to 7 m and at larger distances it started to show some discontinuities.

The following test with a wall instead of a windowpane showed these results: at a distance of 4 m, test device 1 started to disappear and at larger distances, test device 1 was rarely detected. Test device 2 on the other hand, was much longer visible. In a range up to 7 m, test device 2 was still detectable.

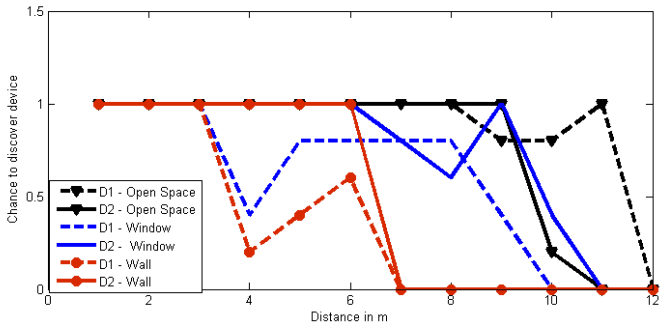


Fig. 3. Results

These results, see Figure 3, showed a general range of 10 m when the base station and test device reside in the same area hence we are working in an open space. Obstacles like walls obviously have some influence on this range. Generally we can decrease the range down to 5 m.

Consequently, when a Bluetooth device detects another Bluetooth device, this estimation will be located in a circular area with a radius up to 10 m in open space. Walls will limit the radius up to 5 m.

3) Signal Strength: RSSI values are often used in order to estimate the proper distance between 2 devices because Bluetooth does not offer an interface to extract the real received signal strength directly [12]. Theoretically, RSSI values should vary exponentially with the real distance but in practice this is not always the case [13].

Although there is no deterministic relationship between

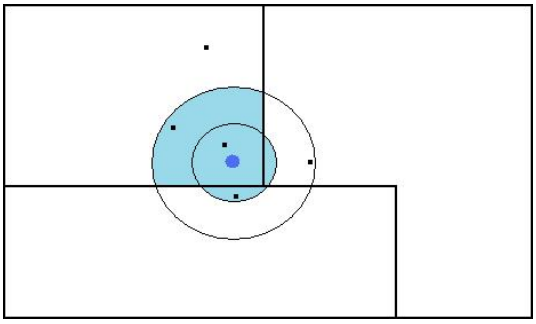


Fig. 4. Range

distance and RSSI, due to fading, reflection etc. ,there is a correlation: when the RSSI value decreases, we know the distance becomes longer and conversely; when the RSSI value increases, the distance diminishes. This information can be used to discover whether devices move away from each other, towards each other or together.

[14] shows that using RSSI values for calculating the distance between 2 devices is not reliable. Nevertheless, RSSI values could be useful to implement object binding. Object binding should only be realised when 2 or more objects are very close. At this point, the RSSI values will be higher. Nonetheless, these values will fluctuate. In this way, it is necessary to use a range of RSSI values in order to decide whether objects should be bound or not.

In this thesis, RSSI values are not used because they bring up another disadvantage: a device needs to set up a connection with the other device and this will increase the scanning time. Considering the fact that we are working with a real-time system, the scanning time should be as short as possible.

B. Opportunistic Seamless Localisation (OSL)

The opportunistic seamless localisation system combines all location related information readily available from multiple technologies such as WiFi, GSM, GPS, accelerometers [15] etc. In this paper we propose a novel method, which allows taking into account also mobile device connectivity via Bluetooth link to other devices as an additional source of location related information which may be successfully utilized by the OSL system for further improvement on location estimation reliability and accuracy. As authors presented in [3] the Bluetooth link connectivity on its own does not provide sufficiently accurate location information for most of the mobile applications. Therefore, to successfully fuse the Bluetooth connectivity information for localising Bluetooth enabled devices, a specific method described in this chapter had to be developed for efficient incorporation into the OSL system fusion location data engine. The OSL fusion engine is based on the recursive Bayesian estimation implemented as a particles filter, therefore, also a likelihood observation function used for the particles weighting was developed.

 ISBN: 978-1-61208-100-7

1) Communication: Firstly, the client scans for all nearby devices. The MAC address of every found Bluetooth device is sent to the server. In the mean time, the client keeps scanning for devices and will regularly send an update.

At the server side, every incoming MAC address will be compared to a list of known MAC addresses. In this list all primarily known Bluetooth devices are saved. Every Bluetooth device has 4 arguments, at first the MAC address, secondly a boolean to indicate whether the device is fixed or mobile, thirdly the coordinates when the device has a fixed place and at last every mobile device has an ID.

When a match between incoming MAC address and a MAC address in the list is found, these MAC addresses are saved in a list.

2) Measurement Model: The Bluetooth measurement model is designed to deal with different situations. A complete overview of this measurement model can be found in Figure 5.

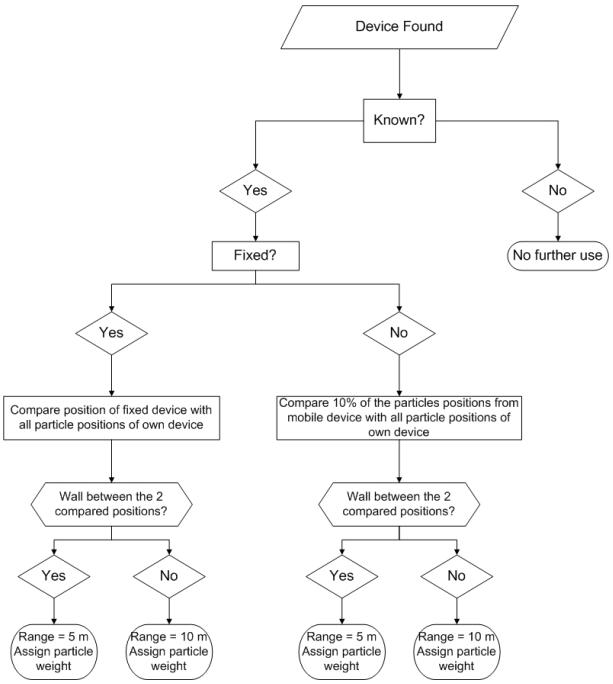


Fig. 5. Flowchart

There are 3 possible options when one or more Bluetooth devices are found. The first option happens when the found devices are unknown. These devices can not be used to localise the client device. Though, these devices can give some interesting information, such as how many devices were present at a certain time in a certain place. This is already implemented at some places such as Brussels Airport [16]. Every Bluetooth device that is discoverable will be detected by fixed antennas. In this way it is possible to measure the time necessary to move from one point to another and consequently it will be possible to calculate the waiting time to pass for example through the safety zone. When the found device is known, there are 2 options left: this device can be a fixed device, this is the second option, or a mobile device which is the third option.

Dealing with the second option, returns a fixed place with the exact coordinates of the fixed device. With the knowledge that a Bluetooth device is only visible within a certain area around that device, the weight of all particles from the client can be adapted.

Calculating the euclidean distance between every particle and the fixed device is the first step. After having calculated the distance between one particle and the fixed device, there will be a wall check. A wall has a big influence on the signal strength and for that reason it is important to know whether there is a wall between the fixed device and the particle. The choice to work with a larger or smaller range depends on the absence or presence of a wall. Based on this range, the new particle weight will be calculated.

If the third option occurs, a known mobile device is found. This device does not show exact coordinates since the location of every mobile device is predicted with a particle cloud. Depending on the situation, a particle cloud can consist out of 100 particles up to 1000 particles. Comparing every particle of the found device with every particle of the client device would be too heavy for a real-time system. For this reason, 10 percent of random particles from the found device are compared to all particles of the client device. Choosing 10 percent still gives us a reliable amount of particles. The coordinates of these particles are loaded and the distance between these particles and the client device particles is calculated. Again, we need to check if there is no wall between the particles. Based on this information, the particle weight can be calculated.

Obviously, it is possible that more than one device is found. For all those devices, previously mentioned options will be looked at and for every device, the correct option will be chosen. Working with multiple found devices, all calculated particle weights are multiplied for every client particle. In this way all found devices are brought into the calculation and the result becomes more accurate.

3) Particle Weight: According to the test results in the section 'Range', a range of 10 m will be used in open space and there will be a range of 5 m when there is an intersection of a wall. It would be inaccurate to assume that discovered devices are always in a range of 10 m with equal chances to be everywhere in that circle. For this reason, using the sigmoid function gives a more realistic image. In this case, the following functions have been used:

$$y = \frac{1}{1 + e^{x-10}} \tag{3}$$

$$y = \frac{1}{1 + e^{x-5}} \tag{4}$$

Equation (3) is used for open space. This function gradually decreases and the particle weight will be based on this

 ISBN: 978-1-61208-100-7

function, see Figure 6. Equation (4) is used when a wall between the 2 devices is detected. This function will decrease earlier because the obstacle has a big influence on the signal strength which consequently will decrease quickly.

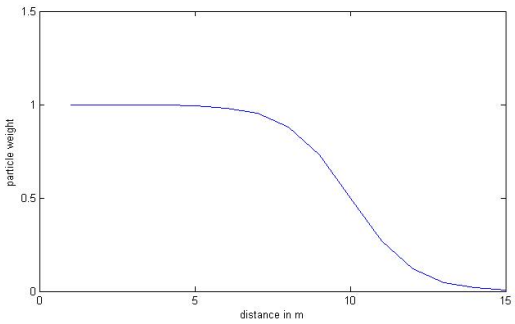


Fig. 6. Sigmoid function

This sigmoid function is S-shaped and by using this function, the use of more complex functions is avoided. Still this function gives a realistic view, due to the smooth curve.

4) Privacy issues: Localising people, often comes with privacy issues and consequently, privacy is a very important subject. Adding the Bluetooth technique does not come with any privacy issues. When scanning for Bluetooth devices, only the unique ID of the device is sent to the server. It is not possible to discover the identity of the owner of the Bluetooth device. There is no connection between the unique Bluetooth ID and the identity of the owner. A connection between those 2 can only arise when this connection is in the system made manually with authorisation of the owner.

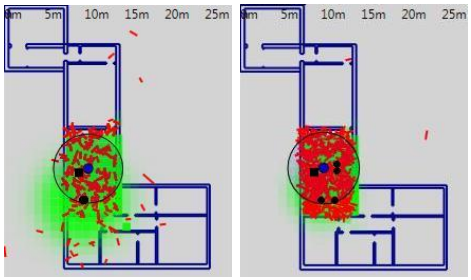
Moreover, every person with a Bluetooth device has the opportunity to shut down his/her device and thus not sending any Bluetooth signals. Most of the time, devices do not need to be shut down. In order to stop sending Bluetooth signals, it is also possible to turn off Bluetooth.

III. RESULTS

For these experiments, indoor localisation is accomplished by using WiFi and Bluetooth. In these tests, the client is only located by using Bluetooth. Multiple tests with fixed and mobile Bluetooth devices were done. The first test was done with one fixed and known device, see Figure 7(a).

The estimated position is located at the center of the circle, the real position is represented by a square and the position of the found and known Bluetooth devices is represented by dots. It shows good room level accuracy, although still some particles -representing different hypothesises- are in adjacent room

Repeating this test, but now with 4 known and fixed devices gives us a better result, see Figure 7(b). You see that all hypothesises, represented by the particles, are now inside the correct room. Using more found and known devices results logically in a more accurate estimation. This is due to



(a) Test with 1 fixed device (b) Test with 4 fixed devices

Fig. 7. Comparison between test with 1 or 4 fixed devices

trilateration. The location of every fixed device will also have an influence on the accuracy, as shown in Figure 8(c) and 8(d). 8(c) shows a good location of fixed devices, the area where the client can be located is very small and consequently more accurate. In 8(d), all fixed devices are close to each other and therefore, the area where the client can be located is still large.

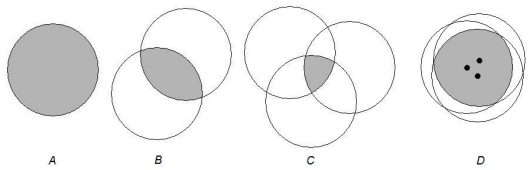


Fig. 8. Trilateration

Obviously the area where the client can be located is a lot smaller when more devices are found. This illustrates why the error rate decreases when the amount of found and known devices increases. Because we are using fixed devices only, it is possible to compare the clients particles with one exact position. Every fixed device has a known position which does normally not change. Therefore the estimated position can be easily calculated with a 100 percent certainty of the location of the fixed Bluetooth device.

Of course this is a kind of localisation which is previously already developed in other research such as [3]. But Bluetooth can be used stronger as a sensor when combined with other technologies to perform object binding.

In dynamic object binding, instead of static devices, other mobile devices will be used as references. Mobile devices do not have one exact and correct position. The likelihood of their position is estimated with a particle cloud. In order to calculate the position of the client, all particles will be compared with 10 percent of the particles from a found and known Bluetooth device. It is possible to increase the threshold of 10 percent, but using more particles will result in heavy calculations, using less particles will make the final result inaccurate.

In this test, the client location, shown in 9(a), is calculated based on the particles of another mobile device, shown in 9(b). Due to the fact that we do not have an exact position of the mobile device, we have to estimate the client position based on another estimation. Consequently, the error rate is increased,

 ISBN: 978-1-61208-100-7

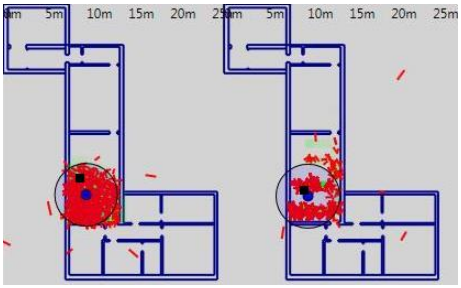


Fig. 9. Test with 1 mobile device

compared to the test with fixed devices. The error depends largely on the correctness and distribution of the likelihood of the dynamic reference device.

Dynamic object binding makes it possible to locate any found Bluetooth device without the necessity to have any other technology embedded in the device itself. Localisation information from all found devices will be used to correctly locate the client device. Merging different technologies improves the final result but within this structure, the position estimation of each device has always been created independent from other devices.

Of course we can combine dynamic reference devices and fixed devices when they are both discovered by the device. This increases the reliability of the estimation.

IV. Conclusion and Future work

In this paper, a method to realise dynamic object binding is presented. We choose Bluetooth to accomplish object binding because of its appearance in many mobile devices. For this project, the Bluetooth technology is fused with multiple other technologies in order to get an accurate localisation system. Some real experiments were done to test the Bluetooth measurement model. These results showed room accuracy when only Bluetooth was used. Obstacles like walls have a big influence on the signal strength which will make it easier to achieve room-level accuracy. This information is incorporated in the Bluetooth measurement model.

Dynamic object biding is used to locate devices which cannot be located by any other technology but can discover other devices which are located by other means. Dynamic object binding can increase the likelihood of the position of these devices.

Further research about acquiring reliable signal strength values can improve the object binding algorithm, since the error rate could be decreased by decreasing the operating range. Object biding can also be used to detect people traveling together to limit the calculations to only 1 object instead of estimating the likelihood of two distinct objects.

Acknowledgements

The research was conducted in the context of the EC FP7 LocON research project.

References

[1] M. Weyn, M. Klepal, and Widyawan, "Adaptive Motion Model for a Smart Phone Based Opportunistic Localization System," *2nd International Workshop on Mobile Entity Localization and Tracking in GPSless Environments (MELT 2009)*, pp. 50–65, 2009.

[2] A. Varshavsky, E. de Lara, J. Hightower, A. LaMarca, and V. Otsason, "Gsm indoor localization," *Pervasive and Mobile Computing*, vol. 3, no. 6, pp. 698–720, 2007.

[3] J. Hallberg, M. Nilsson, and K. Synnes, "Positioning with bluetooth," in *Telecommunications, 2003. ICT 2003. 10th International Conference on*, vol. 2, 2003.

[4] A. Nasipuri and K. Li, "A directionality based location discovery scheme for wireless sensor networks," in *Proceedings of the First ACM International Workshop on Wireless Sensor Networks and Applications*, 2002.

[5] S. Gezici, Z. Tian, G. V. Giannakis, H. Kobaysahi, A. F. Molisch, H. V. Poor, and Z. Sahinoglu, "Localization via ultra-wideband radios: A look at positioning aspects for future sensor networks," *IEEE Signal Processing Magazine*, vol. 22, pp. 70–84, 2005.

[6] F. Gustafsson, F. Gunnarsson, N. Bergman, U. Forssell, J. Jansson, R. Karlsson, and P. Nordlund, "Particle filters for positioning, navigation, and tracking," *IEEE Transactions on signal processing*, vol. 50, no. 2, pp. 425–437, 2002.

[7] J. Haartsen, "Bluetooth-The universal radio interface for ad hoc, wireless connectivity," *Ericsson review*, vol. 3, no. 1, pp. 110–117, 1998.

[8] S. Hay and R. Harle, "Bluetooth Tracking without Discoverability," in *Location and Context Awareness: 4th International Symposium, LoCA 2009 Tokyo, Japan, May 7-8, 2009 Proceedings*. Springer-Verlag New York Inc, 2009, pp. 120–137.

[9] J. Bray and C. Sturman, *Connect without cables*. Prentice Hall PTR Upper Saddle River, NJ, USA, 2000.

[10] A. Huang, "The use of Bluetooth in Linux and location aware computing," Ph.D. dissertation, Citeseer, 2005.

[11] Weidler, "Technical brief: Rugged bluetooth scanners," Motorola, Tech. Rep., 2010.

[12] S. Feldmann, K. Kyamakya, A. Zapater, and Z. Lue, "An indoor Bluetooth-based positioning system: concept, implementation and experimental evaluation," in *International Conference on Wireless Networks*, 2003, pp. 109–113.

[13] U. Bandara, M. Hasegawa, M. Inoue, H. Morikawa, and T. Aoyama, "Design and implementation of a bluetooth signal strength based location sensing system," in *2004 IEEE Radio and Wireless Conference*, 2004, pp. 319–322.

[14] J. Hallberg and M. Nilsson, "Positioning with bluetooth, irda and rfid," *Computer Science and Engineering, Luleå University of technology/2002*, vol. 125, 2002.

[15] I. Bylemans, M. Weyn, and M. Klepal, "Mobile Phone-Based Displacement Estimation for Opportunistic Localisation Systems," in *The Third International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies (UBICOMM 2009)*. IEEE, 2009, pp. 113–118.

[16] B. A. Company, "Accurate real-time information on waiting times," 2010. [Online]. Available: http://www.brusselsairport.be/en/news/newsItems/361700

 ISBN: 978-1-61208-100-7

Push-Delivery Personalized Recommendations for Mobile Users

Quang Nhat Nguyen
School of Information and Communication Technology
Hanoi University of Technology, Vietnam
quangnn-fit@mail.hut.edu.vn

Phai Minh Hoang
School of Information and Communication Technology
Hanoi University of Technology, Vietnam
minhphai87@gmail.com

***Abstract*—In application domains where the availability of items changes quickly and often (e.g., the user problem of receiving relevant promotions, events, etc.), users often find it difficult in keeping track of their desired and interested items. Recommender systems are intelligent decision support tools aimed at addressing the information overload problem, suggesting items that best suit a given user's needs and preferences. In this paper, we present our proposed mobile push-delivery recommendation methodology that is capable of proactively providing recommendations relevant to the user's preferences at appropriate context. The proposed methodology is implemented in a mobile push recommender system that helps users timely receive their desired product promotions.**

Keywords- mobile recommender system; push delivery; context-aware mobile application

I. INTRODUCTION

E-commerce sites often provide huge catalogues of diverse products and services. Hence, without the system support, users of e-commerce sites may find it difficult in making product selection decisions. This information overload problem becomes even harder for mobile users who interact with the system using mobile devices, due to the limitations of mobile devices and mobile users' limited spent time and effort. Recommender systems (RSs) aims at solving the information overload problem by providing product and service recommendations personalized to a given user's needs and preferences [1]. Most existing RSs follow the pull-delivery approach, where the user must explicitly make request for some product or service recommendations. However, in some application domains (e.g., the problem of providing interested product promotions to a given user), the availability of items changes quickly and often. In such application domains, the pull-delivery approach seems less effective in helping users keep track of their interested items, i.e., at the time of a user's request some of his interested items are not available, but when they are available (often in short durations) the user does not know.

In this paper, we present our proposed mobile push-delivery recommendation methodology that is capable of proactively (automatically) providing relevant recommendations to users at right contexts. To provide push-delivery recommendations to users, the system must decide: *what* recommendations should be pushed to a given user, and *when* the system should push these recommendations to the user. To tackle the first problem, our proposed recommendation methodology integrates both long-term and session-specific user preferences and exploits a critique-based conversational approach [2]. The long-term user preferences are inferred from past recommendation sessions, whereas the session-specific user preferences are derived from the user's critiques to the provided recommendations in the current session. To deal with the second problem, the system models a push context as a case, and uses the Case-Based Reasoning (CBR) problem-solving strategy [3], i.e., a machine learning approach, to exploit (reuse) the knowledge contained in the past push cases to determine the right push context for the current case. Our proposed methodology has been implemented in a mobile push recommender system that helps users timely receive their interested product promotions.

The remainder of the paper is organized as follows. In Section 2, we discuss some related work on recommender systems and push-delivery information systems. In Section 3, we introduce the formal representations of product promotions, the user profile and the user query. In Section 4, we present our proposed mobile push-delivery recommendation methodology. Finally, the conclusion and future work are given in Section 5.

II. RELATED WORK

Recommender Systems (RSs) are decision support tools that help users find and select their desired products and services when there are too many options to consider or when users lack the domain-specific knowledge to make selection decisions. Traditional recommendation approaches include: *collaborative*, *content-based*, and *knowledge-based* [1]. RSs have been very effective and popular tools in well-known commercial websites, such as Barnes&Noble.com, eBay.com, Amazon.com, etc.

A push-delivery information system is a system that automatically delivers (i.e., pushes) the information to users without their request. The push-delivery model appears to be effective in application domains where the availability of items changes often and quickly, because it helps users timely receive their interested information. However, if the system pushes uninterested information to a user, or even pushes interested information to the user but at inappropriate contexts, there is a high risk that this push-based delivery will annoy the user (i.e., considered as a spam). Hence, for push-delivery RSs, to provide personalized recommendations and reduce the spamming issue, the system must push *only relevant and targeted* information to the user *at right contexts (time and location)*. In some previous approaches, the system just pushes all objects (or items) that locate near the user's position, without regarding

 ISBN: 978-1-61208-100-7

his preferences [4], [5]. In other previous approaches, the system, though takes into account the user's preferences, but does not estimate right contexts to push, i.e., the system always pushes advertisements to the user when he is close to (or inside) the store [6], [7]. Ciaramella et al. [8] presented a mobile services RS that uses a rules table to determine a user's situation, but the system pushes all services associated with the determined situation to the user without regarding his preferences. The information service system presented in [9] determines the push time based on a decision table that is the same for all users.

In our proposed approach, the pushed recommendations are personalized for each user (i.e., suitable for his preferences), and the push context is determined based on the system's learning from past push cases. Hence, the system's push-context determination is personalized for each user. Moreover, all the push-delivery information systems mentioned above follow the single-shot strategy, where the system computes and pushes to the user the information, and the session ends. In our proposed approach, a push session, after the user accepts to view the pushed recommendations, evolves in a dialogue where the system's recommendations interleave with the user's critiques to these recommendations [2]. Such critiques enable the system to better understand the user's preferences, and hence to provide more suitable recommendations to the user.

III. FORMAL REPRESENTATIONS

A. Product Promotion Representation

In our recommendation problem, the system's recommendation aims at promotions, whereas the information of promoted products and gifts is supplemental. In particular, a promotion, represented *hierarchically*, consists of the three main components: the promotion's information, the promotion's promoted product(s) and the promotion's gift(s). In this hierarchical representation, each component is represented by its own sub-components and features. (Due to the paper's space limit, we present here only the first and second levels of the hierarchical representation.)

PROMOTION = (*PROMOTION-INFO*, *PROMOTED-PRODUCTS*, *GIFTS*)

The component *PROMOTION-INFO* stores the information of the promotion:

PROMOTION-INFO = (*Prom-Type*, *CONDITION*, *DURATION*, *PROVIDER*);

where the feature *Prom-Type* represents the type of the promotion, the sub-components *CONDITION*, *DURATION* and *PROVIDER* represent the promotion's condition, available duration and provider, respectively.

The component *PROMOTED-PRODUCTS* represents the set of the promoted products and their quantity (i.e., required to buy in order to get the promotion):

PROMOTED-PRODUCTS = {(*PRODUCT*, *Quantity*)};

where the sub-component *PRODUCT* is represented by the three features: the promoted product's category (e.g., laptop, TV, etc.), identifier and price.

The component *GIFTS* represents the set of the gifts of the promotion:

GIFTS = {(*Gift-Type*, *GIFT*)};

where the value of the feature *Gift-Type* defines the (structured) content of the sub-component *GIFT*.

B. User Profile Representation

The user profile stores the user's long-term preferences that are exploited by the system to build the initial representation of the user query. The user profile, *hierarchically* represented, consists of the three components that represent the user's long-term preferences on promotions, promoted products and gifts.

U = (*PROMOTION-PREF*, *PRODUCT-PREF*, *GIFT-PREF*);

where the component *PROMOTION-PREF* stores the user's long-term preferences on promotions types, condition types and providers; the component *PRODUCT-PREF* stores the user's long-term preferences on category and price of promoted products; and the component *GIFT-PREF* stores the user's long-term preferences on gift types.

C. User Query Representation

The user query (*Q*) representation encodes the system's understanding (i.e., guess) of the user's session-specific preferences. In a session, at every recommendation cycle the system uses the query *Q* to compute the promotions recommendation list that is then shown to the user.

In our approach, the user query *Q* consists of the two (structured) components: the favorite pattern (*FP*) and the component and feature importance weights (*W*).

$$Q = (FP, W)$$

The favorite pattern *FP*, *hierarchically* represented, consists of the three components that represent the user's session-specific preferences on promotions, promoted products and gifts. The structure of *FP* is similar to the structure of the user profile (*U*) representation, except that *FP* includes additionally the sub-component *DURATION* (i.e., to represent the user's session-specific preference on promotion available duration) and the feature *Distance* (i.e., to represent the user's session-specific preference on distance to promotion provider).

The weights vector *W* is represented *hierarchically* corresponding to the representation of *FP*. For each representation level, the weight of a sub-component (or a feature) models how much important the sub-component (or the feature) is for the user with respect to the others.

IV. RECOMMENDATION METHODOLOGY

In our approach, a recommendation session starts when the system's promotions catalogue is updated (with new promotions) or when the user is close to a promotion provider's store, and ends when the user quits the session. The overview of the recommendation process is shown in Fig. 1.

When the session starts, the system builds the initial query representation (Q^0) exploiting the user's long-term preferences stored in the user profile. In this initialization

 ISBN: 978-1-61208-100-7

step, the values of the features of FP^0 are set by the values of the corresponding features in the user profile (U). In addition, the values of the sub-component *DURATION* and the feature *Distance* are set to unknown to indicate that at the beginning of the session the system does not know about the user's session-specific preferences on promotion available period and distance to provider.

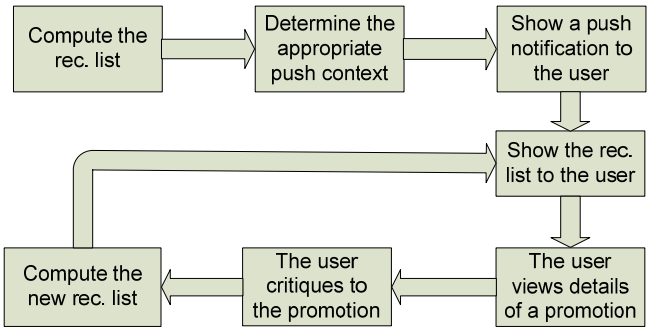


Figure 1. The overview of the recommendation process

The importance weights vector W is initialized by exploiting the history of user critiques. The intuitive idea is that a feature (or sub-component)'s initial importance weight is proportional to the frequency of the user critiques expressed on that feature (or sub-component). In Fig. 2, it illustrates an example of a sequence of critiques that a user makes in a recommendation session. A recommendation session evolves in recommendation cycles, where each cycle comprises the stage where the recommended promotions are shown to the user (see Fig. 3-a) and the successive stages where the user browses the details of a promotion and criticizes it (see Fig. 3-b).

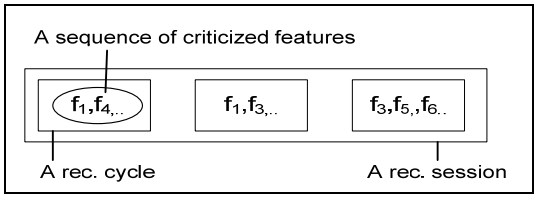


Figure 2. A sequence of critiques in a session

We note that for each representation level (of W) the system computes the importance weights of the features (or sub-components) at that level. For example, for the level of *PROMOTION-INFO*, the system computes the importance weights of the feature *Prom-Type* and the sub-components *CONDITION*, *DURATION* and *PROVIDER*.

First, the system computes the importance weight of feature (or sub-component) f_i, given session s_k of user u_j:

$$w_i(u_j, s_k) = \frac{1}{\lambda_k} \cdot \sum_{l=1}^{\lambda_k} \frac{Ctz(f_i, u_j, c_l)}{\alpha^{(\lambda_k - l)}} \qquad (1)$$

where:

- c_l: a recommendation cycle of session s_k;
- λ_k: the length (i.e., the number of recommendation cycles) of session s_k;
- $Ctz(f_i, u_j, c_l) = 1$, if at cycle c_l user u_j made a critique on feature (or sub-component) f_i ,
  = 0, if otherwise;
- α (>1): a parameter to increase the importance of latest critiques (i.e., those appear later in session s_k).

Next, the system computes the importance weight of feature (or sub-component) f_i over all the sessions of user u_j.

$$w_i(u_j) = \frac{1}{\| S(u_j) \|} \cdot \sum_{k=1}^{\|S(u_j)\|} \frac{w_i(u_j, s_k)}{\beta^{(\|S(u_j)\| - k)}} \qquad (2)$$

where :

- $S(u_j)$: the set of historically ordered sessions of user u_j.
- β (>1): a parameter that shapes how fast the importance of an old session decreases over time.

The system uses the initial query Q^0 to compute the (initial) recommendation list for the user, by ranking the available promotions to their similarity to (FP^0, W^0). The ranking is done, using a similarity function computed over the hierarchical representation described in Section 3, so that the more similar to (FP^0, W^0) a promotion is the higher it appears in the ranked list. In case of ties, the promotion provided by the provider closer to the user's position is ranked higher. Only k best promotions in the ranked list, i.e., those most similar to (FP^0, W^0), are included in the recommendation list.

After computing the recommendation list, the system must determine when it should push this list to the user. In our approach, this push-context determination is done based on the Case-Based Reasoning (CBR) problem-solving strategy [3]. The CBR is used to exploit (reuse) the knowledge contained in the past push cases. In our methodology, each push *case* is modeled by two parts: the *problem description* and the *solution*. In particular, the problem description of a case contains information of: 1) the time-slot of the push, 2) the list of providers that provide promotions contained in the recommendation list, 3) the user's distances to those providers, and 4) the user's (long-term) preferences to those providers. The solution of a case indicates the decision of the user: 1) the user accepts to receive (i.e., view) the recommendation list, or 2) the user rejects to receive.

To estimate (i.e., predict) an appropriate push context, the system identifies: 1) the set of m past push cases most similar to the current one in that the users accepted to receive the recommendation list (denoted as $C^{Accepted}$), and 2) the set of m past push cases most similar to the current one in that the users rejected to receive the recommendation list (denoted as $C^{Rejected}$). Then, the system computes the *push degree* (i.e., the confidence level to push) and the *not-push degree* (i.e., the confidence level to not push) for the current case.

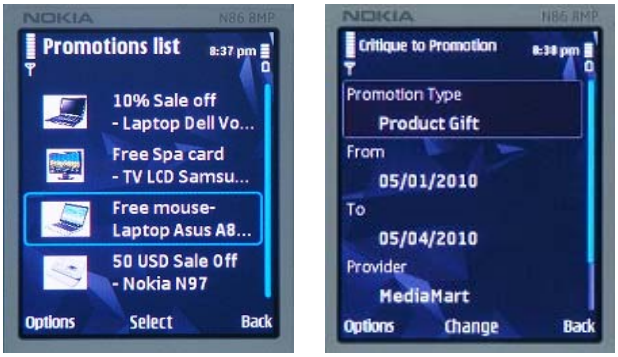
$$push - \deg ree(C^*) = \frac{1}{m} \sum_{C \in C^{Accepted}} Sim(C^*, C) \qquad (3);$$

$$not - push - \deg ree(C^*) = \frac{1}{m} \sum_{C \in C^{Rejected}} Sim(C^*, C) \qquad (4)$$

where C^* is the current case, and C is a past case. If (*push-degree*(C^*) – *not-push-degree*(C^*)) ≥ θ (i.e., θ is a predefined push-confidence threshold), then the system sends a push notification to the user. Otherwise, the system does not, and

the recommendation list is saved in the pending list for the user (i.e., at the next time-slot, the system re-estimates whether or not to send the push notification to the user).

Given the push notification sent to the user's mobile device, he can decide to accept the push, or postpone it, or reject it. In case the user accepts the push, the system first stores (i.e., records) the current push case in its case base for the future uses, and then shows the recommendation list to him (see Fig. 3-a). In case the user postpones the push, he can specifies the later appropriate time slot (i.e., postponed by time) or the appropriate distance to provider (i.e., postponed by distance) to receive the recommendation list. In case the user rejects to receive the push, the system stores the current push case in its case base.



a) Recommendation list b) User critique

Figure 3. The mobile user interface

When the recommendation list is shown to the user (see Fig. 3-a), for each recommended promotion the system shows an icon corresponding to the promoted product's category, the gift's abstract information and the promoted product's name. The user can select a recommended promotion to see its details. After the user views a promotion's details, if he accepts the promotion, then this promotion is added to his Selection List, and he can view another recommended one or quit the session. If the user is somewhat interested in the promotion, but one (or some) of its features does not completely satisfy him, then he critiques to the promotion to specify (i.e., express) his preferences on these unsatisfactory features (see Fig. 3-b). In Fig. 3-b, for example, the user critiques to the promotion to indicate his preference on the promotion's type. By critiquing, the user at the next recommendation cycle (of the current session) is recommended with other promotions that are "closer" to his preferences. Such critiques help the system adapt its previous user-query representation (i.e., guess) (Q) to the user's new preferences, and re-compute some new recommended promotions based on this adapted user query. The new list of recommended promotions is then shown to the user, and the system proceeds to the next recommendation cycle.

When the user quits the session, the system exploits the information of his expressed critiques and selected promotions in the current session to update the user profile. This user profile update allows the system to refine its understanding of the user's long-term preferences, and hence better serve the user in the future.

V. CONCLUSION AND FUTURE WORK

Mobile recommender system aims at providing recommendations to users at anytime and anywhere, exploiting the popularization of mobile devices and their unique features like mobility, high targeting and personality. In this paper, we have presented our proposed methodology for proactively providing personalized recommendations to mobile users at appropriate contexts. The integration of the user's long-term and session-specific preferences enables the system to provide relevant recommendations, and the push-context determination helps the system deliver these recommendations to him at right time and location. This mobile push recommendation methodology has been implemented in a recommender system that helps users timely receive their interested product promotions.

We shall run a usability evaluation of the implemented system to test the effectiveness of our proposed methodology and the usability of the implemented system. In addition, we will need to find the best way to visualize the push notification on the user's mobile device.

ACKNOWLEDGMENT

The financial support for this research work from the Vietnam National Foundation for Science and Technology Development (NAFOSTED) under the grant number 102.01.14.09 is gratefully appreciated.

REFERENCES

[1] R. Burke, "Hybrid web recommender systems", in The Adaptive Web: Methods and Strategies of Web Personalization, P. Brusilovsky, A. Kobsa, and W. Nejdl, Eds. Heidelberg: Springer, 2007, pp. 377–408.

[2] F. Ricci and Q. N. Nguyen, "Acquiring and revising preferences in a critique-based mobile recommender system", IEEE Intelligent Systems, vol. 22, n. 3, pp. 22–29, May-June 2007.

[3] A. Aamodt and E. Plaza, "Case-based reasoning: foundational issues, methodological variations, and system approaches", AI Communications, vol. 7, n. 1, pp. 39–59, March 1994.

[4] N. Hristova and G. O'Hare, "Ad-me: wireless advertising adapted to the user location, device and emotions", in Proc. 37th Annual Hawaii Int. Conf. System Sciences, 2004, pp. 285–294.

[5] L. Aalto, N. Göthlin, J. Korhonen, and T. Ojala, "Bluetooth and WAP push-based location-aware mobile advertising system", in Proc. 2nd Int. Conf. Mobile Systems, Application, and Services, 2004, pp. 49–58.

[6] S. Kurkovsky and K. Harihar, "Using ubiquitous computing in interactive mobile marketing", J. Personal and Ubiquitous Computing, vol. 10, n. 4, pp. 227–240, May 2006.

[7] J. E. de Castro and H. Shimakawa, "Mobile advertisement system utilizing user's contextual information", in Proc. 7th Int. Conf. Mobile Data Management, 2006, pp. 91.

[8] A. Ciaramella, M. G. C. A. Cimino, B. Lazzerini, and F. Marcelloni, "Situation-aware mobile service recommendation with fuzzy logic and semantic web", in Proc. 9th Int. Conf. Intelligent Systems Design and Applications, 2009, pp. 1037–1042.

[9] S. Pinyapong, H. Shoji, A. Ogino, and T. Kato, "A mobile information service adapted to vague and situational requirements of individual", in Proc. 7th Int. Conf. Mobile Data Management, 2006, pp. 20–22.

 ISBN: 978-1-61208-100-7

m-Physio: Personalized Accelerometer-based Physical Rehabilitation Platform

Iván Raso, Ramón Hervás, José Bravo
Technologies and Information Systems Department
Castilla - La Mancha University
13071 Ciudad Real, Spain
Email: ivanrasodiazguerra@gmail.com; ramon.hlucas@uclm.es; jose.bravo@uclm.es

***Abstract*— This paper proposes a rehabilitation system based on new technological tendencies in the mobility and ubiquitous computing areas. Specialists and patients related to rehabilitation area can use this proposed system to improve the fulfillment of exercises and the supervision of rehabilitation tasks. An important current problem is that sometimes these activities cannot be performed efficiently due to the lack of time or the large distances between patient homes and rehabilitation centers. We have developed our system using a mobile device and a bracelet to capture patient's rehabilitation relevant data. As a pre-process procedure, raw data output by mobile device accelerometer is filtered, and then we use the technique called Dynamic Time Warping to train and recognize movements. Based on this recognition, patients can perform rehabilitation without the continuous specialist's surveillance and can be sure of its accuracy. Experimental results show us that our system is able to adapt itself dynamically to the peculiarities of each user and enhance healthy rehabilitation in a proactive way.**

Keywords- ubiquitous computing; accelerometry; physical-rehabilitation; mobility;

I. INTRODUCTION

The ubiquity of mobile devices has led to the emergence of personalized and adaptive services that are able to respond particular needs of each specific user. These services allow us to develop a wide range of proactive applications such as ambient assisted living services (e.g., assistance to elderly people [15] and chronic diseases assistance [22]), entertainment (e.g., mobile quiz games [16]), and smart homes (e.g., personalized home control [17] and visualization services [23]).

A principal characteristic to take into account in our work is the capability of monitoring user movements. Motion recognition is a discipline that has been around us for years in the scientific community. Some of the related works address issues such as handwriting recognition, recognition of hand gestures, and monitoring of the user activities. Some of these researches have in common with our work the use of one particular technology: the accelerometry. Accelerometers are being used in many sectors and, due to the fast development in sensor technology, it is possible the integration of these sensors into every day devices [1], for example, into mobile devices.

Focusing on the rehabilitation area, patients usually have to move about their rehabilitation center several times, but sometimes, factors such as lack of time and large distances affect the number of visits to their specialists, and consequently affect in the quality of the rehabilitation. Moreover, some patients suffer a slight incapacitate and have to perform part of their rehabilitation at home and they also need medical examination to check their evolution. Also, the well-known health care systems overcharge can be lesser by means of this kind of m-Health systems.

The main goal of this paper is the development of a novel system that helps the kinds of patient mentioned above whenever realize their rehabilitation. Besides, physical rehabilitation specialists can improve the monitoring and supervision of tasks by using our web-based system in their rehabilitation center. Our whole system (web-based and mobile applications) lets specialists pay a better attention to patients and reduce the problem of performing rehabilitation without the attentive specialist's eye.

In this paper, we employ the iPhone, one of the first mobile devices equipped with an accelerometer. Later, several mobile devices such as RIM Blackberry Storm, Nokia N95, and Sony Ericsson W910 were equipped with this kind of sensor. They basically use accelerometers for user interaction with games [3]. Few relevant applications employ accelerometry with other purposes. For example, Sony Ericsson's shake control allows the changing of songs by shaking the mobile device. However, this paper presents a novel application that introduces m-Health area into a new challenge that has not been deeply explored, the mobile-based rehabilitation.

This paper is organized in five sections: Section 2 presents some related works that apply accelerometry to the rehabilitation area. Section 3 presents and explains the proposed applications. Section 4 presents evaluation results, and Section 5 discusses about the future works and the conclusions.

II. RELATED WORK

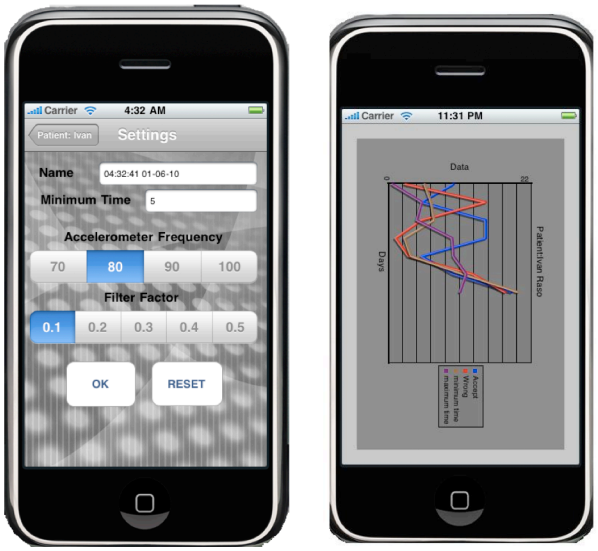
Accelerometry has become a powerful choice for evaluating variability of person's movement using these kinds of sensor that provide a non-invasive method of measurement and have a successful accuracy [4].

The entertainment sector is one of the most influenced by this technology as we can see in the Nintendo Wii game console that uses Wii Remote and Nunchuk to control avatars in games by means of natural gestures, and the PlayStation3 with its SixAxis and DualShock 3 controllers. Other sectors such as motor industry use these sensors to control ABS systems, airbags, and for checking the correct

 ISBN: 978-1-61208-100-7

working of a machine. Transport industry also uses accelerometry to check wherever the merchandises suffer misshapes and their condition or their integrity has been damage by it [2].

Several works [5][6][7][8][10] have shown the accuracy of the accelerometer-based physical rehabilitation monitoring such as limb's motion, gait analysis before strokes, and other illness or accidents that cause malfunction in the physical condition of a patient.



(a) (b)

Figure 1. Application parameters (a) and stadistics gathered from the rehabilitation process (b).

Some related works to this paper are Wiihab [20], Telefonica's Rehabitic [9], and the Arteaga et al. proposal [19]. These projects use the accelerometers to help in the rehabilitation process. The Wiihab is used as an interaction object that encourages people to move their limbs. However, Rehabitic is made up of several accelerometers and a central device that saves the data received from the sensors, all complemented by a web page that the specialist uses to control the patients. This system gives important information relevant to help the patient in the rehabilitation exercises and contribute to the specialist's decisions about the evolution of the patient. Arteaga et al. propose a set of monitoring devices, each of which comprises of an accelerometer and a beeper, LED light and vibrator to provide redundant modes of inappropriate posture warnings that would hopefully trigger self-correction.

Other related contributions are the O'Donovan et al. [7] and Choquette et al. [18] proposals that allow scientific monitoring limb's motion and measures of heart rate using Body Area Network (BAN). Nowadays, the BAN devices use wireless technology and have become into Wireless Area Network (WAN) [5]. We have found many interesting similarities between these systems and our approach. However, the invasive characteristic of these systems prejudices their common use due to its numbers of devices and the tedious task of getting dressed with them. At this point, our application contributes to the rehabilitation area with a less invasive system than the above-described projects do. Moreover, a common problem with these proposals is the high effort needed to deploy the systems. Thus, our solutions achieve the objective of ubiquitous rehabilitation performance and monitoring that enhance the accuracy, less invasively and reducing infrastructural needs.

III. SYSTEM OVERVIEW

The mobile application (Figure 1) has been developed for iPhone 2G devices. This device includes an LIS302DL MEMS smart digital accelerometer [21]. It has 3-axis (X,Y,Z) and includes dynamically user selectable full scales from ± 2g to ± 8g.

According to some related studies [6][7][5], one of the best options to wear the accelerometers to the patient is a wearable system. The examples mentioned above were ruled out because their tedious wear system. We decided to use a bracelet that people use together with the iPhone for jogging or fitness.

Before presenting the principal system's components, we define what kind of rehabilitation exercises are related to this paper and their particular characteristics:

- Exercises end in the same point that they start.
- When a patient performs the exercise, it always starts at the same point approximately.
- The motion of the exercise will be slow due to the fact that the patient is doing rehabilitation.

Additionally, an exercise can be classified in four types:

- Correct exercise: The patient performs the exercise according to the pattern generated in the training process and imposed by the specialist.
- Wrong exercise: The patient performs the exercise according to the time limits but it was not the expected exercise according to the stored pattern.
- Exercise exceeds the maximum time: The patient performs an exercise but out of the maximum time allowed.
- Exercise does not exceed the minimum time: The patient completes an exercise but does not pass the minimum time imposed by the specialist.

A. Filtering

It is necessary to use a filter because the raw data of the accelerometer is noise and redundant. Consequently, we have chosen the following smoothing function for each axis (Equation 1):

$$S(A_t) = \begin{cases} A_t, & \text{if } t = 0 \\ A_t * \alpha + S(A_{(t-1)}) * (1.0 - \alpha)), & \text{if } 0 < t \leq T \end{cases} \quad (1)$$

$S(A_t)$ is the filtered acceleration vector output and At is the acceleration raw vector output, which is acquired by the interaction device at time t. Besides α is a smoothing factor in the range from 0 to 1. The α factor is critical for acquiring valid data to be analyzed in the pattern recognition process.

 ISBN: 978-1-61208-100-7

We have performed several experiments to select an α factor for rehabilitation exercises or other movements with similar characteristics; different kinds of movement may need additional studies to select a valid α factor. Figure 2 shows different graphs captured by the iPhones's accelerometer while a patient is performing a rehabilitation exercise. Each graph represents the same exercise with different α factors. The most representative capture of the rehabilitation exercise was the option (a), with α factor 0.1, because it filters peaks (remarked with circles) that not contribute to define the rehabilitation exercise. In the set of exercises performed for this paper we chose the option (a). On the other hand, fewer values of the α factor are not characteristic to the movement represented by the accelerometer axes.

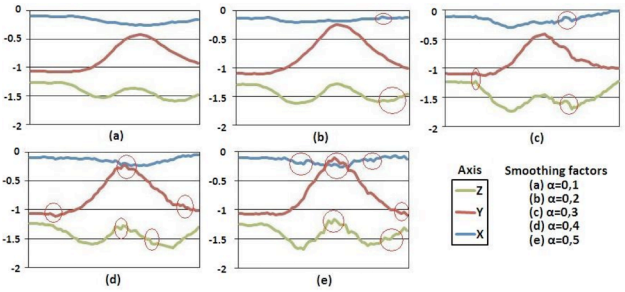


Figure 2. Test of different smoothing factors.

B. Pattern Recognition

The recognition of motion is a kind of pattern recognition. This recognition can be performed in several ways such as brute force, fuzzy logic, Gabor wavelet transform, hidden Markov model, support vector machine, and neural networks [3].

Instead of these examples, we have decided to base our development in the dynamic time warping (DTW) algorithm because it requires a simple training and its effective has been proved in many researches [11], [8]. Besides, it has been used for writing recognition [12] with great results, as well as for speech recognition [13].

DTW computes the distance between two exercises A and B by finding the minimum path that will be represented with a numerical value. In our application, the averaged Euclidean distance defines the cost between two different points Ai and Bj from the rehabilitation exercise.

C. Segmentation

The segmentation mechanism is used to determinate the beginning and end of an exercise. Some related works use the segmentation approach [3][1]. In our case, the segmentation is necessary because the patient has to know if he/she is beginning the exercises in the correct position as well as the device has to know when the exercise begins and ends. In order to achieve the segmentation mechanism, authors such as Schlömer [14] forces users to touch a button for detecting the beginning and end. Other authors [3] use mathematical equations such as the equation (2). According to the author when this equation is bigger than 0.3 the exercise starts and ends when drops to bellow 0.1.

$$D = \sqrt{((x_k - x_{(k-1)})^2 + (y_k - y_{(k-1)})^2 + (z_k - z_{(k-1)})^2} \quad (2)$$

The first method is not directly applicable in our system; patients cannot touch the mobile each time they realize a rehabilitation exercise because it may distort the pattern recognition process. On the other hand, the equation (3) is more interesting but neither applicable to our purposes; we manage several consecutive exercises and it requires that patients cannot stop for a while until the accelerometer drops to the value 0.1.

Our segmentation method is partially based on the push-button approach mentioned above and follows these steps:

- Patients have to touch the mobile's screen when they are ready to start the exercise.
- Once the patient touches the screen, the mobile device starts a countdown (five seconds) to allow the patients gets ready. For example, it could be possible that the patients have to make a rehabilitation exercise with their legs and then they need time for returning to the start position.
- After the mobile device countdown, it starts to calculate the exercise beginning and end, taking enough samples to represent these facts. The number of chosen samples was 30 after a wide testing process. In the rehabilitation and training process the segmentation is different from the capture steps because it is unnecessarily taken this number of samples again.
- As soon as patients finish the exercise, the mobile device uses the last sample to recognize the end of the rehabilitation exercise.

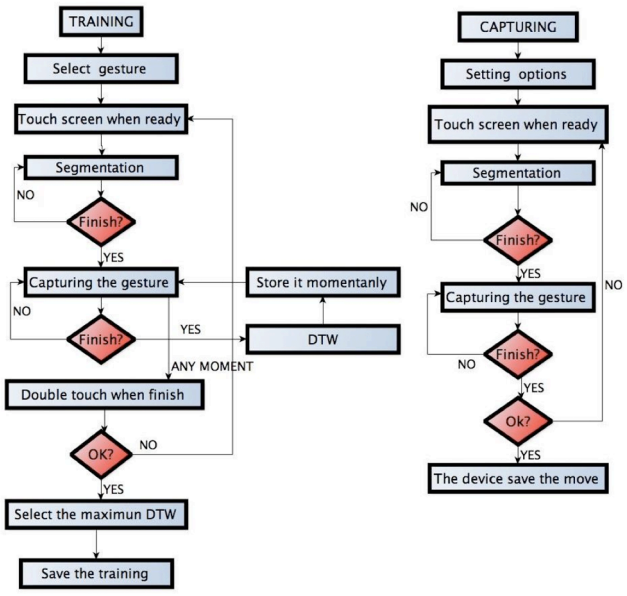


Figure 3. Flow charts representing the steps to capture and train an exercise.

This method has not only been designed to control the beginning and the end of exercises, but also to validate the device's position whenever patients are performing rehabilitation at home. If the position of the mobile device is detected as wrong, the mobile device does not start the countdown and notifies the warning to the patient. In this case, he/she has to wear the device again. Otherwise it could be fatal to the rehabilitation. This step is only performed in the rehabilitation and training process.

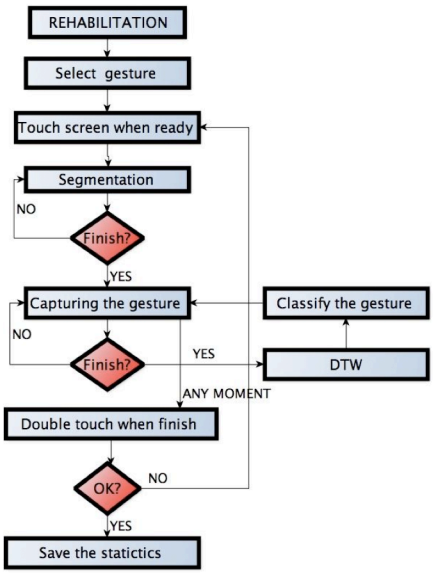


Figure 4. Flow chart that represents the steps to perform the rehabilitation process.

D. Rehabilitation Steps

Three important modules compound our application: exercise capture, exercise training and personal rehabilitation. Each module depends on the preceding results. The relevant modules' steps in response to the interaction of the patients are explained in the flow charts shown in Figure 3 and Figure 4.

- *Exercise capture*: Before using this module, the patient has to wear the mobile device and the specialist explains her/him the steps of the rehabilitation exercise. The specialist can set the movement's minimum time, the accelerometer's frequency, the movement's name and the smoothing factor α. The smoothing factor recommended after our set of test is 0.1 and the frequency is 80Hz. The capture of the exercise give to the mobile device the main configuration of the movement that is used in the following modules. Besides, once the exercise is captured, the mobile device completes the necessary information to enable the next steps of the system and stored the pattern that represents the particular rehabilitation exercise. This information includes the move's maximum time and the accelerometer's data corresponded to the exercise. This process is described in Figure 3 (left)
- *Exercise training*: Once the exercise is stored in the mobile database, it is necessary to train the exercise for being recognized when the patient begins to perform the rehabilitation. Whenever patients are performing the training, the mobile device acquires all the exercises performed by them and applies the DTW algorithm to analyze the movements. This part has to be performed under the supervision of the specialist. Depending on the specialist criteria, the training can be adapted to the patient needs. The specialist can suggest the patient not be accurate in the motion or, on the other hand, the specialist can force patients to perform more precise exercises. In more detail, if the training was hard because the injury was important, the rehabilitation will need an accurate exercise, otherwise if the injury was less relevant, the training will be leak. As we mentioned before, these decision belong to the specialist criteria. This process is detailed in Figure 3 (right)
- *Personal rehabilitation*: The personal rehabilitation is the most important step of our system. This module captures the exercises that patients perform in their rehabilitation process and classify them. This process is presented in Figure 4. There are four kinds of output to a patient's exercise and were defined previously. Once the rehabilitation ends, the mobile phone stores all the outputs and analyzes them to allow the specialist controls the patient's rehabilitation. Additionally, the mobile device synchronizes all the information with a centered database. This information can be accessed via the web application.
- Web application: The incorporation of the web application to the system complements the supervising cycle, giving to the specialists an efficient method to follow the patient's evolution.

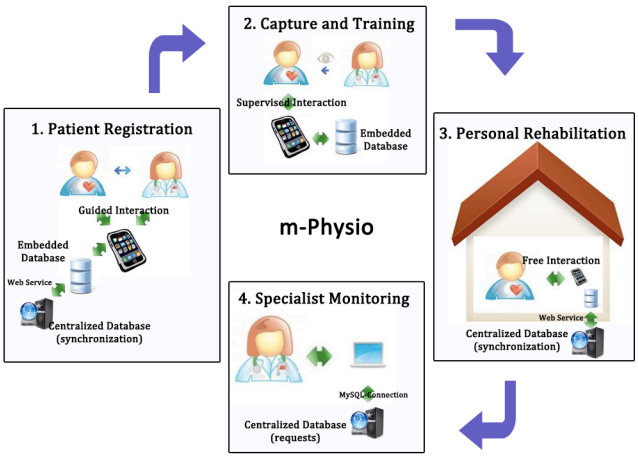


Figure 5. Principal steps on mPhysio rehabilitation process.

We now show the mainly steps for using our application to perform the rehabilitation process at home. First, a patient have to move about the rehabilitation center and the specialist studies the patient's case to decide the suitableness of our system in the specific patient's rehabilitation. Then,

the specialist registers the patient's personal data into the mobile device. Once all the needed information is stored in the mobile device, the specialist gaits the patient with the capturing and training system's steps. When the specialist decides the training is enough, the patient can comeback to his/her home and performs the personal rehabilitation process. Whenever the patient ends his/her rehabilitation session at home, all the data is stored in the mobile device and is sent to the centered database through web services. Finally, the specialist can supervise the patient's rehabilitation evolution by means of the developed web application. If the specialist thinks the patient is performing significant errors, he/she could call or send a message to the patient advising him/her, and if required, set up an appointment in the rehabilitation center. All these steps are summarized in Figure 5.

IV. RESULTS AND EVALUATION

In order to analyze the patient's experience with the application and its accuracy, we have tested it with five patients and two different rehabilitation exercises. The population includes one child, two teenagers, and two adults (with the ages of 43 and 64). The teenager users were familiar with the technology, while the other users were not familiar with this kind of system and device.

TABLE I. REHABILITATION AVERAGE BASED ON THE TWO EVALUATED EXERCISES AND FIVE PATIENTS

Day	Correct	Wrong	Min Time	Max Time
1	23.33 %	33.33 %	26.67 %	16.67 %
2	30.00 %	36.67 %	20.00 %	13.44 %
3	36.67 %	33.33 %	20.00 %	10.00 %
4	36.67 %	33.33 %	20.00 %	10.00 %
5	33.33 %	33.33 %	20.00 %	13.33 %
6	36.67 %	33.33 %	16.67 %	13.33 %
7	40.00 %	30.00 %	16.67 %	13.33 %
8	46.67 %	26.67 %	13.33 %	13.33 %
9	50.00 %	26.67 %	13.33 %	10.00 %
10	56.67 %	26.67 %	10.00 %	6.67 %
11	60.00 %	23.33 %	10.00 %	6.67 %
12	73.33 %	20.00 %	0.00 %	6.67 %
13	76.67 %	20.00 %	0.00 %	3.33 %
14	83.33 %	16.67 %	0.00 %	0.00 %
15	93.33 %	6.67 %	0.00 %	0.00 %

The tested exercise were a shoulder and leg movement shown in Figure 6. The exercises were repeated 30 times along 15 days, which provided 450 examples for each exercises and each patient. The training range was from 10 to 20 repetitions. The results in Table 1 and Figure 7 present the evolution of the patient's rehabilitation. The first days, only one out of every four rehabilitation exercises were performed correctly. Without using m-Physio or another monitoring system, patients are not aware of the incorrect development of the rehabilitation process neither the specialist. This fact brings as consequence an inadequate physical recovery and, in some cases, it may worsen the injury. Our system guides patients since the first day of rehabilitation and enables the enhancement of the performed exercises. Moreover, the specialist can supervise this process and take part whenever necessary.

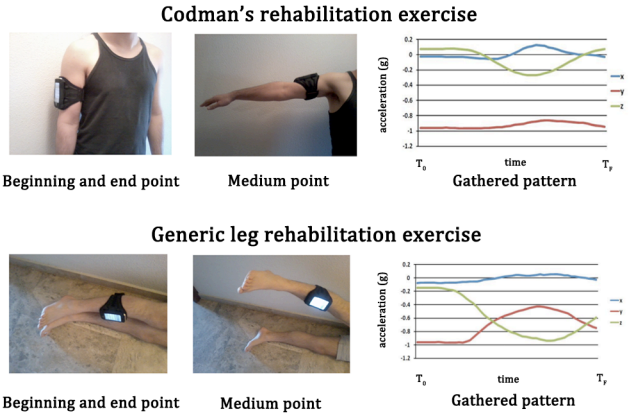


Figure 6. Shoulder and leg rehabilitation exercises.

Focusing again on tested exercises, these results show a high accuracy rate of 76.67% when users were using the application along 13 days and it improves during the next days rising up to 93% in the 15th day.

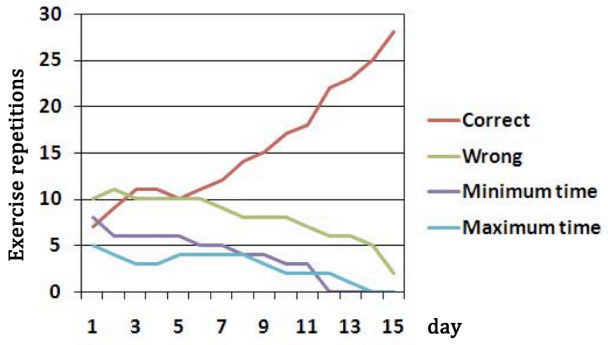


Figure 7. Daily evolution of the patients' rehabilitation during the tests.

V. CONCLUSION

In this paper, we have presented and evaluated a mobile-based rehabilitation system that can be used in rehabilitation centers for improving control and supervision. The proposed system is completely customizable, so the specialist can choose the position of the device, the frequency, minimum and maximum time of the rehabilitation exercises and the accuracy of the patient when they are performing the rehabilitation at home or without the continuous specialist's surveillance at rehabilitation center. Since the applied pattern recognition and segmentation techniques have been proposed and studied previously, we have analyzed their practical application to physical rehabilitation and we have optimized these techniques to this kind of movements.

One of the future works of our system includes the improvement of the segmentation process. A new technology implanted in the new mobile devices can be particularly helpful for the recognition and validation of the exercise's

 ISBN: 978-1-61208-100-7

beginning and end. Moreover this new technology helps patients to wear the mobile device at home. This technology is the gyroscopes that being used together with the accelerometers can enhance the physical rehabilitation.

In summary, our proposal contributes to the ubiquitous health care. Our system improves the physician monitoring, guides patients on the rehabilitation process, and can reduce the problem of health care systems overcharge.

ACKNOWLEDGMENT

This work has been financed by PII1I09-0123-27 and HITO-09-50 projects from Junta de Comunidades de Castilla-La Mancha, and by the TIN2009-14406-C05-03 project from the Ministerio de Ciencia e Innovación (Spain)

REFERENCES

[1] Z. Prekopcsk, "Accelerometer based real-time gesture recognition," Proc. International Student Conference on Electrical Engineering, Prague, Czech Republic, May 2008, pp. 1-5.

[2] J. Doscher and C. Kitchin, "Monitoring machine vibration with micromachined accelerometers," Sensors, vol. 14(5), 1997, pp. 33-38.

[3] M. Joselli and E. Clua, "grmobile: A framework for touch and accelerometer gesture recognition for mobile games," Proc. Brazilian Symposium on Games and Digital Entertainment, Rio de Janeiro, Brazil, Oct. 2009, pp 141-150, doi:10.1109/SBGAMES.2009.24.

[4] K. M. Culhane, M. OConnor, D. Lyons, and G. M. Lyons, "Accelerometers in rehabilitation medicine for older adults," Age and Ageing, vol. 20, Oct. 2005, pp. 556–560, doi:10.1093/ageing/afi192

[5] E. Jovanov, A. Milenkovic, C. Otto, and P. C. de Groen, "A wireless body area network of intelligent motion sensors for computer assisted physical rehabilitation," Journal on Neuroengineering Rehabilitation, vol. 2(6), Mar. 2005, pp 1-10, doi:10.1186/1743-0003-2-6

[6] E. Mpofu and T. Oakland, Rehabilitation and Health Assessment: Applying ICF Guidelines. Springer Publishing Company, 2010, 760 pp.

[7] T. ODonovan, J. ODonoghue, C. Sreenan, D. Sammon, P. OReilly, and K. A. OConnor, "A context aware wireless body area network," Proc. Int. Conf. on Pervasive Computing Technologies for Healthcare, London, UK, Apr. 2009, pp. 1-8, doi: 10.4108/ICST.PERVASIVEHEALTH2009.5987

[8] G. Niezen and G. P. Hancke, "Gesture recognition as ubiquitous input for mobile phones," Proc. Int. workshop on Devices that Alter Perception, Sep. 2008, pp. 25-28.

[9] MovilForum, "Aplicaciones ehealth en la feria movilforum 2009," http://pressoffice.telefonica.com/documentos/Dossier (2010, Feb 20th)

[10] A. S. Alklind Taylor, P. Backlund, H. Engström, M. Johannesson, H. Krasniqi, and M. Lebram. "Acceptance of Entertaiment Systems in Stroke Rehabilitation," Porc. IADIS Game and Entertaiment Technologies, Algarve, Portugal, Jun. 2009, pp. 75-83.

[11] T. S. Leong, J. Lai, J. Panza, P. Pong, and J. Hong, "Wii want to write: An accelerometer based gesture recognition system," Porc. Int. Conf. on Recent and Emerging Advanced Techologies in Engineering, Pan Pacific, Malaysia, Nov. 2009.

[12] R. Niels and L. Vuurpijl, "Using dynamic time warping for intuitive handwriting recognition," Proc. Conference of the international graphonomics society (IGS2005), Salerno, Italy, Jun. 2005, pp. 217–221.

[13] L. A. R. Solano, Verificacion del hablante basado en Dynamic Time Warping. Universidad del Norte, 1998.

[14] T. Schölmer, B. Poppinga, N. Henze, and S. Boll, "Gesture recognition with a wii controller," Proc. 2nd international conference on Tangible and embedded interaction, Bonn, Germany, Feb. 2008, pp 11-14, doi: 10.1145/1347390.1347395

[15] D. Lopez-de-Ipiña, X. Laiseca, A. Barbier, U. Aguilera, A. Almeida, P. Orduña, and J.I. Vazquez. "Infrastructural support for ambient assisted living," Advances in Soft Computing 51, Springer, 2009, pp 66-75.

[16] J. Garduño, L. Pedersen, P. Brodal, T. Sund, G. K. Klungsøyr, T. Konstali, and H. Gundersen. "OMA-BCAST Quiz Gaming Prototype for Mobile". Proc. IADIS Game and Entertaiment Technologies, Algarve, Portugal, Jun. 2009,

[17] M. García-Herranz, P.A. Haya, A. Esquivel, G. Montoro, and X. Alamán. "Easing the smart home: Semi-automatic adaptation in perceptive environments," Journal of Univeral Computer Science vol. 14(9), 2008, pp. 1529-1544.

[18] S. Choquette, M. Hamel, and P Boissy. "Accelerometer-based wireless body area network to estimate intensity of therapy in post-acute rehabilitation," Journal of NeuroEngineering and Rehabilitation vol. 5(20), 2008, doi: 10.1186/1743-0003-5-20

[19] S. Arteaga, J. Chevalier, A. Coile, A. W. Hill, S. Sali, S. Sudhakhrisnan, and S. H. Kurniawan. "Low-cost accelerometry-based posture monitoring system for stroke survivors," Proc. 10th international ACM SIGACCESS conference on Computers and accessibility. Halifax, Canada, 2008, pp. 243-244, doi: 10.1145/1414471.1414519.

[20] M. W. Hinkel. "WiiHab rehabilitative therapy using the Wii." http://wiihabtherapy.blogspot.com . (2008, May 10th)

[21] ST Microelectronics. "LIS302DL Datasheets and characteristics," www.st.com/stonline/books/pdf/ docs/12726.pdf (2010, May 29th)

[22] J. Bravo, D. Lopez de Ipiña, C. Fuentes, R. Hervás, R. Peña, M. Vergara and G. Casero. "Enabling NFC technology for supporting chronic diseases: A proposal for alzheimer caregivers," LNCS 5355. Springer Publishing Company, 2008, pp. 109-125, doi: 10.1007/978-3-540-89617-3_8

[23] R. Hervás, S. W. Nava, G. Chavira, V. Villarreal and J. Bravo. "PIVITa: Taxonomy for displaying information in pervasive and collaborative environments," Advances in Soft Computing 51, Springer Publishing Company, 2009, pp. 293-301, doi: 10.1007/978-3-540-85867-6_34

 ISBN: 978-1-61208-100-7

The Importance of Context Towards Mobile Services Adoption

Shang Gao, John Krogstie
Department of Computer Science and Information Science
Norwegian University of Science and Technology
NO 7491 Trondheim, Norway
shanggao@idi.ntnu.no, krogstie@idi.ntnu.no

***Abstract*—Along with the popularity of mobile devices and advances in wireless technology, mobile services have become more and more prevalent. Although many analysts have predicted that mobile systems will become mainstream, the adoption of mobile services has been slower than expected. The main objective of this research is to study the influence of context on mobile services adoption. The importance of context towards mobile services adoption was explored by looking at two newly developed mobile services. The findings from the exploratory study demonstrate that context is a significant factor to affect people's adoption of mobile services.**

Keywords-Context; Mobile Services Adoption

I. INTRODUCTION

Along with the popularity of mobile devices and advances in wireless technology, mobile services have become more and more prevalent. Although many analysts have predicted that mobile systems will become mainstream [1], the adoption of mobile services has been slower than expected. Despite all the technological possibilities, the number of successful context-aware mobile services in the commercial market is still limited [2]. Building successful strategies for promoting mobile services stems from understanding the context in which potential users prefer to use mobile services. Key factors for the success of mobile services are to identify the actual and potential customers, to investigate how they are influenced and how they behave (i.e., people's behavior) and to uncover what they really expect (i.e., needs, and preference) [3]. Therefore, it is important to study how users' perception on mobile services is affected by context.

The main objective of this research is to study the influence of context on mobile services adoption. Extensive research on the Technology Acceptance Model (TAM) [4][5] has explained why people accept or reject information systems. However, TAM has limitations when investigating users' adoption of mobile services, which is also confirmed by prior research work [6]. An important goal throughout this work is to investigate the importance of context in the adoption of mobile services. By exploring the role of context towards mobile service adoption in two case studies, the findings of this research will not only help mobile services developer to better understand users' expectations on mobile services, but also provide insights into how to promote new mobile services to potential users.

The remainder of this paper is organized as follows. In Section 2, we review prior literature on mobile commerce, mobile services, and context. Section 3 discusses some related work. In Section 4, we illustrate the role of context in mobile services and propose some contextual factors. Section 5 explores the importance of context towards mobile service adoption by looking at two newly developed mobile services. Section 6 concludes this research work and points out directions for future research.

II. LITERATURE REVIEW

A. Mobile Commerce and Mobile Services

Mobile Commerce [7] refers to e-commerce services, conducted through mobile devices using wireless telecommunications networks and other wired e-commerce technologies. Due to its inherent characteristics such as ubiquity, personalization, flexibility, and dissemination, mobile commerce promises business unprecedented market potential, enhanced productivity, and high profitability. Hence, network designers, service providers, vendors and application developers must cautiously take the needs and considerations of various users into account to provide better services and attract them to mobile commerce [8].

Mobile commerce involves mobile services, mobile technologies, and business models. Mobility implies portability. In other words, users can conduct business on real time bases in mobile commerce environment. Customers as well as vendors can be reached at any time via a mobile device. Ubiquity, convenience, localization, and personalization are characteristics of mobile commerce [9].

With the evolution of mobile technologies and the appearance of new innovative business models, we are seeing the growth of mobile services. Over the past 10 years mobile devices have changed the way that we work and live. Many people consider mobile devices as extensions and attachments of themselves [10]. As technology advances, mobile devices are able to be used to do things and fulfill needs in a more efficient and effective manner.

Mobile services provide an entirely new way for services providers to better serve their users through a variety of mobile devices over a wireless network in a wireless environment. Mobile services will enable users to make purchases, request services, as well as access news and information using mobile devices. Some key

 ISBN: 978-1-61208-100-7

features of mobile services are: mobility, reachability, localization, personalization [11].

B. Context

Webster's Dictionary defines context as "whole situation, background or environment relevant to some happening or personality." The definition of context in the Free Online Dictionary of computing is "that which surrounds, and gives meaning to something else." Building on those definition from dictionary, Dey et al, [12] crafted a definition that operationlized the concept in terms of the actors and information sources involved in creating context: "Any information that can be used to characterize the situation of entities (i.e., whether a person, place, or object) that are considered relevant to the interaction between a user and an application, including the user and the application themselves." Context is typically the location, identity, and state of people, groups, and computational and physical objects.

Context is a key issue in the interaction between users and mobile devices, describing the surrounding facts that add meanings. Location can be regarded as one part of the context. In [13], the authors create a working model for context. At the top level of this model, they propose two contexts related to human factors in widest sense and physical environment respectively. Human factors related context is structured into three categories: information on the user (i.e., knowledge of habits), the user's social environment (i.e., co-location of others, social interaction), and the user's tasks (i.e., spontaneous activity). Likewise, context related to physical environment is structured into three categories: location (i.e., absolute position, relative position), infrastructure (i.e., surrounding resources for computation), and physical conditions (i.e., noise, light). Furthermore, how context relates to requirements specification and analysis and design of mobile information system was discussed in [14][15].

III. Related Work

Mobile services adoption is a relatively new field of research. When introducing new information technology, it is critical to study factors that influence user intention to adopt the new services. Developers and vendors can apply this knowledge throughout the design and implementation process to create a better service. Various technology acceptance models and theories, for example, the Technology Acceptance Model (TAM) [4][5], Theory of Planned Behavior (TPB) [16], Innovation Diffusion Theory (IDT) [17], Unified Theory of Acceptance and Use of Technology (UTAUT) [18], have been suggested to assist developers in the evaluation of new software applications. Further, the authors proposed a mobile services acceptance model [19] by extending traditional technology acceptance and diffusion theories above to assess users' adoption to mobile services.

While acceptance and adoption of IT services has been one of the most prevailing IS research topics (e.g., [5], [20], [21]), the pervasiveness of mobile systems raises new questions in exploring the adoption of mobiles services, such as what are the key factors determining the adoption of mobile services, and how contextual factors affect users' adoption of mobile services.

Because of these, some context related theories and frameworks were proposed to address the issue of mobile service adoption. Figge [22] introduced situation dependency as a new concept to adapt mobile services according to spatial, personal, and temporal context in which the user accesses a service. Situation dependency may be conceived as a three dimensional space, with user identity (personal profile, background, preferences, etc.), access position, and access time. In [23], they proposed contextual perceived usefulness as a new construct to enhance the understanding of an individual's mobile commence acceptance behavior.

However, the number of studies using individual consumer samples to investigate and observe the importance of contextual factors towards the adoption of newly developed technologies and services specifically provided on mobile devices and ubiquitous systems is small. Most previous studies on the significance of contextual factors towards mobile services adoption focus on general mobile services like voice, data services and messaging. Therefore, we believe that the contextual determinants for mobile services adoption with some newly developed mobile services is still worthy of examination. In this research work, we examine some contextual factors for mobile services adoption on two mobile services.

Concerning the two mobile services used in our exploratory studies, it might be equal to some existing mobile services on the commercial market. Moreover, the functions in these two mobile services are quite advanced and are able to offer some interesting applications to university students. Considering the fact that university students will become major customers on mobile business market soon, the role of context in these two mobile services is worthy to explore.

IV. The Role of Context in Mobile Services

The term context has been extensively used in the research of mobile related technologies. A unique feature of mobile services is that it can be applied in different contexts. A context often describes the surrounding circumstances of mobile services, which receives increasing attention in mobile computing.

Context provides an understanding of the way and circumstances for performing an activity [24]. Mobile services are often developed to provide an alternative channel for accessing services, not to replace the existing channels completely. The use of mobile services is able to provide time and place independent service access. When a service needs to be accessed immediately regardless of time and place, the usefulness of the mobile service is perceived as the highest, so that it would implicitly influence user's intention to use the service. Because a user's concerns and needs vary with the context in which he/she uses a service, the services that can meet the user's

 ISBN: 978-1-61208-100-7

needs in a specific context will provide the best value to the user [22]. Therefore, we believe that context plays an important role in the adoption of mobile services. Some contextual factors would influence the usage of mobile services.

Based on the context, a user can decide whether the mobile services are useful or not. For example, if people have no access to a desktop computer, they will perceive accessing information systems via mobile devices as useful. Prior research [25] found that there were significant differences between experienced users and inexperienced users in the influence of intention to use. In [26], the authors also indicated that, for experienced users, there was a stronger intention to use the technology/service. It is also believed that users' perception of the ease of use and usefulness of mobile services may vary in different contexts.

The growing interest on the part of practitioners and academics alike in developing context-aware mobile services underlines the importance of context [27]. It is believed that the added value of mobile services depends on the context in which users are using. This inspired us to study the potential contextual factors which may impact mobile services adoption.

Furthermore, the authors in [28] classify the frequent changes in the context with regards to the usage of mobile information systems into six categories. We list three of them which are of relevance to the design of mobile services. Firstly, the environmental (physical) context, which captures the entities that surround the user, for example, the absolute or relative location of the user, plays an important role. Secondly, the task context describes what the user is doing. The task context may refer to the tasks people are interested. This view is also empirically confirmed by [29]. Thirdly, the social context describes the social aspects of the user context. It may, for instance, contain information about friends, neighbors, co-workers, and relatives. The role that the user plays is an important aspect of social context.

People's profile/lifestyle plays an important role in mobile services adoption. The extent to which users are inclined to mobile services adoption in a given situation may vary depending on people's profile/lifestyle. For example, party people may be much more interested to use mobile services in social situations, while as professional individuals may prefer to use mobile services which are able to help them with their daily tasks.

Task oriented context is also important for all groups of mobile users. For instance, students may like to use mobile social services to keep in touch with their friends and use mobile student information systems to keep updates about class related information, while professional individual may aim to use mobile services to ease their daily tasks whenever they are on the move.

Based on the definitions of context provided in Section 2 and context related concepts in this section, we decompose context into two dimensional constructs: people-centered context and place-centered context. Each dimension can then further divide into four categories (see Table 1). People-centered context mainly refers to personal profile (e.g., gender, personal preference, cultural background), past experiences people have (past impression or perception with similar systems), social status and roles, and personal tasks or goals. Place-centered context refers to a specific location, what kinds of resources the place has, what kind of environment people are in (e.g., weather, sound-level), and network condition (e.g., network connectivity).

TABLE I. CONTEXTUAL FACTORS

Context	
People-Centered Context	***Place-Centered Context***
Personal Profile/Lifestyle	Location
Past Experience	Physical Condition
Social Status	Available Resources
Personal Tasks or Goals	Network Condition

V. EXPLORATORY STUDIES

We have carried out two exploratory studies to study users' impressions and perceptions on mobile services in different contexts. The aim of these two exploratory studies is to study the importance of context towards mobile services adoption and test some of the contextual factors proposed in the last section. Two mobile services were presented to the participants in two daily life scenarios at a Norwegian university campus.

A. Study 1- Mobile Students Information Systems(MSIS)

The main purpose of the Mobile Students Information Systems (MSIS) is to offer a number of mobile services that can assist students in their daily activities in a university campus environment. The system makes use of contextual information such as location, time, and personal preferences to provide the user with relevant and timely information. MSIS consists of three parts: a lightweight client application for deployment on mobile devices, a Web-based portal for system configuration, and a backend server which provides database storage, business logic, and a number of public web services.

Three basic functions are offered by the system:

1). Location Finder: Allow users to search for different type of locations on campus, e.g. lecture rooms, computer labs, dining halls, etc. It provides a short description of the location with an option to show the position of the location on a map.

2). Lecture Planner: Allow users to view current lectures for a given day.

3). Announcement: News, notifications, and other information relevant to the user are published on an announcements board. The list supports sorting according to different "flags", such as importance or category.

Figure 1 shows screenshots of the MSIS main menu and the location finder service as they appear on a Windows Mobile 6 Professional emulator. This is quite similar to how it appeared on the actual test devices.

In order to assess the importance of contextual factors on the MSIS system, a survey was conducted to all the

 ISBN: 978-1-61208-100-7

(a) The main menu (b) The location finder

Figure 1. The screenshots of the MSIS

invited participants after using the MSIS in two specific realistic scenarios in the university campus environment for around 45 minutes. The first scenario utilizes the location finder and map services within campus, whereas the second scenario utilizes the course schedule service. Respondents were also informed that the data being collected was part of a research study.

25 university students participated in this study. The students were from various study programs, including students with both technical and non-technical background. Fifteen of the participants were students majoring in computer science, whereas the other 10 participants were students with non-computer science background. Most of the survey participants had at least one mobile device and had some previous experience with mobile services.

It is believed that the adoption of mobile services is likely to be more affected by context than traditional desktop applications. As expected, our findings show that students are more likely to use the system if they are in a situation where they do not have access to a desktop computer or a laptop. According to the survey results, all the participants would use the system if they were out of their office or home. Both of the situations above can be considered as place-centered context. The first situation is related to the available resources in place-centered context, while the second situation is related to the location in place-centered context.

Another interesting observation was made from another contextual related measurement item, which concerned the users' previous experience with mobile services. 36% of the respondents did not regard this to be a critical factor. 12% were neutral to this matter, while 52% agreed that they would more likely use the system if they previously had had a nice experience with mobile services. This demonstrates the importance of people-centered context (i.e., past experience) on mobile services adoption.

Further, most respondent (24 out of 25 participants) indicated that they would also more likely use the service if it would be meaningful in the current situation and help increase task efficiency. This finding proves the significance of fulfilling personal goals, which is a people-centered contextual factor, on the adoption of mobile services.

36% of the participants agreed that they would use the MSIS system if most people around them are using the system. 32% were neutral to this determinant, while another 32% disagreed with this. According to our survey result, this people-centered contextual factor is the least important contextual determinant for mobile service adoption. This shows that, given the fact that the service has a value for them, the users are generally not affected by others' decisions to use a mobile service or not.

B. Study 2- FindmyFriends

FindMyFriends was a project developed by Accenture for UKA-07 (a student festival in Trondheim, Norway) that allowed students to locate each other at Samfundet, the building where the student society is located. Samfundet was constructed in 1929, has three and a half floors, and contains 10 main rooms. The most prominent arrangement at Samfundet is the biennial student festival UKA.

It is a known problem among students in Trondheim that it is difficult to find each other inside Samfundet. The system was particularly aimed towards the more than 2000 voluntarily workers of UKA-07, to make it easier for the workers to keep track of each other.

In brief, the FindMyFriends system offered the possibility of keeping track of your friends in the main venue of the festival. Just before UKA-07 started, the users received their tag used for positioning. In order to connect to other users, the user needed to link the tag with his/her profile, and registers the tag at the FindMyFriends system. Then, users could start connecting to each other,

 ISBN: 978-1-61208-100-7

much like Facebook or any other social network service. When a user moved around Samfundet wearing the tag, the user's friends could log on the FindMyFriends or one of the terminals to check out the user's position. A user can only locate the users that have accepted to be friends with him/her. In addition to the FindMyFriends system, there are some terminals placed at Samfundet, which allowed the users to log into the system and keep track of their friends. Moreover, the system could generate statistics based on the user profiles, which allowed the users to see which rooms that had most girls, the average age of the users in a room, where you should be if you want to meet most single boys and similar statistics.

The technology used for positioning of tags inside Samfundet is ultrasound indoor positioning system (IPS). Ultrasound makes it possible to locate users precisely by room using wireless detectors. Each tag has its own unique identification sound, which is transmitted periodically or by moving. This sound is detected by one of the 63 detectors ("microphones") spread around in the rooms of Samfundet.

As reported in [30], we did a study to investigate the usability of this system. A questionnaire was distributed to the registered users at the FindMyFriends system after the student festival and face-to-face interviews with some respondents were carried out as well. Here we only present the observations and results that are of relevance to this paper. More specifically, we did some follow-up studies in connection to one of the research questions in the questionnaire: RQ1. Are people willing to use a system with functionality for locating and interacting with their friends and family using a mobile device connected to a wireless network in a city environment?

There were 2769 users registered in the FindMyFriends system, but only 1661 registered tags. 207 users answered the questionnaire. Over one third had between 10 and 29 friends, and approximately 60 % had 10 or more friends. This number corresponds well with the overall distribution of friends for all users. 55% of the participants indicated they would use this kind of system if it was available in the city environment.

The results show that the more a user visited Samfundet, the more friends the user has. This can indicate that the users who did not use the system so much, actually never got the chance to use it, because they visited Samfundet none or only a few times during UKA. This situation is related to the location and available resources aspects of place-centered context. This finding indicate that the users are inclined to keep using the system once they get chance to know and use the system.

Some users expressed great enthusiasm about using the system, and many of them would probably use the system without thinking too much about privacy mechanisms. Most users, who indicated the statement above, thought that it would be useful tool for finding their friends, especially when they were out partying. And six respondents explicitly mentioned that they only would use the system when they were out partying, which is also confirmed by the interviewees of this study. It is believed that this is of relevance to the place-centered context in terms of partying.

Some respondents mentioned that they had no joy from using FindMyFriends and they did not think that this service would give them any value individually. However, one of the interviewee indicated that he may attempt to use the application if some of their closest friends or family is starting to use it. We believe that this impression is of relevance to people-centered context (i.e., personal profile, personal task).

VI. CONCLUSION AND FUTURE WORK

This study presents the results from an exploratory study of the importance of contextual factors on mobile services adoptions in two newly developed mobile services. The theoretical background for the proposed contextual factors was adopted from the existing theories on context.

The findings of our study provide some contributions to mobile services adoption research. First, the study proposed some contextual factors which might influence people's adoption of mobile services based on existing research work on context. Second, the observations obtained from two studies provide support for the fact that context is a significant factor to affect people's impression and perception on mobile services. People tend to use mobile services in the situations, such as, when the services need to be accessed immediately and when other more advanced and convenient alternatives are not available. In these situations the usefulness of mobile services and benefits of mobility are the highest. Third, the results also imply that the general research model on mobile services adoption and diffusion needs to be augmented with contextual related factors which affect the use of mobile services. In the MSIS study, most respondents perceived the MSIS service as useful when it allows them to access lecture information and location of the classroom in a timely manner on the move, particularly in the case that mobile devices as the only possible means to access information. Last but not least, it is believed that the proposed contextual factors in Section 4 would be useful as a foundation to create contextual related instrument items to assess people's adoption of mobile services. It can also offer some insights to compose contextual related questions to test the usability of mobile services.

While our study provided some interesting findings on the importance of context towards mobile services adoption, we are also aware of some limitations of this research work. The respondents in two explorative studies were students at a university. This means that the results do not represent views from other users. Therefore, the generalizability of the results to other potential users remains to be determined. Further, the current study only examined two mobile services in university based environment. More research is needed to test the importance of context in some other commercial mobile services.

 ISBN: 978-1-61208-100-7

There exist some opportunities for future research. First, generalization can be increased by expanding the study to include individuals representing different countries and cultures. Second, we have improved the instrument developed in [31] by taking these contextual factors into consideration. Then, we will try to use the enhanced instrument to measure the importance of the contextual factors towards mobile services adoption in some other context-aware mobile services.

REFERENCES

[1] S. Balasubraman, *et al.*, "Exploring the Implications of M-Commerce for Markets and Marketing," *Journal of the Academy of Marketing Science*, vol. 30, pp. 348-361, 2002.

[2] C. Carlsson, "ECRA - Special issue on mobile technology and services," *Electronic Commerce Research and Applications*, vol. 5, pp. 189-191, 2006.

[3] S. J. Barnes, "The mobile commerce value chain: analysis and future developments," *International Journal of Information Management*, vol. 22, pp. 91-108, 2002.

[4] F. D. Davis, "Perceived usefulness, perceived ease of use and user acceptance of information technology," *MIS Quarterly*, vol. 13, pp. 319-340, 1989.

[5] F. D. Davis, *et al.*, "User acceptance of computer technology: a comparison of two theoretical models," *Manage. Sci.*, vol. 35, pp. 982-1003, 1989.

[6] J.-H. Wu and S.-C. Wang, "What drives mobile commerce? An empirical evaluation of the revised technology acceptance model," *Inf. Manage.*, vol. 42, pp. 719-729, 2005.

[7] K. Siau, *et al.*, "Mobile Commerce – Promises, Challenges, and. Research Agenda," *Journal of Database Management*, vol. 12, pp. 4-13, 2001.

[8] P. Pedersen and L. Methlie, "Understanding Mobile Commerce End-User Adoption: A Triangulation Perspective and Suggestion for an Exploratory Service Evaluation Framework," in *Proceedings of the HICSS'02*, Hawaii, USA, 2002.

[9] G. S. Mort and J. Drennan, "Marketing m-services: Establishing a usage benefit typology related to mobile user characteristics," *The Journal of Database Marketing & Customer Strategy Management*, vol. 12, pp. 327-341, 2005.

[10] K. Wehmeyer, "Assessing Users' Attachment to Their Mobile Devices," in *Proceedings of the International Conference on Mobile Business (ICMB 2007)*, 2007.

[11] K. Siau and Z. Shen, "Mobile communications and mobile services," *Int. J. Mob. Commun.*, vol. 1, pp. 3-14, 2003.

[12] A. K. Dey, "Understanding and Using Context," *Personal Ubiquitous Comput.*, vol. 5, pp. 4-7, 2001.

[13] A. Schmidt, *et al.*, "There is more to context than location," *Computers and Graphics*, vol. 23, pp. 893-901, 1999.

[14] J. Krogstie, "Requirements Engineering for Mobile Information Systems," in *the Seventh International Workshop on Requirements Engineering: Foundations for Software Quality (REFSQ'01)*, Interlaken, Switzerland, 2001.

[15] J. Krogstie, *et al.*, "Mobile Information Systems - Research Challenges on the Conceptual and Logical Level," in *Proceedings of the MobiMod'02*, Tampere, Finland, 2002.

[16] I. Ajzen, "The theory of planned behavior," *Organizational Behavior and Human Decision Processes*, vol. 50, pp. 179-211, 1991.

[17] E. M. Rogers, *The diffusion of innovations*. New York: Free Press, 1995.

[18] V. Venkatesh, *et al.*, "User Acceptance of Information Technology: Toward a Unified View," *MIS Quarterly*, vol. 27, pp. 425-478, 2003.

[19] S. Gao, *et al.*, "Mobile Services Acceptance Model," in *Proceedings of the 2008 International Conference on Convergence and Hybrid Information Technology*, 2008.

[20] S. Taylor and P. A. Todd, "Understanding Information Technology Usage: A Test of Competing Models," *Information Systems Research*, vol. 6, pp. 144-176, 1995.

[21] K. Ven and J. Verelst, "The Impact of Ideology on the Organizational Adoption of Open Source Software," *Journal of Database Management*, vol. 19, pp. 58-72, 2008.

[22] S. Figge, "Situation-dependent services--a challenge for mobile network operators," *Journal of Business Research*, vol. 57, pp. 1416-1422, 2004.

[23] T. Lee and J. Jun, "Contextual Perceived Usefulness? Toward an Understanding of Mobile Commerce Acceptance," in *Proceedings of the International Conference on Mobile Business*, 2005.

[24] R. C. Basole, "The value and impact of mobile information and communication technologies," in *IFAC Symposium on Analysis, Modelling & Evaluation of Human-Machine Systems*, Atlanta GA, USA, 2004.

[25] I. Ajzen and M. Fishbein, *Understanding Attitudes and Predicting Social Behavior*. Englewood Cliffs, NJ: Prentice-Hall, 1980.

[26] S. Taylor and P. Todd, "Assessing IT usage: the role of prior experience," *MIS Q.*, vol. 19, pp. 561-570, 1995.

[27] M. de Reuver and T. Haaker, "Designing viable business models for context-aware mobile services," *Telematics and Informatics*, vol. 26, pp. 240-248, 2009.

[28] J. Krogstie, *et al.*, "Research areas and challenges for mobile information systems," *Int. J. Mob. Commun.*, vol. 2, pp. 220-234, 2004.

[29] H. Bouwman and L. van de Wijngaert, "Coppers context, and conjoints: a reassessment of TAM," *Journal of Information Technology*, vol. 24, pp. 186-201, 2009.

[30] A. Kofod-Petersen, *et al.*, "An empirical investigation of attitude towards location-aware social network service," *Int. J. Mob. Commun.*, vol. 8, pp. 53-70, 2010.

[31] S. Gao and J. Krogstie, "Development of an Instrument to Measure the Adoption of Mobile Services," in *8th Global Mobility Roundtable conference (GMR 2009)* Cairo, Egypt, 2009.

 ISBN: 978-1-61208-100-7

Human Behaviour Detection Using GSM Location Patterns and Bluetooth Proximity Data

Muhammad Awais Azam, Laurissa Tokarchuk, Muhammad Adeel
Department of Computer Science and Electronic Engineering,
Queen Mary University of London,
UK, E1 4NS
{muhammad.azam, laurissa.tokarchuk, muhammad.adeel}@elec.qmul.ac.uk

Abstract— **Human behaviours are multifarious in nature and it is a challenging task to predict and learn from daily life activities. The profusion of Bluetooth enabled devices used in daily life has created new ways to analyze and model the behaviour of individuals. Bluetooth integrated into mobile handsets can be used as an efficient short range sensor. The aim of this research work is the detection of unusual human behaviours from cell tower and Bluetooth proximity data using neural networks. The primary purpose is to find anomalies in individual's daily life routines that will further help us to detect and predict unusual behaviour of elderly people and patients such as dementia patients.**

Keywords- Behaviour; Cell tower ID; Bluetooth proximity; Neural Network; Jaccard Index

I. INTRODUCTION

Detection and prediction of human behaviour is a hot issue nowadays in social research circles. Modelling human behaviour such as individual routines from proximity data and relations with gathered data of daily life activity patterns is an emerging realm in ubiquitous computing. There can be different sensing devices e.g., Radio Frequency Identification (RFID), motion sensors, GPS enabled tracking devices [5], and other context aware devices that can be used for real time proximity detection and daily life data gathering purposes. In particular, devices such as mobile phones provide a rich platform for various forms of data gathering by using its integrated sensors such as Bluetooth ID, digital camera, microphones and GPS transceivers. These sensors can give an individual's location, movement and proximity information for the whole period of cell phone usage. Specifically, Bluetooth radios are frequently incorporated into mobile devices [3].

Human behaviours and activities are analyzed by researchers using different sensor devices such as accelerometers, digital cameras and microphones. Frameworks have been presented to identify close proximity social behaviours [16], group actions in meetings [17] and audio visual perception of a lecture in smart environment [18]. In most studies, the majority of the sensing devices that are used in limited environment restrict their usage and thus it is only useful when activities take place in their proximity. This does not suit our case as we are interested not only indoor and limited environments but also outdoor movement and activities.

The enormous penetration of Bluetooth devices have enabled them to be used as a personal identifier. Many researchers have exploited this capability by using the mobile as a sensing device. The mobile phone nowadays is an indispensable part of our society with many integrated sensors. Researchers have investigated using these sensors in social proximity sensing [8][19], social behavioural modelling and routine classification [1][2][3][4] and movement prediction [6][7]. The significance of these studies is that they have identified new techniques to recognize an individual's behavioural patterns and abnormal movements. In [1][2], Author Topic Model (ATM) and hierarchical Bayesian topic models like Latent Dirichlet Analysis (LDA) are used for routine classification. A framework for daily life activity recognition based on the user's location and group affiliation is then presented. In [6][7], neural networks are used to detect and predict user movement based only on cell tower IDs. Our work is similar in one aspect with their work and that is; we have also utilized the probabilities of user being in different locations. Difference between our work and the work presented in [6][7] is that we have used real time data for our experiments and used both cell tower ID and Bluetooth proximity data.

This work is an extension of [9], in which repeated patterns and behaviour of an individual was detected by using n-gram technique and considering only Bluetooth proximity data. The primary purpose of this research is to detect unusual daily life activity patterns and individual behaviours that deviate from their normal routines in order to aid in the detection of abnormal behaviour such as wandering behaviour, a behavioural disorder in dementia patients. Another aim is to determine the reliability of behaviour detection using only Bluetooth proximity data. So, behaviour detection is done by considering the record of cell tower ID's and Bluetooth proximities because of the easy and economical availability of

 ISBN: 978-1-61208-100-7

Bluetooth enabled devices such as mobile phone as sensing device. This detection of Bluetooth Proximate devices shows the regularity of user's behaviour as discussed in [4]. According to [4], if the user in his daily life, repeat the activities and routines with less change, it will be known as 'low entropy' behaviour. While a more change in daily routine patterns is considered as 'high entropy' behaviour. Now, if we consider the elderly people and patients specifically dementia patients, they have somewhat fixed and regular routines to follow [20] that make those individuals a low entropic user based on [4]. Therefore our interest is to study primarily the users with low entropy from the reality mining dataset [4].

The results presented here use the reality mining dataset collected at MIT for the year 2004-2005. Nokia 6600 cell phones were used to record the data of 100 users over duration of 9 months. This research uses cell tower ID and Bluetooth proximity data to analyze the routines and behaviour of an individual that deviate from their normal routines. This paper presents the techniques used and corresponding analyses of the data that show the level of behavioural abnormality of individual's routine by using cell tower IDs and Bluetooth proximity information.

The rest of the paper is as follows: Section-II contains related work on abnormal activity detection and usage of Bluetooth as a sensing device. Section-III explains the methodology of our analysis that we have adopted to get the results. Section-IV discusses the results and Section-V contains the summary of the work and notes on the direction planned for our future work.

II. STATE OF THE ART

Detection of abnormality in human behaviour is very intricate and challenging task for researchers and has been in the past. Recently, with improvements in network systems and information technology, people have more easily been able to study the behaviours and activities of humans. Researchers have tried to detect the abnormal routines and daily life patterns of an individual inside the home and restricted environments [10][13]. Majority of work in this area has used sensing devices that either have short range of detection, less battery power and storage, or not very common that every person can use it without adding extra hardware, which is not possible for the scenarios outside the home. In [10], researchers have presented a framework for the detection of unusual human behaviour inside an intelligent house that is different from our case as we are considering the scenarios outside the home as well. They used motion sensors to detect the activities and unusual patterns based on Markov Chain. Vector quantization is used to reduce the sensor states and the change between these states is observed by transition probability. They detect the unusual behaviour by computing the distance between the state transition probabilities or by the likelihood of user action. The distance between the state transition probabilities was calculated by using either Kullback-Leiber distance or Euclid distance.

In [13], researchers detect abnormal event in solitary elder's daily life by mining the related data gained by sensors. They employ the association rules finding algorithm with time cluster to analyze the elder's activities. In first step, they cluster each item of elder activity with time and then in the second step, all frequent item sets were found and strong association rules were created. Researchers in [15] work on the recognition of abnormal activities based on the Hierarchical Dirichlet Process Hidden Markov Model (HDP-HMM). They incorporate a Fisher Kernel into the One-Class Support Vector Machine (OCSVM) to filter out the most likely normal activities. Then from those normal activities, they derive a model to detect abnormal activities and tried to reduce false positives. In [14], a model for abnormal behaviour detection is presented. That model considers user's location based on the cell tower ID and used Dynamic Bayesian Networks (DBN) to predict user's location. They proposed an X-Factor model, which is a DBN with a hidden variable. User's location according to this model not only depends on the hour of the day and day of the week but also this latent variable that represents the abnormal behaviour. In our work, we have applied neural network to detect abnormal behaviour of an individual by using cell tower ID and Bluetooth proximity data. This detection of behaviour will help us to aid elderly people and dementia patients.

III. METHODOLOGY

The aim of this research is the detection of abnormal behaviour in an individual's daily routines in order to aid in the detection of unusual behaviours in patients such as dementia patients, by using cell tower ID data and Bluetooth proximity data. This section starts by evaluating the use of only cell tower ID data and describes the methodology used in this research. Same methodology is then applied on the Bluetooth proximity data with some changes that are discussed in more detail in results section.

Cell tower ID gives information about the user's location and movement. The cell tower ID data that is used in this study is classified into four different locations; i.e., Home (H), Work (W), Elsewhere (E) and NoSignal (N). This data is divided into twenty four time slots. Each time slot is represented by the associated presence information of the user (H, W, E, and N) during the one hour period. The presence of user at specific location depends on the hour of the day and day of the week. For example, if the user has a regular routine of going to office, then location of the user at 10a.m on Saturday morning can not be the same at 10a.m on Monday morning. The daily life activities of an individual depend on the entropy level of the user as discussed in [4]. If the user is a low entropy user, his routines do not change much as compared to high entropy users, whose routines and activity

 ISBN: 978-1-61208-100-7

patterns change continuously. Figure 1 show the basic architecture used to get the behaviour of an individual. Data Base (DB) contains the classified information of cell tower ID data into H, W, E and N.

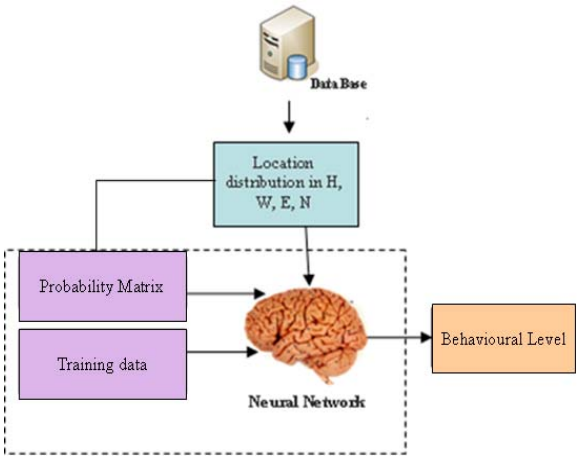


Figure 1. Data Processing Design

A probability matrix is generated depending on the hour of the day and day of the week from this classified information. This means that every entry of this matrix depends upon the specific hour of the day and whether it is a weekday or a weekend. This matrix is utilised for preparing training data for the neural network that is used for the detection of level of abnormality in the user behaviour. The neural network used for this purpose is Multilayer Perceptrons (MLP). There are four inputs and one output of this MLP. Output is the behaviour of the user and the format of the input used for this neural network is:

$$[Loc_i, Hour_j, Day_k, Abn\_Level_m]$$

where Loc_i, gives the location i.e., H, W, E, N. $Hour_j$, gives the hour of the day i.e., between 1 and 24. Day_k, gives the day of the week, i.e., between 1 and 7. Abn_Level_m, gives the behavioural levels. Four different levels are assigned to behaviour depending upon the probability of being at any of the four places on a specific day and hour, shown in Table 1.

This generates twenty four samples of training data for one day. So for each user, total training samples are (24 x number of days). 70% of this training data is used for training the neural network whilst the remaining is used for cross validation and testing purposes. Training of the neural network is done till the cross validation error becomes less than 0.02, by using Mini-Batch training process [11], an advantage of using Mini-Batch training is that it is a compromise between batch and incremental training. Output of this neural network will give the level of abnormality of an individual for each hour of the day.

TABLE 1

Probability	Behaviour
$0 < p < 0.25$	Abnormal
$0.25 < p < 0.5$	Low Abnormal
$0.5 < p < 0.75$	Average Normal
$0.75 < p < 1$	Normal

IV. RESULTS

The results discussed here are of one user with the entropy level 23.06, calculated by using the Shannon's entropy equation given below.

$$H(x) = -\sum_{i=1}^{n} p(i) \log_2 p(i)$$

One month data is used to detect the behavioural levels of the user after training the neural network on about 70% of the data available for this specific user. First, only the cell tower ID data is used to detect the behaviour of the user. Figure 2 shows the daily distributions of (H, W, E and N) transitions based on cell tower ID data of one month that is further used to detect the behaviour of the user through neural networks.

Figure 3 shows the comparison of behaviour of an individual for two days. The trained neural network provides the behavioural levels for twenty four hours. As the entropy level of the user is quite low, this figure shows that most of the time the behaviour of the user is average normal. Now look at day-10 in Figure 2, there is an unusual detection of 'Elsewhere' during 5-6am in the morning, which doesn't happen normally in usual daily routine of the user. Figure 3 shows the detection of that unusual behaviour for day-10 in that specific time duration.

Figure 4 and Figure 5 shows the behaviour of the user for one month time duration. In week-1, the behaviour of the user remains average normal and this can be verified from Figure 2 that shows the regularity in the distributions of 'Home' and 'Work' patterns and shows that user did not make any unusual movements.

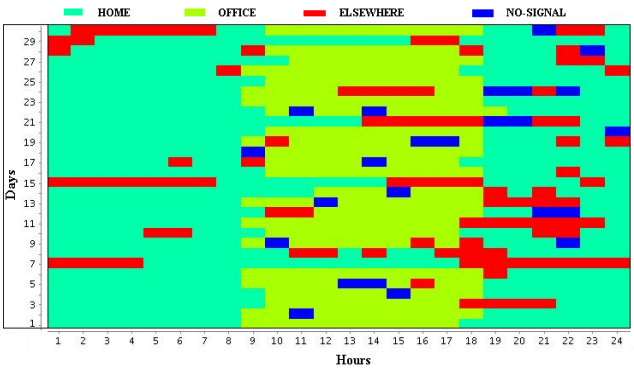


Figure 2. Distribution of (H, W, E and N) Transitions of Cell Tower ID Data

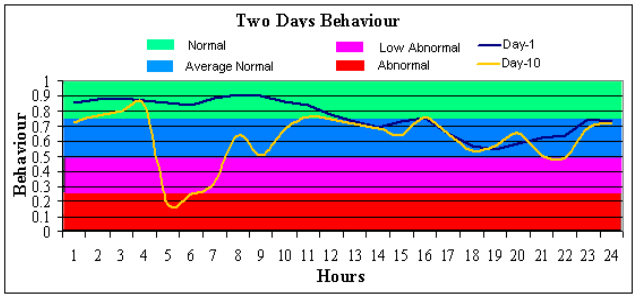


Figure 3. Comparison of Two Days of Behaviour Detected from Cell Tower ID Data

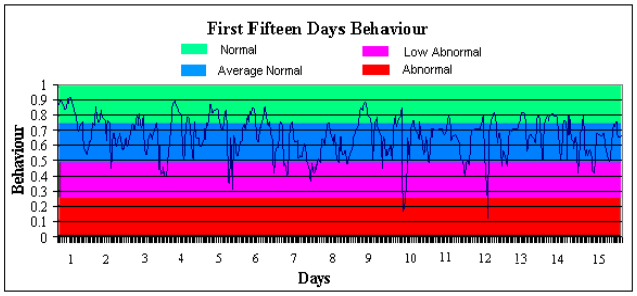


Figure 4. Week-1 and Week-2 Behaviour Detected from Cell Tower ID Data

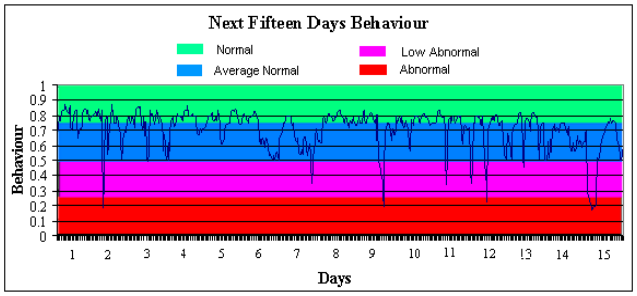


Figure 5. Week-3 and Week-4 Behaviour Detected from Cell Tower ID Data

In week-2, there is a change in behaviour on third and fifth day of the week when the user's (H, W, E and N) distributions show an irregular routine activity. In week-3 and week-4, some unusual routines are also detected in the behaviour of the user. The results presented above show only the cell tower ID data. We will now discuss the results obtained by using the same technique on only the Bluetooth proximity data of the same user. Each time slot for Bluetooth proximity data represents one hour as in the case of cell tower ID data.

Bluetooth proximity data is available in the form of detected devices as a result of a scanning performed by the user's cell phone after every five minutes. Each scanning results a list of devices present within the range of 5-10m. The first aim was to cluster the data in 'Home', 'Office' and 'Other Devices', so that the above technique can be used with Bluetooth data. The Bluetooth proximity data is clustered into only three categories so that the results obtained from cell tower ID data and the Bluetooth proximity data can be combine together to see if we could get some interesting anomalies in behaviour of the user. In future work, the aim is to cluster the Bluetooth proximity data into more finer grained time periods and try to detect the anomaly in user's behaviour in smaller time slots. The reason behind the clustering of Bluetooth data on finer scale is to classify the user behaviour in different activities that will provide one step further in the identification of unusual routines without using cell tower data.

After analysing the data, user's home computer device was given the name 'Home' (H). That means that all those time slots in which user detects his home computer device, considered as 'H' because it shows user's presence in the home. For office, there are many devices that user detect during office hours. To obtain a cluster of devices that belong to office, we remove the weekends from one month data and use Jaccard index [12], to detect how similar the detected devices are throughout the office hours for all remaining weekdays. Jaccard similarity equation is:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

where, 'A' and 'B' are sets of detected devices in two consecutive days. First, 'A' and 'B' represent day-1 and day-2, then day-2 and day-3 and so on up to the all remaining weekdays that left after removing the weekends from one month Bluetooth data. This gives us the similarity of detected devices during the office hours between the pairs of consecutive days, shown below in Figure 6. The average similarity between the detected devices is above 0.5. This means there are many devices that user detect repeatedly during his office hours. All those devices that user detect for at least 70% of the days during office hours goes in 'Office' cluster. All other devices go in 'Other Devices' cluster. After classifying the devices, a new data matrix is generated that contains twenty four time slots for each day as were in the case of cell tower ID data. Each time slot is assigned one of these clusters (i.e., Home, Office, Other Devices, No Devices Found) depending upon the number of detection of the devices belonging to a specific cluster. The same technique as used on cell tower IDs, described previously, is also used with Bluetooth proximity data. Figure 7 shows the Home/Work distribution of locations depending on the Bluetooth data clusters while Figure 8 shows the fifteen days behaviour of the user detected from both cell tower ID and Bluetooth proximity data. An interesting observation can be made by analysing the results of both cell tower ID and Bluetooth ID data. It is observed that sometimes when behaviour detected from cell tower ID data is not unusual, a change in behaviour is detected from Bluetooth proximity data. It can be said that it is more likely to be detecting unusual behaviour because during a regular routine of office hours of a weekday, user is supposed to detect 'Office Devices'. Cell tower ID data will show his normal behaviour as the user is in Office, but may be there is some gathering or meeting of students or staff that is not part of the regular routine. Behaviour detected from Bluetooth proximity data can be pointing towards that activity.

 ISBN: 978-1-61208-100-7

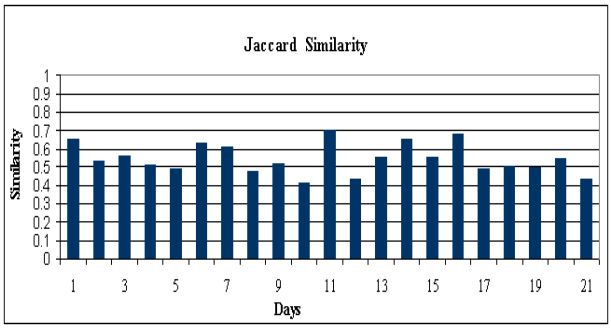


Figure 6. Jaccard Vertex Similarity

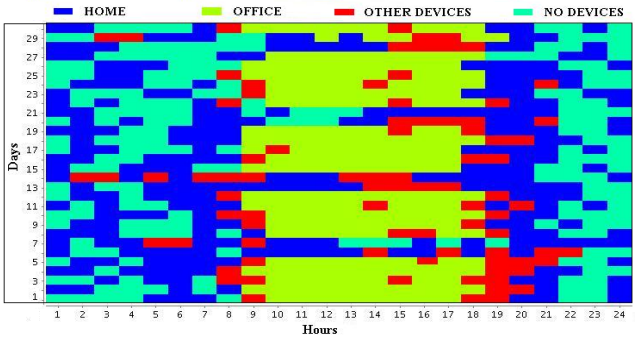


Figure 7. Distribution of Home/Work Transitions of Bluetooth Data

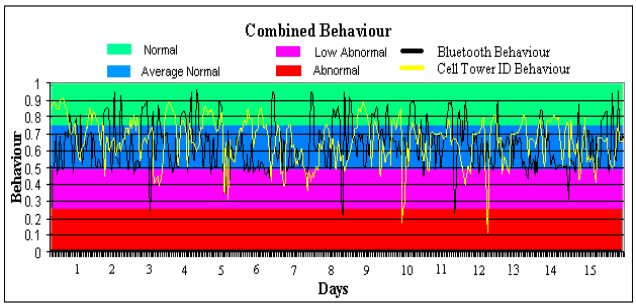


Figure 8. Fifteen Days Behaviour of Both Bluetooth and Cell Tower ID Data

These results show that for low entropy users, the detection of unusual routines and behaviours by using only Bluetooth data is possible. These low entropy users follow specific routines as compared to high entropy individuals, who live more diverse lives. This study aims to aid elderly people and patients to detect abnormal and unusual behaviours to avoid any accidents. Normally patients and elderly people have fixed and limited routines to follow that can likely be detected using Bluetooth devices by clustering the Bluetooth device detections into different activities or communities.

V. SUMMARY AND FUTURE WORK

In this paper, real time Bluetooth proximity and cell tower ID data is used to detect abnormal and unusual activities and routines of an individual by using neural networks. A low entropy user was selected for experiments due to the regularity and constancy in his routines. A successful detection of abnormal behaviour in this user's routines is done by using cell tower ID's and Bluetooth proximity data. Bluetooth proximity data is only clustered into three different categories. The idea was to combine the results of behaviour detection from Bluetooth proximity data with the results of cell tower ID data. To detect anomalies in more specific and lower level activities and routines, we need to cluster the Bluetooth proximity data into temporal clusters. In future work, we will try to cluster the Bluetooth proximity data on temporal scale to cover the minute details of the user's behaviour and will also try to predict the behaviour based on these clusters and communities detection. This will help us to facilitate elderly people and patients who need more care and concern about their behaviour and unusual routines that can cause serious accidents.

REFERENCES

[1] K. Farrahi and D. Gatica-Perez, "Daily Routine Classification from Mobile Phone Data". In: Popescu-Belis, A., Stiefelhagen, R. (eds.) MLMI 2008. LNCS, vol. 5237, pp. 173–184. Springer, Heidelberg (2008)

[2] K. Farrahi and D. Gatica-Perez, "What did you do today? Discovering daily routines from Large-Scale Mobile Data". In: MM 2008: Proceeding of the 16th ACM international conference on Multimedia, pp. 849–852. ACM, New York (2008)

[3] M. Hermersdorf, H Nyholm, J Perkiö, and V Tuulos. "Sensing in Rich Bluetooth Environments"- Workshop on WorldSensorWeb, in Proc. SenSys, 2006 - sensorplanet.org

[4] N. Eagle and A. Pentland, "Reality mining: sensing complex social systems". Personal and Ubiquitous Computing 2006 – Springer, Vol. 10, # 4, 255-268

[5] GPS enabled tracking devices, retrieved on 14-07-2010, http://www.elderoptionsoftexas.com/article_alzheimers_gps_tracking_devices.htm

[6] M. Vukovic, G. Vujnovic, and D. Grubisic, "Adaptive User Movement Prediction for Advanced Location-aware Services", Proceedings of the 17th international conference on Software, Telecommunications and Computer Networks, pp. 343-347 (2009)

[7] M. Vukovic, I. Lovrek, and D. Jevtic. "Predicting user movement for advanced location-aware services". In 15th International Conference on Software, Telecommunications and Computer Networks, pages 1–5. SoftCOM 2007., 2007.

[8] N. Eagle, "Machine Perception and Learning of Complex Social Systems", Ph.D. Thesis, Program in Media Arts and Sciences, MIT, June 2005

[9] M. A. Azam and T. Laurissa (2009). Behaviour Detection Using Bluetooth Proximity Data. Proceedings of Networking & Electronic Commerce Research Conference pp. 46-52 (NAEC 2009).

[10] K. Hara, T. Omori, and R. Ueno, "Detection of unusual human behaviour in intelligent house"; Proceedings of the 2002 12th IEEE workshop on Neural Networks for Signal Processing, pp. 697-706, 2002.

 ISBN: 978-1-61208-100-7

[11] Linear Neural Networks, retrieved on 14-07-2010, retrieved from http://www.idsia.ch/NNcourse/linear2.html

[12] Jaccard Index, retrieved on 14-07-2010, retrieved from http://www.statemaster.com/encyclopedia/Jaccard-index

[13] T. Yiping, Z. Zhiying, G. Hui, L. Huiqiang, W. Wei, and X. Gang, "Elder Abnormal Activity Detection by Data Mining", SICE Annual Conference in Sapporo, August 4-6, 2004, vol. 1, pp. 837–840 (2004) Japan

[14] N. Eagle, A. Clauset, and J. A. Quinn, "Location Segmentaion, Inference and Prediction for Anticipatory Computing", AAAI Spring Symposium, 2009

[15] D. H. Hu, X. Zhang, J. Yin, V.W. Zheng, and Q. Yang, "Abnormal Activity Recognition Based on HDP-HMM Models", AAAI Publications, 21st International Conference on Artificial Intelligence, pp. 1715–1720, 2009

[16] C. Wren, Y. Ivanov, I. Kaur, D. Leigh, and J. Westhues, "SocialMotion: Measuring the Hidden Social Life of a Building". In: J. Hightower, B. Schiele, and T. Strang, (eds.) LoCA 2007. LNCS, vol. 4718, pp. 85–102. Springer, Heidelberg (2007)

[17] I. McCowan,, D. Gatica-Perez, S. Bengio, and G. Lathoud,.: Automatic Analysis of Multimodal Group Actions in Meetings. IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI) 27(3), 305–317 (2005)

[18] R. Stiefelhagen, K. Bernardin, H.K Ekenel, J. McDonough,, K. Nickel, M. Voit, and M. Woelfel: Audio-Visual Perception of a Lecturer in a Smart Seminar Room. In: Signal Processing - Special Issue on Multimodal Interfaces, vol. 86 (12). Elsevier, Amsterdam (2006)

[19] T. Nicolai, N. Behrens, and E. Yoneki, "Wireless Rope: Experiment in social proximity sensing with Bluetooth". In *fourth annual IEEE International Conference on Pervasive Computing, LNCS 4277, 2006*

[20] T. M. Gill, M.M. Desal, A. Evelyne, T. R. Holford, and C. S. Williams, "Restricted-Activity among Community-Living Older Persons: Incidence, Participants, and Health Care Utilization", Annals of Internal Medicine. http://www.annals.org/content/135/5/313.full.pdf+html

 ISBN: 978-1-61208-100-7

Towards Radio Localisation of Running Athletes

Lawrence Cheng[1], Gregor Kuntze[2], Huiling Tan[3], Stephen Hailes[1], David G. Kerwin[2], A. Wilson[3]

[1]Computer Science, University College London, Malet Place, London, UK, WC1E 6BT. {l.cheng, s.hailes}@cs.ucl.ac.uk

[2]Cardiff School of Sport, University of Wales Institute, Cardiff, Cyncoed Road, Cardiff, UK, CF23 6XD, {gkuntze, dkerwin}@uwic.ac.uk

[3]Royal Veterinary College, Structure and Motion Lab, Hawkshead Lane, Herts, UK, AL9 7TA. {htan, awilson}@rvc.ac.uk

***Abstract*—The use of ubiquitous computing for sport performance monitoring was demonstrated in recent work. In this paper, a custom-designed radio-based localisation system and its applicability for tracking a running athlete were presented. The system uses 2.4GHz radio with a Time-of-Arrival (ToA)-based localisation protocol to locate a running athlete on an indoor running track. Through a real-time analysis on the raw localisation data, the location results could be used to support other useful coaching support applications, such as automatic video tracking. The system was experimented against gold-standard technologies. The results show that the presented system achieves a positional accuracy within 21.959cm for tracking running athletes.**

***Keywords*—Application; athletes; localisation; sports; ubiquitous sensing.**

I. INTRODUCTION

The use of ubiquitous sensing to support sports performance monitoring has been demonstrated in recent years. The system design principles for developing ubiquitous sports equipment were identified in [6]; whereas in [7], a wearable system that detects kicks in martial arts was reported. In [1][8][11], the use of pervasive sensing for performance monitoring of athletes (i.e., sprinters) was reported. The most interesting performance information of an athlete is speed, which is derivable from location data. Although the use of radio-based localisation systems to locate objects have been reported, however, most existing work focus on locating static objects (e.g., locating static sensor nodes in an office) or tracking (relatively) slow moving subjects (e.g., people in a hospital); also, most of the reported work were carried out in indoor and (relatively) confined environments (e.g., offices, hospital rooms, etc.). The applicability of radio-based localisation system for locating *running* subject in a *large* indoor environment, for example a sprinter running in an indoor stadium, is unknown.

It should be noted that besides the athletes' speed profile, video footage is considered as an important element in coaching support. Traditionally, coaches hold hand-held cameras to capture footage of athletes in motion. The drawback is that such manual approach distracts coaches from the coaching session. An automated solution is therefore preferred. It is possible to install multiple cameras along-side the track to capture video footages of athletes running for long range (e.g., over 60m), but the drawback of this solution is that the cost increases substantially. An alternative solution that involves fewer cameras is one that spins a camera towards the athlete. To do so, the system must be able to approximate the location of the athlete on the track in real-time. One solution is to carry out real-time image processing of the video footage to track athletes. The authors suggest that radio-based localisation systems – such as the one presented in this paper – could provide location information of a running athlete in order to drive a camera in real-time.

To investigate the applicability of radio localisation system for locating a running athlete, the custom-designed SEnsing for Sports and Managed Exercise (SESAME) [1] nanoLoc (NNL) 2.4GHz radio-based localisation system was developed and evaluated. Through experimentations, the system shows that radio-based localisation is a promising approach with an average positional error of 21.9589cm using minimal equipment setup. The accuracy of the results is promising comparing to the 0.5m to 1m positional accuracy reported in existing literature. This paper is organised as follow: firstly, related work and the design challenges will be presented; secondly, the design assumptions and the NNL system will be presented; thirdly, the experiments and the results will be presented and analysed. The paper finishes with a conclusion and future work.

II. BACKGROUND

A. Related Work

Motion-capture optical-based systems [4][14] have been used in existing biomechanics research to capture 2D/3D motion data of athletes, including positional data. Although highly accurate (i.e., millimeter-level accuracy) [5], they are very expensive, have limited Field of View (FoV), and unless permanently installed, would require one calibration per setup. Also, multiple markers must be attached to the subject. Thus, to cover longer distance runs, multiple scanners would be needed which means cost would rocket. Also, since they are infrared-based systems, they could only be used indoor. The high monetary cost involved and the level of complexity in the setup process mean these systems are impractical for regular data collection. These systems are generally used for small-scale studies involving small number of sprinters over small number of trials, or used as gold-standard for system evaluation. The same restrictions apply to high speed video cameras (i.e., >2k frame rate) as well.

A split time monitoring system is reported in [10]. Split time is the time it takes for one to run for 10m. The system reports gold-standard comparable split time information of athletes using cost-effective light-sensors. But split time does not provide continuous location information. Although more light sensors could be added to improve the granularity of the information, this may create a scalability issue. The same applies to conventional Light-Gate (LG)-based systems.

 ISBN: 978-1-61208-100-7

There are several other methods for continuous location tracking: GPS-based systems have been used; with a repeater, it could be a solution for indoor tracking. However, the use of repeaters is illegal in many parts of the world including the UK and Europe. Also, high precision GPS systems are expensive. An alternative would be laser range finders [9]. Coaches have been using laser range finders for continuous location tracking of athletes, however, only on an occasional basis. A laser range finder is placed at the end of the track, and emits an infrared beam to a flat subject, commonly the lower back of an athlete. The time-of-flight of the (reflected) signal is used to determine the position of the athlete relative to the laser range finder. However, laser range finders suffer from the following drawbacks: a) a laser range finder must be placed behind an athlete, and the operator (i.e., the coach) must manually adjust the finder to point at the flattest surface of the athlete (i.e., the lower back); a task which is increasingly difficult when the athlete runs further away from the finder; and b) they are expensive. Although automated laser range finders - such as Total Station - are available, they are even more expensive than conventional laser range finders.

Another approach for continuous location tracking would be to the use of on-body sensors such as inertial sensors [15]; high quality inertial sensing systems, however, are expensive. In [11], an integrated on-body and track-side sensing system which detects step/stride length was reported. Stride length is the forward displacement of on the same foot during a stride. A series of accurately measured stride length would enable one to determine the location of the subject, provided that the starting position and the direction of the run are known. An alternative approach would be radio-based localisation systems. These systems use different types of radio [12][13], each with different types of characteristics and accuracy. Radio-based systems, however, subject to noise. There are reports on radio localisation data analysis protocols, such as the Curvilinear Component Analysis (CCA) [16] for processing multi-dimensional localisation data. In [13][17], the algorithms for interference-aware radio-based localisation systems were presented. In [18], the work on locating a relatively slow moving pedestrian in an indoor environment was reported. The use of (extended) Kalman Filter for robot localisation was reported in [19][20]; however, the speed of the robots was relatively slow and the results are insufficiently accurate for the purpose of this study.

B. *Design Challenges*

Spaces available on athletes for attaching on-body sensors are limited: the authors' interviews with the coaches and athletes suggest that they would prefer to minimise the number of on-body equipment to avoid affecting athletes' performance. Athletes do have a more open attitude towards placing sensors at more static and rigid locations, such as the lower back, where motion obstruction caused by sensor attachment is negligible. Thus, care must be taken to design systems so that they are small in size and light in weight. Also, since the track is a shared environment, the number of on-track equipment should be minimised to avoid disturbance to other users.

III. SYSTEM DESIGN

A. *System Assumption*

The NNL system is deployed and evaluated in an indoor environment; this is because training commonly takes place in indoor stadiums in the UK; also, optical-motion systems, such as CODA, could be used for evaluation purpose. Since the system is portable, radio-based, and uses a small number of track-side equipment, the system is provisioned to be deployed in outdoor environment as well. As a preliminary study, the investigation will start with using the NNL radio-based localisation system for locating an athlete sprinting on a straight (i.e., a 60m straight). Energy consumption of devices is not important, as each sprint is no more than a few seconds and athletes only train for a few sprints per day training session; thus, batteries could be re-charged or replaced at the end of a training session. Since the track is a shared environment, the number of on-track and on-body equipment should be minimised. The track is assumed to be clear when the athlete runs; this is a valid assumption for safety purpose. Safety and security issues are not addressed.

Radio-based localisation systems are sensitive to changes in the surrounding environment, such as new additional infrastructure; however, it is fair to assume that during a day raining session, the surrounding environment does not change. In Section VII, provisionings in the finalised system that minimise the effect on the system's accuracy due to changes of the surrounding infrastructure will be presented.

B. *System Design*

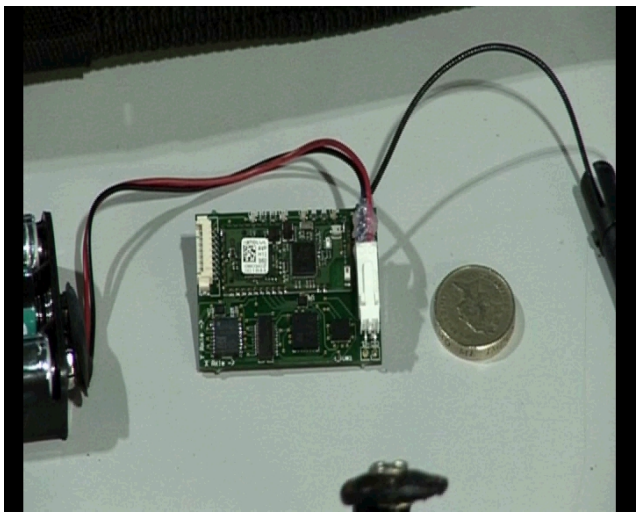


Figure 1 – The custom-designed NNL board

The NNL system has track-side anchor(s), on-body tag(s), and a track-side sink which is connected to a laptop. The on-body tag sends radio signal to the track-side anchor(s) which is(are) placed at known position(s), and uses Time-of-Arrival (ToA) to determine the distance between itself and the anchor(s). The (raw) distance value calculated by the tag is uploaded to the track-side sink. The system supports both Multiple Anchor (MA) and Peer-to-Peer (P2P) mode. The former requires multiple anchors to be placed at known positions on the track and uses triangulation to locate a subject,

 ISBN: 978-1-61208-100-7

whereas the latter uses only one anchor (which is placed behind the athlete) and is therefore suitable for straight runs. Another advantage of the P2P mode was that it involves less equipment on the track, which is ideal for experimentation in a shared environment at this early stage. Thus, in this paper, the results of the system using the P2P setup are reported.

The custom-design localisation board (Figure 1) is used for implementing the anchors, tags, and sink. Thus, the only difference between the three types of nodes lies within their functionalities (i.e., software). Each board is only a few millimeter thick and half the size of a credit card, and each has a nanoLoc AVR chip from nanotron [3]. Each board is equipped with an ATmega644 processor and a 3-axis accelerometer and gyros. The boards operate in the 2.45GHz ISM band for both localisation and wireless data transmission. The effective wireless range in an indoor environment with a 12dBi directional antenna was sufficient enough to cover a 60m indoor track. The use of a directional antenna improves signal quality, hence reduces packet loss, and the range of coverage. For ranging, the system uses the Symmetric Double-Sided Two-Way Ranging (SDS-TWR) protocol. Due to space limitation, readers are referred to [3] for details on the protocol. The double-side and two-way approach of the SDS-TWR protocol enables compensation of internal hardware delays, time drifts, and wireless transmission delays; hence eliminating the need of explicit wireless synchronisation between devices. Thus, the major advantage of the protocol is that it is asynchronous, which means no synchronisation is needed among the boards involved in ranging.

IV. EXPERIMENT

A. *Experiment Setup*

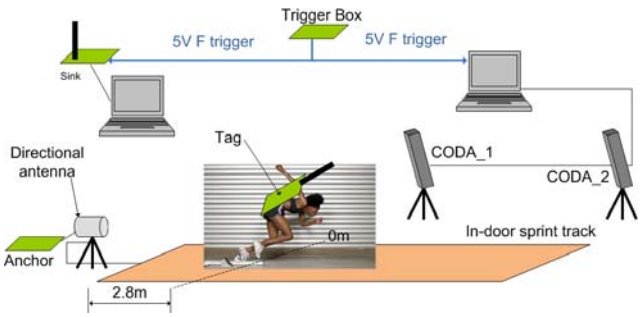


Figure 2 – Experiment equipment setup

Figure 2 shows the equipment setup. The anchor (with a directional antenna) is placed on a tripod (i.e., 0.76m above ground) and is placed at 2.8m behind the 0m line. The on-body tag is securely mounted to a belt which is worn by the subject (i.e., 1.12m above ground); thus, the tag is close to the Centre-of-Mass (CoM) (i.e., lower back) of the subject (Figure 3). Gold-standard passive optical motion tracking system, CODA from CODAmotion, was used for evaluation. CODA was chosen because of its well-documented high accuracy (i.e., millimetre-level). The CODA scanners were placed horizontally to the track, each has a viewing FoV of ~7m; thus, the total FoV was ~13m (with some overlapping between the scanners). A CODA marker is placed on the right side of the tag, so that the marker is within direct line-of-sight with the track-side CODA scanners in order to track the forward displacement of the tag (i.e., CoM of the subject). CODA was set to sample at 400Hz.

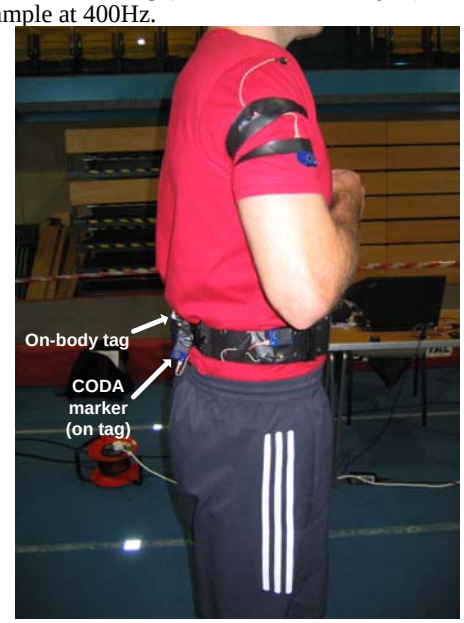


Figure 3 – On-body equipment

It should be noted that a unique feature of sprinting experiment is that the experiment runtime is very short and there is no need to capture positional information beyond each sprint. Since crystal clocks drift linearly and the experiment runtime is very short, the effect of clock drift is minimal. TRIG IN was therefore chosen as the cross-subsystem synchronisation method between CODA and NNL: a 5.5V falling edge trigger was delivered to both systems through a BNC cable. An alternative method to TRIG IN is SYNC IN, in which all systems are driven to sample by the same clock. CODA provides a SYNC OUT function which could be used to drive other systems to sample. However, SYNC IN is currently not supported in NNL because NNL was designed operate independently. Note that for longer experiment, multiple triggers could be sent to synchronise the systems to address clock drift: this is possible because both NNL and CODA logs the incoming trigger signal in a separate ADC channel from their data channels. A series of common triggers could be generated as square waves by a signal generator. This arrangement facilitates for continuous and concurrent data sampling and trigger signal sampling (i.e., continuous synchronisation through multiple triggers).

B. *System Calibration*

It is well known that radio-based localisation systems are subject to noise and bias. The question is the repeatability of these parameters. Calibration is needed to address bias in the system. To calibrate the system, the subject was asked to stand at different known positions on the track for 30s (i.e., at every 5m away from the 0m line, up to 60m). The raw distance

 ISBN: 978-1-61208-100-7

values reported by the NNL system at each known position are averaged to determine the bias value associated with that position.

Note that two sessions of the experiments were conducted. This is because, even though radio subjects to noise, calibration is only needed *per experiment setup*, or *per infrastructural change* (i.e., major constructional changes in the surrounding environment), but not per repetition (rep.). Repetition is the term used by coaches to refer to a sprint. This argument is justified by carrying out the experiments over two sessions: the calibration on the second session reports a similar bias.

V. Results and Analysis

A. *Low-Pass Filtering and Correction*

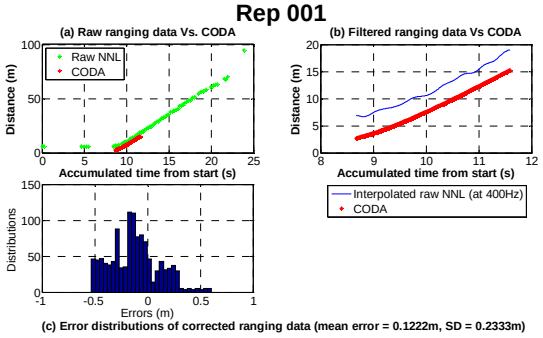


Figure 4 – Error distributions of corrected ranging data (rep. 1)

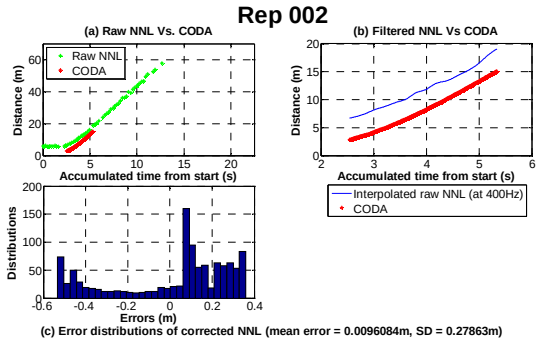


Figure 5 – Error distributions of corrected ranging data (rep. 2)

One approach to correct bias in the data is through modeling. However, given the level of variability caused by a relatively significant change in pace (i.e., acceleration) at the start-up phase of a run, a modeling approach would be difficult, as reported in [2]. In this section, the raw ranging data are first low-pass filtered, then corrected using the calibration data collected as described in Section IV.B. Fast Fourier Transform (FFT) was used to determine a suitable cut-off frequency (i.e., 1Hz), and the filtered ranging data are corrected using a piecewise linear model and the calibration data. The corrected ranging data are then compared with CODA data for error analysis. Some of the selected results are shown in Figure 4 to Figure 8 respectively. Note that because the NNL ranging data and CODA data were sampled at different rates, the corrected NNL ranging data are interpolated at 400Hz for error analysis.

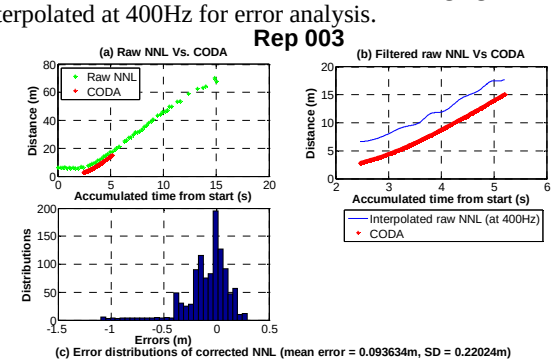


Figure 6 – Error distributions of corrected ranging data (rep. 3)

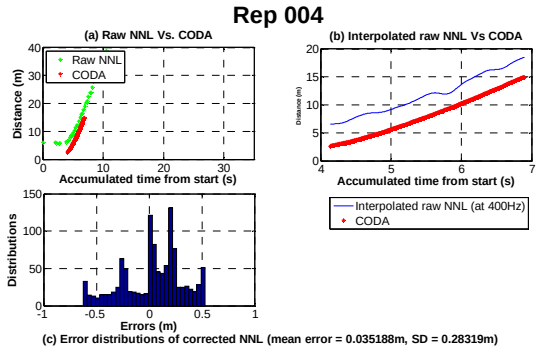


Figure 7 – Error distributions of corrected ranging data (rep. 4)#

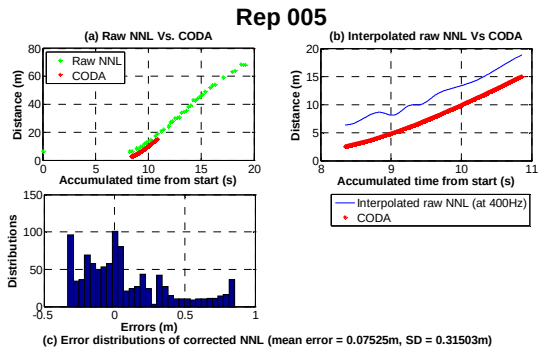


Figure 8 – Error distributions of corrected ranging data (rep. 5)

The averaged positional error of all reps. is 6.7172±26.6078cm (mean±SD), which gives an overall positional error of 20.0211cm. There were a number of factors which would have contributed to the error: a) mean error and SD were useful to evaluate systematic error and noise; but the lower the cut-off frequency, the smaller the SD, the estimated trajectory is smoother. To justify this point, the analysis process was repeated but with a higher cut-off frequency at 4Hz, the mean error was reduced but the SD was increased (i.e., 6.03248±31.138cm); b) it was assumed that the direction of movement of the tag is along the forward plane only. This is a valid assumption because sprinters are trained to run in a straight line (to maximise their speed) and lane crossing is strictly prohibited; in reality, however, sprinters could drift slightly off the centre of their assigned lane, and each lane has

 ISBN: 978-1-61208-100-7

a width of 1.21m; c) track-side equipment's positional measurements were done manually; which could contribute to the error; and d) because NNL and CODA has a different sampling rate, corrected NNL data are interpolated at 400Hz for error analysis; should some of the corrected values were out-of-range, interpolation would contribute to the error as well.

B. *Real-time Approximation*

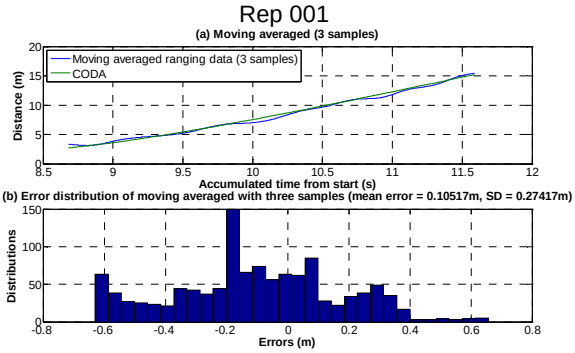


Figure 9 - Error distributions of the moving averaged data (rep. 1)

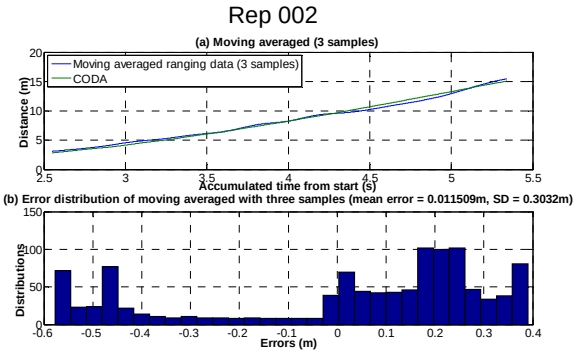


Figure 10 - Error distributions of the moving averaged data (rep. 2)

It should be noted that, a light-weight localisation protocol for approximating the location of an athlete in real-time would be preferred; one that is sufficiently accurate enough for spinning the camera at the athlete. Note that a typical camera placed on the track side has a FoV of ~5m (dependent on size of lens used). The FoV increases when the camera is placed further away from the track, which is the likely scenario (i.e., permanently installed cameras in a sport stadium – such as photo-finishing cameras - are usually installed in the roof or somewhere high on the side walls). Thus, one could tolerate a relatively larger error in the localisation system if the data were used for driving a camera. It should be that the design of the mechanical mechanism to spin a camera is out of the scope of this paper.

Thus, the use of an alternative light-weight algorithm for approximating the location of a running athlete in real-time is investigated in this section. The algorithm involves moving average and correction in which three consecutive raw ranging data are averaged (i.e., a raw ranging data is averaged with its neighbouring value immediately before and after itself). The averaged value is subsequently corrected using the known calibration data. Since the typical FoV of camera is ~5m, the delay incurred by waiting for three samples is therefore acceptable. It should be noted that this is a proof-of-concept experiment, which means the presented system is not restricted to a moving average of three samples. Figure 9 to Figure 13 shows the error distributions of the selected reps. (i.e., interpolated NNL results against CODA). Note that interpolation is not needed to spin the camera, but needed for error analysis only.

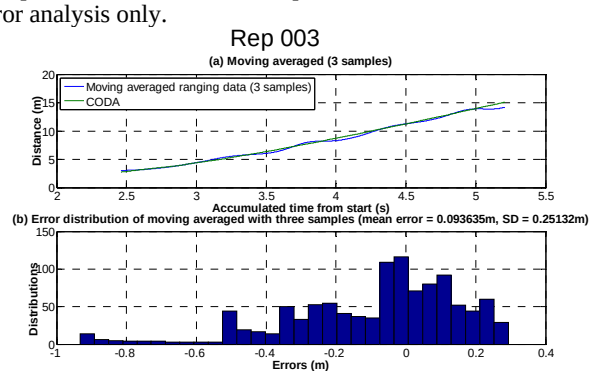


Figure 11 - Error distributions of the moving averaged data (rep. 3)

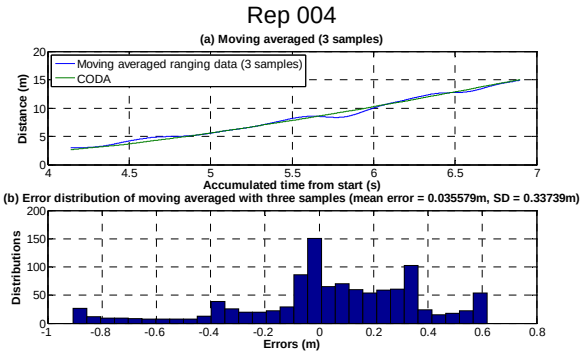


Figure 12 - Error distributions of the moving averaged data (rep. 4)

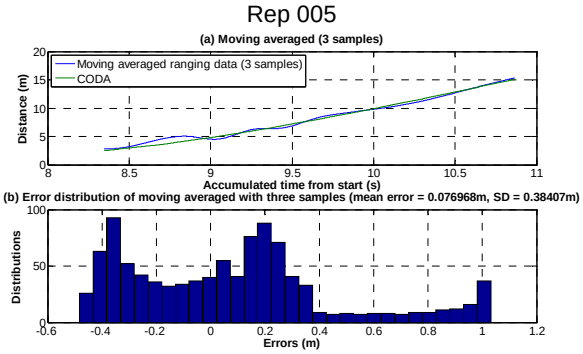


Figure 13 - Error distributions of the moving averaged data (rep. 5)

The results show that the error is 6.4574±31.003cm (mean±SD); thus, an average error of 21.9589cm. Both the mean error and SD are higher than the results using filtered and corrected approach reported in Section IV.A. This is expected because of the presence of noise due to the decision to compensate accuracy for performance (i.e., a faster response time for spinning a camera). Consider the typical FoV of a camera is in excess of a few meters, the authors

argue that this level of accuracy is sufficiently enough for spinning a camera to follow an athlete in motion.

VI. CONCLUSION

In this paper, a radio-based localisation system that is capable of accurately locating a running athlete on an indoor running track was presented. The system operates in the 2.4GHz band and uses ToA as the ranging protocol for localisation. The system supports either a multi-anchor mode, which includes multiple anchors placed at known positions on the track, or a P2P mode, in which only one anchor would be needed. A range of experiments using the system in P2P mode were conducted and the results show that radio localisation technique is a promising approach with an average positional error of 21.9589cm. The authors suggest that such level of accuracy is sufficiently enough for supporting an automated camera-based video tracking system to track athletes.

VII. FUTURE WORK

It was discussed in Section II.B that the presented system is not limited to indoor but also outdoor. With a directional antenna, the range of coverage in an outdoor environment could reach over 130m. To evaluate the system's performance in an outdoor environment, a light-weight GPS board is being developed. Part of the future work is to develop a spinning motor what spins a camera at an athlete based on real-time localisation data from the NNL system using a gumstix computer. Gumstix was chosen because of its wide range of functionalities, although the system will not be restricted to gumstix. Another part of the future work includes installing the anchors in the roof of the stadium. The purpose of doing so is to avoid the need of placing equipments on the track, hence minimising the level of disturbance to other users. Also, calibration would be needed only when there was a substantial infrastructural change to the system's location. This arrangement would also enable one to experiment with the multiple-anchor setup of the NNL system, which would allow one to do 2D tracking via triangulation (i.e., for oval track localisation).

With regard to the on-body equipment, the authors' observation is that coaches and athletes prefer tiny, light-weight on-body equipment. This requirement means three further areas of work: a) a new version of the on-body sensor board is underdevelopment, the new version of the board is the size of a one-euro coin; b) the omni-directional antenna of the on-body tag must be replaced; probably by a chip-antenna on the NNL AVR chip; and c) currently, the sensor is attached to a belt which is worn by the subject; the sensor must be firmly attached to the subject's body for safety reason as well as reducing the fluctuation of orientation of the antenna; thus, a better, less intrusive, user-friendly sensor attachment will be investigated.

ACKNOWLEDGEMENT

The authors would like to thank Rae Harbird, Scott Simpson, Michelle Manning, Philipa Jones, Tim Exell, Alex Atack, Gen Williams, Dawn Tighe, Karin Jerpers, Sharon Warner, Ashweeni Beeharee, Graeme McPhillips, Simon Julier, and Venus Shum for their contributions and support. The authors would also like to thank the athletes who kindly agreed to participate in the studies and experiments. This work was funded by EPSRC grant number EP/D076943.

REFERENCES

[1] The SEnsing for Sports And Managed Exercise (SESAME) project, http://www.sesame.ucl.ac.uk, 1st Jun 2009.

[2] H. Tan and Wilson AM. 2008. Measurement of stride parameters using a wearable GPS and inertial measurement unit. Journal of Biomechanics, Volume 41, pp. 1398-1406.

[3] nanoLoc Development Kit v1.4, nanotron Technologies, http://www.nanotron.com/EN/PR_nl_dev_kit.php, 1st Jun 2009.

[4] CODAmotion, http://www.codamotion.com, 1st Jun 2009.

[5] Charnwood Dynamics Ltd. (2006). CODA cx1 User Guide

[6] M. Kranz, W. Spiessl, and A. Schmidt, "Designing Ubiquitous Computing Systems for Sports Equipment", in Proceedings of IEEE PerCom 2007, pp. 79-86.

[7] E. Chi, "Introducing Wearable Force Sensors in Martial Arts", in Pervasive Computing Magazine, 04(3):47-53, July-Sep 2005.

[8] L. Cheng and Stephen Hailes, "Analysis of Wireless Inertial Sensing for Athlete Coaching Support", in Proceedings of IEEE Global Communications Conference (GLOBECOM), New Orleans, USA, Dec 2008.

[9] Loke Engineering, http://www.loke-engineering.de/3.0.html, 1st Jun 2009.

[10] L. Cheng, H. Tan, G. Kuntze, I.N. Bezodis, S. Hailes, D.G. Kerwin, and A. Wilson, "A Low-cost Accurate Speed Tracking System for Supporting Sprint Coaching", in Proceedings of the Institution of Mechanical Engineers, Part P, Journal of Sports Engineering and Technology.

[11] L. Cheng, G. Kuntze, H. Tan, D. Nguyen, K. Roskilly, J. Lowe, I. N. Bezodis, T. Austin, S. Hailes, D. G. Kerwin, A. Wilson, and D. Kalra, "Practical Sensing for Sprint Parameter Monitoring", in Proceedings of the 7th IEEE Sensor, Mesh and Ad Hoc Communications and Networks (SECON), Boston, Massachusetts, USA, June 2010.

[12] A. Ladd, K. Bekris, A. Rudys, D. Wallach, and L. Kavraki, "On the Feasibility of Using Wireless Ethernet for Indoor Localisation", in the Proceedings of IEEE Transactions on Robotics and Automation, Vol. 20, No.3, June 2004.

[13] B. Song, H. Lee, and K. Chung, "Toward A Totally Solving Interference Problem for Ultrasound Localization System", in the Proceedings of Optical Internet and Next Generation Network (COIN-NGNCON), Jeju, South Korea, July 2006, pp. 162-164.

[14] Qualisys, http://www.qualisys.com, 1st Jun 2009.

[15] MTx, xSense, http://www.xsens.com, 1st Jun 2009.

[16] L. Li and T. Kunz, "Localisation Applying an Efficient Neural Network Mapping", in Proceedings of the 1st International Conference on Autonomic Computing and Communication Systems, Rome, Italy, 2007

[17] Y. Shen, Y. Cai, and X. Xu, "Localized Interference-aware and Energy-conserving Topology Control Algorithms", in the Proceedings of Wireless Personal Communications: An International Journal, Vol. 45, Issue 1 (April 2008), pp. 103-120, 2008.

[18] A. Ladd, K. Bekris, A. Rudys, D. Wallach, and L. Kavraki, "On the Feasibility of Using Wireless Ethernet for Indoor Localisation", in the Proceedings of IEEE Transactions on Robotics and Automation, Vol. 20, No.3, June 2004.

[19] S. Kwon, K. Yang, and S. Park, "An Effective Kalman Filter Localization Method for Mobile Robots", in Proceedings of the 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems, October, 2006, Beijing, China.

[20] L. Freeston, "Applications of the Kalman Filter Algorithm to Robot Localisation and World Modelling", Final Year Project Report, Electrical Engineering, University of Newcastle, NSW, Australisa, 2002.

 ISBN: 978-1-61208-100-7

Experience and Vision of Open Innovations in Russia and Baltic Region: the FRUCT Program

Sergey Balandin
Nokia Research Center, Helsinki, Finland, Sergey.Balandin@nokia.com

Abstract — **This paper discusses our vision and experience of developing Open Innovations framework program, how it was implemented and why. The Open Innovations paradigm is a new trend for performing research and development that was proposed only a few years ago in 2003. The main emphasis of this approach is on setting cooperation in form of direct industry-to-academia joint R&D project and development of the corresponding competence incubators around the most relevant technologies. As an example this paper presents the open innovations framework program FRUCT that is targeted in development of telecommunications R&D ecosystem in Russia and Baltic region. The need for such cooperation is already recognized by the local industry, academic community and government authorities. In particular, Nokia and Nokia Siemens Networks expressed readiness and interest to invest into this project by contributing competences and providing some financial support.**

Keywords: Open Innovations; FRUCT; Industry-to-Academia; R&D cooperation; new trends in innovations; Russia-Finland partnering; Russia-Baltic cooperation.

I. INTRODUCTION

The paper presents a project targeted in developing mobile R&D ecosystem of Russia and Baltic region, with the main emphasis on setting up industry-to-academia competence incubator operating in the format of open innovations. Continues development of the strategic partnership between industrial and academic research is a key success factor of the modern innovation ecosystem. There are a few success stories of such strategic partnership frameworks functioning in different parts of the world. These programs bring significant benefits to the involved parties and fueling their further R&D units. As we know the fundamental science driven by the universities and other academic organizations should not be directly attached to the existing industries, but industrial research would benefit by early access to the results and information about main trends and weak signals. At the same time many universities also active in the applied research, but to be efficient they need feedback channel from the industry. Another key driver for setting stronger connection between academia and industry is that the time between a moment of innovation and its adoption by the industry is getting shorter and shorter. An interesting new trend for addressing this need is by building open innovation frameworks targeted in developing strategic partnership between industrial and academic research. Such framework programs help to find right research partners and jointly incubate new competences.

Nowadays the USA universities are the recognized leaders in adaptation of the academic research and education to the existing industrial needs, just look to the density of industrial presence in Silicon Valley. This situation creates a strong demand for quick and adequate actions from universities in Russia and Europe. A number of cooperation frameworks have been built inside the EU, e.g., Framework Program 7 [1]. However, the cooperation between Europe and Russia is still leaves a lot to be desired. This creates a historical chance for Finnish universities to use geographical proximity and traditionally good relations with Russian colleagues to strengthen Finnish science. Such cooperation is in clear mutual benefit, as among other advantages it will give to Finnish academia a priority path for accessing the huge pool of highly qualified talents and new innovative competences and help Russian universities to better integrate into the EU academic institutions and consequently will contribute in development of the bridge between academic and R&D worlds of EU and Russia.

There is a number of well known and hidden thresholds on the road to long-term collaboration and partnering. The most critical are need of mutual trust, lack of awareness about partners' capabilities, need for significant initial investments, and so on. As a result many good collaboration opportunities die at the very initial phase. This is especially true when thinking of R&D cooperation in countries that do not have long history of cooperation with global industry players. At the same time these regions have large undiscovered R&D potential, e.g., "non-traditional" solutions, new bright ideas that are not well known outside of a particular team and so on. Industrial research can benefit by getting early access to this "box of secrets" and the first players that manage to achieve it will win the most.

This paper describes our experience of building, managing and developing Finnish-Russian University Cooperation in Telecommunications (FRUCT) Open Innovations Framework program [2]. The FRUCT program was established in 2007 by a group of enthusiasts supported by Nokia and two universities. By now the program unites teams from 18 universities, Russian Academy of Science and is supported by Nokia and Nokia Siemens Networks, the companies that are recognized long-standing leaders and drivers in their segments of the ICT industry.

In particular, the paper focuses on two key aspects of the FRUCT program, which are less understood by the

 ISBN: 978-1-61208-100-7

outside observers, but are the key factors contributing to the success of program:

- Main elements and principles of FRUCT framework as a whole and our approach to the program management and technological steering;
- Principles of organizing, managing and sharing results of the joint R&D project between FRUCT member teams.

The experience of presenting FRUCT principles to externals and followers shows that the solutions for these two points are the most complicated for understanding of FRUCT and Open Innovation paradigm. So this was an original motivation to prepare the paper and put these topics to the open discussion in R&D community.

The paper is organized as follows. The next section gives an overview of FRUCT mission, motivation for the member organizations, general principles of operation, expected deliverables and achievements. The third section specifically addresses core principles of FRUCT management on the level of whole program and each particular project. The main points of this paper are summarized in conclusion section, which is followed by acknowledgements and the list of references used in the study.

II. General Overview of the FRUCT Program

This chapter gives an overview of FRUCT program, its mission and main principles of operation. The theoretical basement of the program is the principle of open innovations proposed by Prof. Henry Chesbrough from UC Berkley [3]. Open Innovations is a paradigm that assumes that firms can and should use external ideas as well as internal ideas, and internal and external paths to market, as the firms look to advance their technology [4]. The boundaries between a firm and its environment have become more permeable; innovations can easily transfer inward and outward. The central idea behind open innovations is that in a world of widely distributed knowledge, companies cannot afford to rely entirely on their own research, but should instead buy or license processes or inventions from other companies. In addition, internal inventions not being used in firm's business can still give benefits outside the company, e.g., through licensing, joint ventures, spin-offs [5]. However, implementation of the open innovations principle requires significant ecosystem preparation work and framework that provides trust and other components that fuel open innovations.

Nowadays we can claim that the FRUCT program implements main principles of open innovations plus we done further theoretical and practical development of the basic principles in the part related to the ecosystem preparation to open innovations. The program aims in increasing level of competences and visibility of the member organizations, especially awareness about Russian research in Europe and wise versa. Russian universities have good reputation and traditions in fundamental science. However, visibility and presence of young Russian scientist in the international scientific community is a clear area for further improvement. Historically the first area of the corresponding development selected by FRUCT was development of competences and infrastructure that allows young talents to participate and publishing papers at the top international conferences. At the same time European, e.g., Finnish universities gets this way the great opportunity to access the largest pool of talents and resources in Europe.

The FRUCT program was established in 2007 by the group of individuals, Nokia Research Center, Saint-Petersburg University of Airspace Instrumentation and University of Turku. Originally the program was targeted in facilitating cross-boarder R&D cooperation between industrial and academic organizations of Finland and Russia. By now FRUCT become the most significant and actively growing cooperation framework between leaders of ICT industry and universities in the Baltic region. At the moment the FRUCT community consists of representatives of 18 universities from Russia, Finland and Denmark, three industrial companies represented by their R&D units, and R&D institute of Russian Academy of Science. The main FRUCT principle is in removing cooperation entering thresholds by setting R&D projects run by students under direct joint supervision and tutoring of industrial and academic experts. The FRUCT principle of building cooperation through the joint student-driven R&D projects has been proven to be very efficient for identifying and incubating the demanded competencies. The big advantage is that initially all cooperating sides take minimal obligations, with low financial and image risks. As a result a number of cooperation activities have been started and successfully developing. Care of the personal development of involved people and teams is the key priority of the program. The FRUCT students have successful represented the program in a series of international contests, including Symbian Student Essay contests, Widsets coding contests, and so on. Also a number of good international publications have been done within the FRUCT scope and we scale of activities is growing fast.

Generally FRUCT program promotes mobile device oriented research, telecommunication and information technologies. The directions of research within the FRUCT program include (but are not limited to) open source solutions and MeeGo/MAEMO mobile OS, smart spaces, physical air interface, embedded networks, mobile device software and service solutions, energy management and green technologies, security, and so on.

FRUCT program is based on two modes of cooperation: cooperation of R&D team in joint projects and regular face-to-face meeting (i.e., conferences, seminars, trainings, etc.) for gathering together all members of the FRUCT community. The R&D projects cooperation helps to the involved teams to learn about the capabilities of each other, building thus basic trust and understanding. The fact that projects are done by students is important as it minimize the involved costs and risks. Our approach focused on creating international groups of students supervised by industrial and university experts, which help directing the R&D work of the students in the most interesting and challenging areas of ICT R&D. In other words,

 ISBN: 978-1-61208-100-7

the program implements the project-based training, where students are oriented towards real creativity and contributing to the final concrete deliverables. Generally it is hard for industry and academy people to find the right partner for cooperation, but such small joint R&D projects provide the required basic interface and topic for setting direct contacts and find ground for commercial projects.

The key enabler factor of success in development of the strategic cooperation is to identify solid and well established niche where partners have unique and supplementing competences. The FRUCT program creates an environment to highlight the existing relevant R&D niches and what even creates new R&D niches around recently emerging technologies, which can be generated as a product of open innovation cooperation between industrial and university experts. The FRUCT also helps universities to incubate new competences demand for which is emerging on the industry side. The main goals of FRUCT program include:

- Identifying world-class R&D teams that are looking for partners and interested in open innovation cooperation;
- Creating the new competences and corresponding niches for R&D cooperation;
- Developing long-term strategic partnership between industry and universities;
- Providing chance for more students to realize scientific, R&D and career ambitions at the university through direct academia-to-industry cooperation;
- Promoting idea of Europe without borders and corporate social responsibility.

It is also well known problem that universities experience difficulties in keeping the best students, but close partnering with industry provides association with the strong brands, challenging and very concrete research tasks and additional resources, which attract students and help to solve the resource problem. On the other hand, the industry companies are interested to have long-term and high-risk research done by the universities, and benefit from getting closer to the edge of science, so that adoption of new key finding could be done even faster. Also the early industrial feedback is in mutual benefit as it allows right tuning and presentation of the new technologies. So driving into a stronger cooperation between the academic and industrial research, being more open and involved in joint activities, getting stronger visibility by making joint publications and so on, these all are also in the strong mutual benefit.

One of the most interesting questions is how this rather informal structure exists, how to define and control vector of development, what are the main management principles of FRUCT program.

III. MANAGEMENT OF THE FRUCT PROGRAM

The chapter presents two key aspects of FRUCT framework management solution: management of FRUCT program and management of R&D projects inside FRUCT. When looking from the program deliverables point of view, these two areas looks like one big task, but in fact in order to achieve efficient management and build robust and scalable solution we shall clearly separate these areas. The main focus of framework management is on maintaining and developing efficient and cozy environment that attracts best talents and creates motivation for people to actively contribute within scope of the open innovations principle. For that the framework must guaranty reliable and equal access to a set of benefits on both individual and group levels. The short summary of the FRUCT member benefits is as follows:

- access to new competences and professional growth opportunities for the involved individuals and teams;
- professional network, trust building and direct contacts with relevant top experts from academia and industry;
- contribution to the positive business and scientific visibility of the member teams and individuals.

Cooperation in the joint projects is the crystallization point that allows providing the above listed benefits, but in order to make it work a proper infrastructure and resource allocation has to be provided. The main FRUCT engine is a group of highly motivated enthusiasts of the open innovations paradigm, who believes that this kind of framework in future will result in incubation of innovative and highly-competitive businesses. Also FRUCT receives some financial support from the industrial members that are interested to develop the overall ecosystem of the region and also care of having image of the good corporate citizen. Together these two factors provides strong enough basis for delivery of the above stated benefits to the program members. It is important to mention that for academic members the program is free of charge and even industrial members have no direct financial obligation, but by default certain level of support is expected. FRUCT makes a lot of investment into development of the efficient environment to support the open innovation principles so that all FRUCT members can for free enjoy the following open innovations project enablers:

- FRUCT conferences (200+ attendees) are organized every half a year, plus FRUCT seminars, topical conferences, workshops and sessions are regularly organized as independent events or within scope of other well respected events in the field of communications;
- technological trainings, exchange lectures and courses, winter and summer schools, open repositories with various materials;
- support of the active FRUCT teams with the research enabling things, such as books, dif-

 ISBN: 978-1-61208-100-7

ferent kinds of devices, measurement tools, software licenses, project management tools, office equipment, and so on;

- travel grants and free registration/accommodation packages for presenting FRUCT project papers at the recognized conferences and exhibitions;
- R&D contests with good prizes and opportunity to gain high recognition and good publicity of the project results;
- R&D grants for students and in certain cases special student stipendiums.

In addition the members get a lot of other supporting benefits that keeps them attached to the program, e.g., discounts on different industrial, academic and scientific events, FRUCT facilitates PhD and MSc exchange between member universities, help students to find good academic and industrial consultants, team of co-authors for publications and thesis opponents, help forming consortia for participation in EU and national grants, etc. So let's now discuss in details how does it work.

A. Management of the FRUCT framework

Let's first define what is the FRUCT framework nowadays. FRUCT unites R&D teams from 22 organizations, located in 3 countries and 4 time zones, with at least 4 major cultural backgrounds. The number of people actively working in FRUCT exceeds 70 persons plus over 200 active followers. Five FRUCT laboratories are located at four Russian universities, in three different cities. The core FRUCT management team is 8 persons. The program is active for more than 3 years and over this time 12 projects were successfully completed and went to further collaboration phase. Over 70 publications were produced as a result of FRUCT activities and over 40 projects currently are under development. FRUCT is leading R&D activities of Russian MAEMO community and represents interests of Symbian Foundation community in Russia. Internet visibility is supported by the group of FRUCT sites, which includes the main web site, FRUCT forum, five fully operational sites of the second level and three more sites of the second level are under development. Only the main FRUCT site recently gets over 700 views per day. FRUCT follows EU and local calls for project grants, identify the most relevant topics and facilitates formation of the consortia. Also FRUCT carefully follows all new trends and weak signals in ICT industry, which is used for steering technological development of the community. At the same time FRUCT is quite unofficial organization that follows principles of the new Internet era, e.g., it does not have permanent staff 100% allocated for FRUCT bureaucracy, as all management positions are occupied by people working in the member organizations and having other duties beside the FRUCT activity. At the same time the core management team has significant work load in FRUCT, which create a number of personal and organizational challenges. This is the FRUCT scope. As you can see it is quite broad, challenging and non-trivial solution, so how it is managed?

Formally FRUCT framework is built on the principle of eManagement in matrix organizational structure. The key tools that support FRUCT management are: the existing group of web site and other Internet tools, well defined structure of on-site representation in the member organizations and procedure for internal communications, and regular face-to-face meetings, including social events plus joint free-time activities of FRUCT activists.

The regional representation of FRUCT is based on a network of FRUCT laboratories and teams that are created in the most active partner organizations. The decision to setup FRUCT laboratory should be seen primary as recognition of the existence of strong local FRUCT team and its ability to identify R&D niche with competitive level of expertise. The lab creation starts from selection of the future lab leader. There are two ways how the lab leader and correspondingly the lab organization can be formed. The lab leader can be recommended and supported by the official authorities of the member organization. Then FRUCT can either accept lab with the proposed leader or refuse it, or start negotiation about other potential candidates. The lab structure in this case is formed by the member organization according inline with the internal guidelines and traditions.

Another approach is used when initiative comes from the department level. Then the laboratory leader is selected via the "natural competition" when at the beginning all key members of the partner team get equal level of recognition and attention from the FRUCT board. This helps to identify natural leader of the team. If the team does not have natural leader, then such case is put on ice till the moment when the leader will appear. If team has more than one natural leader this approach helps us to identify the strongest leader. In certain cases FRUCT might even have two independent FRUCT laboratories in the same university.

Selection of the right person to the laboratory leader position is a key task for any organization, which is absolutely viable requirement for the open innovations type of organization, which by definition are connected by the weak ties and have small and weak bureaucracy structure that definitely cannot handle additional management overhead created due to lack of initiative at the laboratory level.

FRUCT labs get certain amount of money as a budget for internal needs. The money can be used for general expenses of the lab (e.g., traveling, small office expenses, etc.), for organizing lab-level events and support exchange activities, and so on. This gives to the lab and university people some level of financial independency, so that it internally can decide who deserve additional recognition, what additional equipment they need to buy, etc. The lab leaders regularly report the status and devel-

 ISBN: 978-1-61208-100-7

opment to the FRUCT board and budget stakeholders. The lab leaders together with the FRUCT general chair (and optionally other core team members assigned by the chair) define the lab development plans with horizon of half a year and two years. These plans are reviewed, re-evaluated and updated every half a year. Also every half a year the general chair organizes personal development discussion with the leader of each team, which is often held at the time of the main FRUCT conferences. The purpose of this discussion is to understand personal development priorities of the lab leader and team members. The lab lead is responsible for organizing similar discussions with all members of the team. As a result FRUCT cares about interests of the involved people, gets efficient mechanism for getting feedback and minimize probability of unwanted surprises due to loss of people that do not see for themselves benefits in maintaining FRUCT membership in their future. The lab leaders form a sub-team of local leaders; it is recommended that one person is responsible for each area of the strategic development of the laboratory. The package leaders are responsible for running package management tasks, defining and maintaining the work breakdown structures for the projects in their packages and be the main interface for the people involved in the package project work.

Another important aspect is how to handle cross-cultural management, defining common language and communication principles. We selected English to be the official language of FRUCT, even despite dominance of Russian participants. However to once again lower the entering barrier, recently FRUCT introduced couple of satellite resources (e.g., maemo.fruct.org) in Russian. But all official email exchange is in English. English is used on all main pages of FRUCT program and in the official news line.

Synchronization of the framework activities is achieved by setting regular telcos in Skype for the members of the core team. In addition, every day the general chair reserves open timeslot for telcos that is convenient for participants in all time-zones. This time is reserved specially for ad-hoc telcos so that any FRUCT member can call to the chair and provide direct feedback, ideas and recommendations. Also every month FRUCT publishes newsletters and sends major announcements via the community email distribution.

The above listed tasks are facilitated by use of Web 2.0 solutions at www.fruct.org, e.g., professional social network engine, blogs, event pages, project pages, lab pages, forum, wiki, inline commenting of news, etc.

The risk management is handled at both framework and project level. At the project level FRUCT use standard risk management planning, plus the lab risk management plan and budget provide the second line of defense against major risks. The third line is provided at the framework level, when all teams operating in cooperation project should have a plan what to do if peer's deliverable will not be available in time.

B. Management of R&D projects in FRUCT

As it was said in introduction, the main task of all FRUCT projects is creation of new competences and cooperation contacts between the involved parties. The actual R&D challenge should be mainly seen as the crystallization point around which the competences and cooperation activity is built.

According to FRUCT rules any representative of the full-member organization can propose a new topic for R&D project. Every FRUCT project should involve representatives from two or more member organizations. In the simplest case the project work group consists of students from one university, which are supervised by local professor and assigned industrial tutor. In more advanced cases the team might consist of the members from a number of FRUCT universities, industrial experts can directly participate in the research work and more than one supervisor and/or tutor from different universities and companies can assist to the team.

The key idea for such organization of the project teams is to lower the commitment threshold on the way to start real cooperation. It is obvious that before entering into a serious cooperation it would be comfortable for the involved parties to get know each other without any commitments. FRUCT projects provide such "sandbox" framework and additional time to learn about strong and weak sides of the partner. It also helps to get the required knowledge for fine tuning of the cooperation proposals. As a result the proposal can be presented in the format and under conditions that are most favorable for setting the full-scale long-term cooperation.

Another idea behind FRUCT projects is to promote project-based training to help young specialist be more creative and understand how to balance high-risk research with a need to deliver results within the agreed timeline. The eventual results of the project are not just classical deliverables like novel algorithms, signal structures, architectural solutions, software code, etc., but creation of the "competence incubation" infrastructure and a set of well-prepared teams capable of continuing challenging research and design work on their own.

The new FRUCT projects can be initiated by a professor, industry expert or even by a student. Of course, depending on the initiating party there are some differences in the procedure: when a professor initiates project, he/she has to take an obligation to be a supervisor of the project team, when the project proposal comes from the industry expert, the expert takes obligation to be team's tutor. In both these cases FRUCT performs lite-analysis of the project proposal, as the main responsibility for the quality of project proposal is on the initiating expert. When a student comes with the idea of a new project he/she should ask for a supervisor and tutor from the

 ISBN: 978-1-61208-100-7

FRUCT board. Such project proposals are evaluated by FRUCT Advisory Board members and invited experts. If the selected domain and problem definition look reasonable and challenging, the proposal is called accepted and FRUCT board helps student to find supervisor and tutor.

The main expected deliverables of FRUCT projects are: article in a good journal or/and conference, and/or code contribution to one of the open source initiatives. However depending on the project scope some other deliverables could be also possible.

As was discussed in the previous section, FRUCT does not provide direct financial support of the projects, but all projects can get support in expertise, books and device donations, tools for developing and managing projects, travel grants and free-participation packages for presenting project results at the approved conferences (with reasonable quality and visibility) and in certain cases even student stipendiums. The project results belong to the developer team, so FRUCT does not pretend to the results ownership, only to the parts the FRUCT leadership team contributed to.

In order to be accepted in FRUCT the project must pass standard project definition and evaluation procedure. Taking into account specifics of FRUCT principles and targets we developed own procedure for entering to FRUCT and special form that must be filled in by all candidate projects. The form and procedure steps are available at FRUCT web [1] and in fact represent the first tool that we give to the news teams to correctly and efficiently formulate the project proposal.

Each project must openly report main results and overall progress comparing it to the original targets every half a year at the FRUCT conferences. Plus the status progress reports are regularly sent to the corresponding lab and/or team leaders.

The project steering is performed by the FRUCT advisory board under personal control and responsibility of the selected tutor. For controlling project progress we recommend to the project teams prepare the timeline plan using Gantt chart [6], with clear specification of the main project phases and dates, names of responsible persons, summary of involved resources, plus also reflect major milestones in the status section of the project web page. In addition we introduced web-based Concurrent Versions System (CVS) solution as a part of the new FRUCT project management framework that is currently in pilot use. The web solution is also combined with the commenting, brainstorming, project and individual calendar tool and ideas aquarium tools.

The above listed tasks are facilitated by use of Web 2.0 solutions provided at fruct.org and the laboratory sites: project pages, professional social network engine, event pages, forum, wiki, CVS, calendar, ideas aquarium, brainstorming, feedback tool, whiteboard, etc.

IV. CONCLUSIONS

The paper gives an overview of the FRUCT framework, its scope and targets, with special attention to the approach for addressing the framework and project management tasks and what tools and solutions are used for that. As FRUCT was one of the first attempts to implement the open innovation principles in Baltic region. In this project we have faced a lot of challenges and had to implement a lot of program components without having reference examples for adoption of the best practices. Development of the proper management framework was the key element of the performed investigation and this study is not finished. FRUCT has working solution in hands, but still a number of questions are open for further study and discussion (e.g., degree of the solution scalability). The described management framework is under development and even some practices and solutions mentioned in the paper are still in piloting and not yet applied not to the whole FRUCT framework and all projects, but only to the selected subset. FRUCT is a “living” organizational structure and we all time try new things and approaches. Unfortunately due to the space restriction it was not possible to describe many other aspects of FRUCT program. Many management activities, such as management of the professional developer communities (maemo.fruct.org), organization of scientific and educational events (conferences, training, courses, PhD and MSc exchange, etc.), formation of consortia for applying to public funding and many other aspects were just mentioned in the paper. We believe that the paper provides strong enough arguments in favor of the open innovations paradigm, which are further supported by the reference to successful implementation.

ACKNOWLEDGMENT

Author thanks all active members of the open innovation framework program FRUCT, and especially Alexey Dudkov, Veronika Prokhorova and members of the FRUCT Advisory Board for their contribution to the overall success of the program. The special thanks also go to Nokia and Nokia Siemens Networks for donations provided to FRUCT program and all FRUCT experts who contribute their time and expertise to support development of open innovation principle in Russia and Baltic region.

REFERENCES

[1] Official web site of European Commission. Research: Framework Program 7: the future of European Union research policy, 2010. http://ec.europa.eu/research/fp7. Retrieved: 16.7.2010.

[2] Official web site of Finnish-Russian University Cooperation in Telecommunications (FRUCT) Open Innovation Framework program, 2010. http://www.fruct.org. Retrieved: 16.7.2010.

[3] H.W. Chesbrough, Open Innovation: The new imperative for creating and profiting from technology. Boston: Harvard Business School Press, 2003.

[4] H.W. Chesbrough, Open Innovation: The new imperative for creating and profiting from technology. Boston: Harvard Business School Press, p. xxiv, 2003.

[5] H.W. Chesbrough, The era of open innovation. MIT Sloan Management Review, Vol. 44 (3), pp. 35-41, 2003.

[6] KIDASA software. Gantt Charts, 2010. http://www.ganttchart.com. Retrieved: 16.7.2010.

 ISBN: 978-1-61208-100-7

Tracking Recurrent Concepts Using Context in Memory-constrained Devices

João Bártolo Gomes*, Ernestina Menasalvas* and Pedro A. C. Sousa †
*Facultad de Informática, Universidad Politécnica de Madrid, Spain
joao.bartolo.gomes@alumnos.upm.es, emenasalvas@fi.upm.es
† Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa
Lisboa, Portugal
pas@fct.unl.pt

***Abstract*—The dissemination of ubiquitous devices with data analysis capabilities motivates the need for resource-aware approaches able to learn in reoccurring concept scenarios with memory constraints. The majority of the existing approaches exploit recurrence by keeping in memory previously learned models, thus avoiding relearning a previously seen concept when it reappears. In real situations where memory is limited it is not possible to keep every learned model in memory, and some decision criteria to discard such models must be defined. In this work, we propose a memory-aware method that associates context information with stored decision models. We establish several metrics to define the utility of such models. Those metrics are used in a function that decides which model to discard in situations of memory scarcity, enabling memory-awareness into the learning process. The preliminary results demonstrate the feasibility of the proposed approach for data stream classification problems where concepts reappear and memory constraints exist.**

***Keywords*-Ubiquitous Knowledge Discovery, Data Stream Mining, Concept Drift, Recurring Concepts, Context-awareness, Resource-awareness**

I. INTRODUCTION

Learning from data streams in ubiquitous devices where the data distributions and target concepts may change over time is a challenging problem, known as concept drift [13]. In real world classification problems it is common for previously seen concepts to reappear [14]. This represents a particular type of concept drift [13], known as concept recurrence [1], [5], [10], [14], [15].

Prediction models usually change over time, for example in product recommendations where customer interests change due to fashion, economy or other *hidden context* [8], [14]. Several methods have been proposed to detect and adapt to concept drift [6], [13]. The usual approach is to use a forgetting mechanism and learn a new decision model when drift is detected [6]. A possible solution to exploit recurrence, is to store previously learned models that represent observed concepts, thus avoiding relearning a previously learned concept when it reappears [1], [5], [15]. The main drawback of this approach is associated with its memory consumption cost. Depending on the capabilities of the device where the learning algorithm is executed, this can be a critical factor. Thus resource-awareness [2], [3] should be considered in memory constrained scenarios. Examples of ubiquitous applications that use resource-constrained devices include intelligent vehicles, personal digital assistants (PDA), wireless sensor networks (WSN) or ambient intelligence systems to name a few.

Resource-awareness means monitoring the availability of the required resources and adapt accordingly. This has been addressed in [3] where a generic framework that uses the mining algorithm granularity settings in order to change the resource consumption patterns according to availability of different resources. For example the parameter that controls the number of clusters is defined as a function of the available memory.

To address the problem of concept recurrence in resource-constrained devices, we propose a method that adapts its behaviour according to available memory. It extends existing drift detection methods [6] by associating context information with learned decision models, under the assumption that recurring concepts are related with context [1]. Such context information is used to improve the adaptation of the learning process to drift. The method used to decide which models to discard according to memory constraints constitutes the new contribution of this paper. This method uses pre-defined criteria to assess model utility in order to discard low utility models, allowing the storage of new ones in situations of memory scarcity.

This paper is organized as follows. Section II describes the problem of concept recurrence with memory constraints, with its assumptions and requirements, followed in Section III by the proposed resource-aware solution. In Section IV the preliminary experimental results obtained are presented and discussed. Finally we provide some concluding remarks and outline future research work in Section V.

II. RECURRING CONCEPTS LEARNING PROCESS

Let us assume a system learning from a stream of records where concepts can change over time, based on an incremental version of the Naive Bayes algorithm [4]. Note that any other incremental classification algorithm can be used instead without loss of generality. The performance of the learning algorithm is monitored using a drift detection mechanism [6] that triggers an event when drift is detected.

 ISBN: 978-1-61208-100-7

To make this system resource-aware and able to handle recurring concepts, the following requirements must be met:

- **i)** adapt the classification model to concept drift.
- **ii)** recognize and use past models from previously seen concepts when these reappear.
- **iii)** use contextual information to improve adaptation to drift.
- **iv)** adapt memory consumption according to availability of resources.

We assume that:

- the target concepts are related to available context. This is the main motivation behind storing models together with context.
- the memory available in the device running the learning process can be obtained anytime through a known interface.

Drift Detection method - When the records distribution is stationary the classifier error-rate decreases as the number of training records grows. This assumption is shared by most approaches dealing with drift [13], as it is a natural property in real world problems where periods of stable concepts are observed followed by change to a new period of stability (with a different target concept). Proposed in [6] a method for drift detection uses this assumption to find drift events. This method stores the probability of misclassifying p_i = (F/i) and the standard deviation s_i = $\sqrt{pi(1-pi)/i}$, where i is the number of trials and F is the number of false predictions. These values are updated incrementally. Two levels are defined, a warning level and a drift level. Each of them is reached according to pre-defined conditions based on p_i and s_i and its minimum values $p_{i_{min}}$ and $s_{i_{min}}$. Note that it is possible to observe an increase in the error-rate reaching the warning level, followed by a decrease. This is considered a false alarm.

The continuous learning process consists of the following steps, as presented in [1]:

1) process the incoming records from the data stream using an incremental learning algorithm (base learner) to obtain a decision model capable of representing the underlying concept.
2) the drift detection method monitors the error-rate of the learning algorithm.
3) when the error-rate goes up the drift detection mechanism indicates,
   - **warning level**: store the incoming records into a warning window and prepare a new learner that processes incoming records while the warning level is signaled.
   - **drift level**: store the current model and its associated context into a model repository; use the model repository to find a past model with a context similar to the occurring context that performs well with the new data records (i.e., represents the current concept, this was presented by the authors in[1]). Reuse the model from the repository as base learner to continue the learning process as in point 1. If no model is found use the learner that was initiated during warning level period as base learner.
   - **false alarm (normal level after warning)**: the warning window is cleared and the learner used during the warning period is discarded. The learning process continues as in point 1, which is also the normal level in terms of drift detection.

III. Proposed memory-aware approach

Adaptation to concept recurrence depends on storing past models. In memory constrained situations, the challenge consists in assessing model utility and adapt the learning process to memory availability. Having a decision function will enable to decide which model to discard in such situations, freeing memory for a new model.

A. Preliminaries

1) Context: Context information depends on the data mining problem and is accessed through a known interface. We assume context modeling to be based on the Context Spaces model [11] where a context is represented as an object in a multidimensional Euclidean space. A context state C_i is defined as a tuple of N attribute-values,

$$C_i = (a_1^i, ..., a_n^i)$$

a_n^i represents the value of attribute n for context state C_i. A context space defines the regions of acceptable values for these attributes. In this work, we use numerical context attributes and the Euclidean distance as the measure to compare context states formulated as:

$$|C_i - C_j| = \sqrt{\sum_{K=1}^{N} dist(a_k^i - a_k^j)^2}$$

a_k^i represents the k^{th} attribute-value in a context state i. For numerical attributes distance is defined as:

$$dist(a_k^i, a_k^j) = \frac{(a_k^i - a_k^j)^2}{s^2}$$

where s is the estimated standard deviation for a_k. For nominal attributes distance is defined as:

$$dist(a_k^i, a_k^j) = \begin{cases} 0 & if \quad a_k^i = a_k^j \\ 1 & otherwise \end{cases}$$

We define comparison of contexts using the following formulation:

$$similarContext(C_i, C_j) = \begin{cases} true & if \quad |C_i - C_j| \leq \epsilon \\ false & if \quad |C_i - C_j| > \epsilon \end{cases}$$

ϵ is a pre-defined threshold (using zero as threshold means that the contexts are equal).

 ISBN: 978-1-61208-100-7

2) Model Storage: In the Naive Bayes learning algorithm, P_C is the table storing $P(C)$ which represents the observed frequency for each class C, and P_A is the vector that stores P(A_n|C) that represents the frequency table of each feature A_n given class C. This can be expressed as:

$$P_C = P(C) \quad and \quad P_A =< P(A_1|C), ..., P(A_n|C) >$$

In our proposed approach, storing a decision model M using this learning algorithm also requires storing:

- The most frequent context state observed during model M learning period as $freqC$=(a_1, ..., a_n), where each attribute value of $freqC$ is the most frequent value that each attribute takes in that learning period.
- The accuracy Acc_k of M is the accuracy value obtained during the learning period $_k$, with $numCRecords_k$ being the number of correctly classified records by M and $numRecords_k$ being the total number of records processed during $_k$. The accuracy value is updated if M is reused. This is formulated as:

$$Acc_k = \frac{numCRecords_k}{numRecords_k}$$

- T is the timestamp that records the time when the model M was stored.

Consequently each decision model M stored in the model repository is defined as the tuple:

$$M = \{P_C, P_A, freqC, Acc_k, T\}$$

3) Model utility metrics: The definition of model utility metrics is needed so that the mechanism can decide which models to keep or discard in situations of memory scarcity.

We propose the following metrics to define a selection function that chooses which model to remove from the repository,

- Model Accuracy - This metric represents the accuracy of the decision model for the period it was used.
- Mean Square Error - This metric measures the error of the decision model for a set of records (we use the records available in the warning window). The error prediction of model M_i, using the window W_n of n records in the form of (x, c), where c is the true class label for that record. The error of M_i on record (x, c) is 1 - $f_c^i(x)$, where $f_c^i(x)$ is the probability given by M_i that x is an instance of class c. The MSI_i metric can be expressed as:

$$MSE_i = \frac{1}{|W_n|} \sum_{(x,c) \in W_n} (1 - f_c^i(x))^2$$

- Context Similarity - We exploit the context stored within the decision models by using the similarity between contexts as a metric. This is used to maximize context heterogeneity between the models in the repository.
- Timestamp - The timestamp that is stored along with models can be used as a metric of model utility in cases where all the other metrics give the same value. Thus this metric can be used to break ties where the oldest model receives lowest utility according to the situation.

In situations of memory scarcity we make use of a function that selects from the model repository, using the proposed metrics, the model to be discarded. This function searches the model repository for the models with the lowest context distance, in order to maximize the context heterogeneity in the repository. From this subset the ones with highest mean square error are selected. If more than one model remains after this step, the model with lowest accuracy is returned. If still more than one model exists (i.e., the MSE and Accuracy metrics have the same value) the model with lowest timestamp is selected. In algorithm 1 the pseudo-code of this function is presented.

Algorithm 1 Model Selection Function: Selects the model to discard

Require: ModelRepository $MR \rightarrow M$:(P_C,P_A,C,Acc,T)

1: For each M_i, $M_j \in MR$ with $_i \neq _j$, compute $distance(C_i, C_j)$;
2: Let $Context_{min} \subset MR$, where $min(distance(C_i, C_j))$;
3: Let $Error_{max} \subset Context_{min}$, where $max(MSI_i)$;
4: Let $CandidateSet \subset Error_{max}$, where $min(Acc_i)$;
5: **return** $M_i \in CandidateSet$, where $min(T_i)$;

B. Proposed Extension of the Drift Detection Mechanism

We propose as an adaptation strategy to discard a stored model when the memory limit is reached and it is not possible to store further models. This simple strategy enables us to handle the problem of concept recurrence in scenarios with different memory constraints, which limits how many models can be kept for reuse. We are able to test and measure the accuracy loss in such scenarios.

The **sufficient memory** condition is defined as:

$$\begin{cases} if \quad (usedMemory + storageCost \leq MemoryLimit) \\ then \quad true \\ otherwise \quad false \end{cases}$$

The used memory is accessed through an interface with the device as assumed in Section II, $storageCost$ depends on the mining schema of the data records. Since all the stored models share the same data schema this value is a constant. This is a consequence of using the Naive Bayes algorithm as the size of the frequency tables that are stored (see Section III-A2) is determined by the mining schema. The $MemoryLimit$ value depends on the ubiquitous device memory given to run the learning process algorithm.

The proposed learning process is extended in the **drift level** case with the condition:

 ISBN: 978-1-61208-100-7

- if (not **sufficient memory**): discard the decision model returned by function (algorithm 1);

IV. Experimental results

In order to test the proposed learning process, an implementation was developed in Java, using the MOA [9] environment as a test-bed. The available evaluation features have been used and the *SingleClassifierDrift* class that implements the drift detection method [6] has been extended into our proposed approach. We used artificial and real world datasets to evaluate the proposed method.

A. Datasets

1) Artificial Dataset: As artificial dataset, the SEA Concepts [12] with MOA [9] as the stream generator was used. SEA Concepts is a *benchmark* data stream that uses different functions to simulate concept drift, allowing control over the target concepts and its recurrence in our experiment. The SEA Concepts dataset has two classes {class0, class1} and three features with values between 0 and 10 but only the first two features are relevant. The target concept function classifies a record as class1 if $f_1 + f_2 \leq \theta$ and otherwise as class0, f_1 and f_2 are the two relevant features and θ is the threshold value between the two classes. Four target concept functions are defined with threshold values 8, 9, 7 and 9.5 as proposed in the original paper [12].

2) Real World Dataset: As real world dataset we used the Electricity Market Dataset [7]. The data was collected from the Australian New South Wales Electricity Market, where the electricity prices are not stationary and are affected by the market supply and demand. The market demand is influenced by context such as season, weather, time of the day and central business district population density. The supply is influenced primarily by the number of on-line generators. An influencing factor for the price evolution of the electricity market is time. During the time period described in the dataset the electricity market was expanded with the inclusion of adjacent areas (Victoria state), which lead to more elaborated management of the supply as oversupply in one area could be sold interstate. The ELEC2 dataset contains 45312 records obtained from 7 May 1996 to 5 December 1998, with one record for each half hour (i.e., there are 48 instances for each time period of one day). Each record has 5 attributes, the day of week, the time period, the NSW demand, the Victoria demand, the scheduled electricity transfer between states and the class label. The class label identifies the change of the price related to a moving average of the last 24 hours. The class level only reflects deviations of the price on a one day average and removes the impact of longer term price trends. As shown in [7] the dataset exhibits substantial seasonality and is influenced by changes in context. This motivates its use as a real world dataset in our experiments.

B. Context and recurrent concepts definition

As context for the SEA dataset we used a numerical context feature space with two features a_1 and a_2 with values between 1 and 4. It was generated independently as a context stream where the context attribute a_1 is equal to the target concept function number, and a_2 value equals the target concept function 9 in 10 times, which introduces noise in the context stream. We generated 250000 records and changed the underlying concept every 15000 records. The test was executed with a 10% noise value as in the original paper [12], this means the class value of the training record is wrong in 10% of the records, testing how sensitive is the approach to noise.

For the Electricity Market dataset we have considered the classification problem to predict the changes in prices relative to the next half hour, using as predictive attributes, the time period, the NSW demand, the Victoria demand and the scheduled electricity transfer. As context we used the day of week attribute, as in [7] experiments using it lead to 10 different contextual clusters. We expect that the association of this context with the stored models achieves good accuracy results. However, one drawback of a real world dataset is that we do not know for sure what the actual *hidden context* is and when such changes occur, which makes it more difficult to evaluate the obtained results. This dataset was also used in [6] to test the drift detection method in real world problems, achieving good performance results.

C. Experiments

1) Reference experiment: For both datasets the approach proposed in this paper is compared in terms of accuracy with the *SingleClassifierDrift* implemented in MOA[9]. This represents a scenario without memory constrains to be used as reference for the two boundary cases (i.e., case where it is possible to store all the required models vs no models stored). The *SingleClassifierDrift* approach also uses the Naive Bayes algorithm and detects drift using the drift detection method [6]. When it occurs, the system learns a new model by forgetting the old one (i.e., it represents the case where the memory available to store additional models is zero). In the real world dataset we also compare results with an incremental Naive Bayes algorithm [4] (without any mechanism to adapt to drift), again to be used as reference.

2) Experiment with memory constraints: We compared the memory-aware approach in situations with memory constraints, with 7KBytes and 5KBytes and 3KBbytes of available memory. This allowed to test different scenarios, ranging from ones where it is possible to store enough models to represent different target functions and others where due to the strong memory constraints only a reduced number of models can be stored. Note that in the SEA of concepts dataset, this means being forced to store less models than existing target functions. This allowed us to observe and measure how the accuracy declines as the

 ISBN: 978-1-61208-100-7

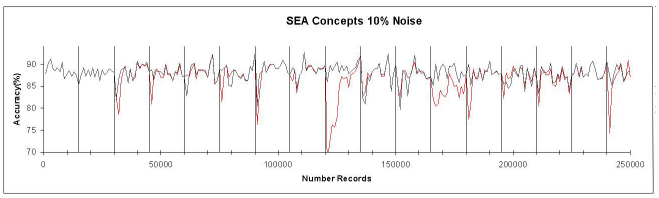


Figure 1. Comparison of accuracy with Proposed approach(black) vs SingleClassifierDrift(red) using the SEA concepts dataset. Black lines show when drift occurs

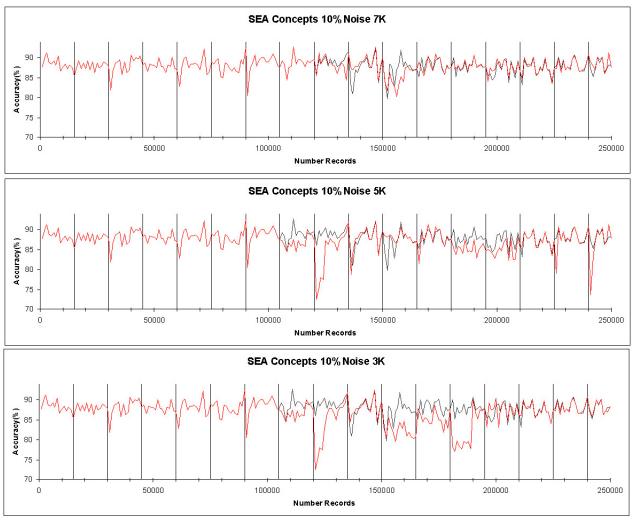


Figure 2. Comparison of accuracy with of the proposed approach with NoConstraints(black) vs Memory-constrained(red) using the SEA concepts dataset. Black lines show when drift occurs.

memory available is reduced and fewer models can be kept. Also it is important to understand the impact of discarding models and the utility of the stored models in the adaptation to recurrence in such memory constrained scenarios.

D. Results with SEA of Concepts Dataset

As can be seen in Figure 1 our approach leads to better accuracy than *SingleClassifierDrift*. In general our approach adapted to drift faster and the models selected using context integration were able to represent the target concepts. This is not observed in the *SingleClassifierDrift* approach that always has to relearn the underlying concept from scratch after drift is detected. It is also noticeable that our approach achieves a more stable accuracy over time, as it recovers much faster from drift than the approach without stored models. The proposed approach obtained 2046 more correct predictions. The integration of context enables to exploit the associations between recurrent concepts and context as a way to track concept recurrence. In situations where this association exists it is possible to achieve better results.

In Figure 2, the memory-aware approach is compared in scenarios with different available memory values. As expected, when memory is reduced, the accuracy is reduced.

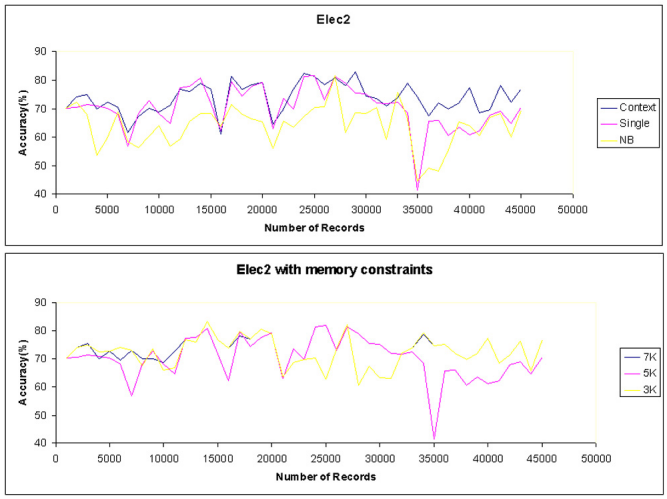


Figure 3. Above(No memory contraints): Comparison of accuracy between the proposed approach(Context), single classifier drift (Single) and incremental NaiveBayes(NB). Bellow: Comparison of accuracy using the proposed approach with different memory values.

In the test scenario with 7Kbytes it is possible to store 7 models, which allow us to keep more than one model for each concept. As a result the performance was very close (with only more 278 misclassified records) to the scenario without memory constraints where 10 models are stored. In the scenario with 5Kbytes, the reduction in accuracy is more significant, with more than 1656 misclassified records and the accuracy curve starts to resemble the *SingleClassifierDrift* seen in 1 especially around records 120000 and 165000 (where the concept with θ=8 is the target concept). Finally in the scenario with 3Kbytes only 3 models can be kept in memory and as a result the performance is further reduced, with more 2881 misclassified records.

E. Results with Electricity Market Dataset

As can be seen in Figure 3, the proposed approach obtained better accuracy results (i.e., 73,4%), which represents a gain of 3,7% and 11% when compared with *SingleClassifierDrift* and incremental NaiveBayes respectively. We can also observe that the proposed approach achieves a more stable accuracy and recovers faster from changes. This can be seen clearly around record 35000.

In relation to the experiments with memory constraints the results show that the overall accuracy is similar between the experiments, in the order of 71% correctly classifier records. For all the tested scenarios the proposed approach still obtains better overall results than the memoryless approach (i.e., *SingleClassifierDrift*) but the accuracy over specific periods depends on the model that is reused and which ones were previously discarded in situations of memory scarcity. This is a direct result of the proposed decision function, and again such difference can be seen around record 35000,

 ISBN: 978-1-61208-100-7

where the tests that kept the adequate model (i.e., 7K and 3K) are able to show improved adaptation.

V. Conclusions and Future Work

In this work, we have proposed a memory-aware approach for the problem of data stream classification that integrates context information with learned models, improving adaptation to drift and exploiting concept recurrence.

Several metrics to define model utility for the challenge of memory adaptation were presented. These were developed in order for the proposed adaptation strategy to suffer minimal loss in accuracy and adaptation to drift when compared to approaches where resources are unbounded and more models can be kept in memory.

We have also tested our approach with the artificial benchmark dataset SEA Concepts and the real world dataset Electricity Market. The experimental results show the advantages of the proposed approach for situations with memory constraints. This is a consequence of minimizing the accuracy loss. As the available memory is reduced while keeping the more relevant models, these are available when the concepts they represent reoccur. We should also note that this is a general approach and better adaptations are expected when using domain sensitive metrics and context. Despite the promising results, we are not exactly sure when drift occurs in the Electricity Market dataset and what really affects change. This limits the depth to which we can evaluate such results. However, such drawbacks are a consequence of learning from real world data.

As future work we plan to test the current approach in a real ubiquitous device with real world problem, and further develop the idea presented in this paper using domain sensitive criteria and context along with more sophisticated adaptation strategies.

Acknowledgments

This research is partially financed by project TIN2008-05924 of Spanish Ministry of Science and Innovation. The work of J.P. Bártolo Gomes is supported by a Phd Grant of the Portuguese Foundation for Science and Technology (FCT).

References

[1] J.P. Bartolo Gomes, E. Menasalvas, and P. Sousa. Tracking Recurrent Concepts Using Context. In *Rough Sets and Current Trends in Computing, Proceedings of the Seventh International Conference RSCTC2010*, pages 168–177. Springer, 2010.

[2] M.M. Gaber, S. Krishnaswamy, and A. Zaslavsky. Ubiquitous data stream mining. In *Current Research and Future Directions Workshop Proceedings held in conjunction with PAKDD*. Citeseer, 2004.

[3] M.M. Gaber and P.S. Yu. A holistic approach for resource-aware adaptive data stream mining. *New Generation Computing*, 25(1):95–115, 2006.

[4] J. Gama and M.M. Gaber. *Learning from data streams: processing techniques in sensor networks*. Springer-Verlag New York Inc, 2007.

[5] J. Gama and P. Kosina. Tracking Recurring Concepts with Meta-learners. In *Progress in Artificial Intelligence: 14th Portuguese Conference on Artificial Intelligence, Epia 2009, Aveiro, Portugal, October 12-15, 2009, Proceedings*, page 423. Springer, 2009.

[6] J. Gama, P. Medas, G. Castillo, and P. Rodrigues. Learning with drift detection. *Lecture Notes in Computer Science*, pages 286–295, 2004.

[7] M. Harries. Splice-2 comparative evaluation: Electricity pricing. Technical report, The University of South Wales, 1999.

[8] M.B. Harries, C. Sammut, and K. Horn. Extracting hidden context. *Machine Learning*, 32(2):101–126, 1998.

[9] G. Holmes, R. Kirkby, and B. Pfahringer. MOA: Massive Online Analysis, 2007 - http://sourceforge.net/projects/moa-datastream/.

[10] I. Katakis, G. Tsoumakas, and I. Vlahavas. Tracking recurring contexts using ensemble classifiers: an application to email filtering. *Knowledge and Information Systems*, pages 1–21.

[11] A. Padovitz, SW Loke, and A. Zaslavsky. Towards a theory of context spaces. In *Pervasive Computing and Communications Workshops, 2004. Proceedings of the Second IEEE Annual Conference on*, pages 38–42, 2004.

[12] W.N. Street and Y.S. Kim. A streaming ensemble algorithm (SEA) for large-scale classification. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 377–382. ACM New York, NY, USA, 2001.

[13] A. Tsymbal. The problem of concept drift: definitions and related work. *Computer Science Department, Trinity College Dublin*, 2004.

[14] G. Widmer and M. Kubat. Learning in the presence of concept drift and hidden contexts. *Machine learning*, 23(1):69–101, 1996.

[15] Y. Yang, X. Wu, and X. Zhu. Combining proactive and reactive predictions for data streams. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, page 715. ACM, 2005.

 ISBN: 978-1-61208-100-7

Modeling and Analysing Ubiquitous Systems Using MDE Approach

Amara Touil*, Jean Vareille*, Fred Lherminier†, and Philippe Le Parc*
*Université Européenne de Bretagne, France
Université de Brest - EA3883 LISyC
Laboratoire d'Informatique des Systèmes Complexes
20 av. Victor Le Gorgeu, BP 809, F-29285 Brest
Email: {amara.touil, jean.vareille, philippe.le-parc}@univ-brest.fr
†Terra Nova Energy
Z.I. de Kergaradec, 28 rue Victor Grignard, F-29490 Guipavas
Email: fredl@terranov.com

***Abstract*—The growth of industrial activities during the last decades and the diversity of industrial products require standards and common methodologies for building and integrating different parts. It is also required that working groups use the same terminologies and concepts needed for each domain. The Model Driven Engineering approach aims to give an answer while using a high level method based on models and transformations. In this paper, we use this approach to model ubiquitous systems. Those systems are composed of devices interconnected through various kinds of network and offer to get and send information. We present a model for this class of system and its use in the field of energy while studying real cases from our industrial partner Terra Nova Energy. This company aims to give solutions to monitor energy use and to reduce consumption. First results, where the use of a Model Driven Engineering approach makes it possible for our partner to improve and to get other points of view of his systems, are presented.**

Keywords - Domain Specific Language, Model Driven Engineering, Analysis, Telecontrol Systems.

I. INTRODUCTION

Ubiquity is often define as being able to be in several places at the same time. In the field of Information Technologies, this definition can be refined in, at least, two ways that may be apparently opposites. The first approach is to considered as in [1] that users are surrounded with "intelligent" systems, that may delivers them needed information: in this case, computing facilities are used to locate users, to understand their environment, to anticipate their needs and to make it possible that all information they require is available anywherere at anytime. The second approach, is to offer people to be able to get information from a system located somewhere and to be able to act safely on it: in this case computing facilities are used to make it possible to be "virtually" present in several places at the same time. We place our work in the second approach in order to model and analyze telecontrol application as a kind of ubiquitous systems.

Terra Nova Energy (TNE) is an innovative company that provides solutions for data mining in the fields of electricity, hydraulic and pulse-energy. It integrates sensing, acting and communicating devices in telecontrol systems, for collecting data from different sites using various technologies. TNE's systems allow real-time operating and provide processes for handling different data treatments.

In order to improve its development, this company is facing two problems: how to conceive remote monitoring systems as automatically as possible and how to fasten their on-site deployment ?

To answer these questions, we propose to use a model based approach, relying on specific components for telecontrol domain, and libraries to integrate industrial parts coming from various providers. Our idea is to build a framework [2] following the Model Driven Engineering (MDE) methodology [3][4] to create a telecontrol ontology and proper transformations for building, generating, analyzing and deploying such ubiquitous telecontrol systems.

Our aim is to define a generic meta-model and to use it for various systems, as case study examples, one of them being the TNE's system.

This paper is organized as follows: first of all, the MDE approach, Domain Specific Languages (DSL) and transformations that may be conducted are briefly introduced. Second, we introduce the generic meta-model proposal. Third, this general meta-model is applied to our case study, while defining an instance for TNE. Then, we explain how to carry a static analyzis on a given instance and show first results. Finally we discuss our method and we finish by further work.

II. RELATED WORKS

In this section, we introduce some basic notions about the MDE approach and DSL, and how to conduct transformations on models.

A. The MDE approach

A model is an abstracted view of a system that expresses related knowledge and information. It is defined by a set of rules, used to interpret the meaning of its components [5][6]. A model is defined according to a modeling language that can give a formal or a semi formal meaning description of a system depending on modeler's intention. Modeling languages can be textual or graphical.

 ISBN: 978-1-61208-100-7

The model paradigm has gained in importance in the field of systems engineering since the nineties. Its breakthrough was favoured by working groups like the Object Management Group (OMG) [7] that has normalized modeling languages such as Unified Modeling Language (UML) and its profiles (such as SysML and MARTE for real-time systems). This group also provides the Model-driven architecture (MDA) software design standard. The MDA is the main initiative for Model Driven Engineering (MDE) approach.

According to OMG, four abstraction levels have to be considered: a meta-meta-model that represents the modeling language, which is able to describe itself; a meta-model level that represents an abstraction of a domain, which is defined through the meta-meta-model; a model level that gives an abstraction of a system as an instance of the meta-model; finally, the last abstraction level is the concrete system.

Tools have been defined to implement OMG standards. Some have general and wide purpose like those for UML language, others have been defined for specific and reduced class of applications like in [8] that aims to specify wireless sensor network systems in order to generate code. Another approach is to use Domain Specific Language (DSL) to specify a special semantic and/or syntactic rules for a class of application. One of DSLs definition is given by [9]: *"A domain-specific language (DSL) is a programming language or executable specification language that offers, through appropriate notations and abstractions, expressive power focused on, and usually restricted to, a particular problem domain"*. In our case we define our own DSL for telecontrol, according to Eclipse Core (Ecore) meta-model [10].

B. Model transformation

In order to breath life in models [11], model transformations aims to exceed the contemplative model view to a more productive approach, towards code generation, analysis and test, simulation, etc. Models are transformed into other models that may be manipulated by specific tools or that may be transformed again into other models. Two kinds of transformation are used, exogenous if source and target transformation do not have the same meta-model and endogenous if source and target have the same meta-model. We use the latter here in order to perform static analysis of the built models.

Generally, static analysis deals with observing system and checking some properties before real system creation, production or code implementation. As example, formal verification and model checking techniques [12][13] are used for verifying models. Behavioural analysis are not beyond in the scope of this article. Static analysis are only conducted in simple cases, but relevant ones for TNE. In this paper we will focus on placement but others studies have been conducted (cost, bandwith,...). As soon as the model is created, it may give answers to the engineer: Are my sensors, repeaters, routers placed in a proper location? Are there better configuration? etc..

Concerning the first question, there are several works treating these points like [14], which study Wireless Network Sensors (WSN) placement for smart highways; it gives a solution based on geometric resolution algorithms in outdoor and open space. In our case, indoor deployment environments, including various fix and mobile obstacles, have to be studied. Component's placement depends on monitored zone, on the chosen topology, on the network capacity, etc. The second question can be adressed by optimyzing the placement configuration.

Next section describes the meta-model proposed for telecontrol and its derivation for TNE's remote monitoring needs.

III. PROPOSED META-MODEL

In this section, we, describe the meta-model we are proposing to specify systems based on communicating objects. Only a high level view and the main concepts are presented. Starting from this meta-model, we derive a specific instance for TNE.

A. Generic meta-model

At a first abstraction level, we built our domain around systems containing entities that communicate with each other as stated in [15]: *"A system is a construct or collection of different elements that together produce results not obtainable by the elements alone"*. In the meta-model proposal (Figure 1), a system, modeled by *System*, can contain other systems according to the *ownedSystem* reference, in order to provide the "system of the system" notion. Regarding entities, they are modeled by *Entity* and its system containment is referenced by the *itsEntity* relationship. Entity paradigm in our meta-model aims to be a generic and general concept formed by extracting common features from our specific telecontrol domain. It can be a logical or physical and can be also composed of other entities (*ownedEntity*).

Moreover, entities are described using several related concepts trying to keep information about their physical properties, communication facilities or behaviour. These concepts are modeled as follows:

- *Structure* element defines the interrelation or arrangement of different physical parts or the organization of elements that provides coherence, shape and rigidity to an entity. It may describe mechanical, chemical or biological aspects.
- Actions performed by entities are described by the *Task* concept. They can be internal, such as making computation, external such as sending information and also connected to the real world such as sensing or acting on a real device.
- *ContactInterface* element allows connection between entities. It can be a physical contact, or a logical interface depending on the specification level, on the refinement degree or on its own structure.
- An entity may contain *Data* related to itself or to its environment.
- *Message* element provides data exchange between entities. To send (or to receive a message) from entity A to entity B, A and B must be connected through *ContactInterfaces*, directly or, it may exist a path of entities that may forward messages.

 ISBN: 978-1-61208-100-7

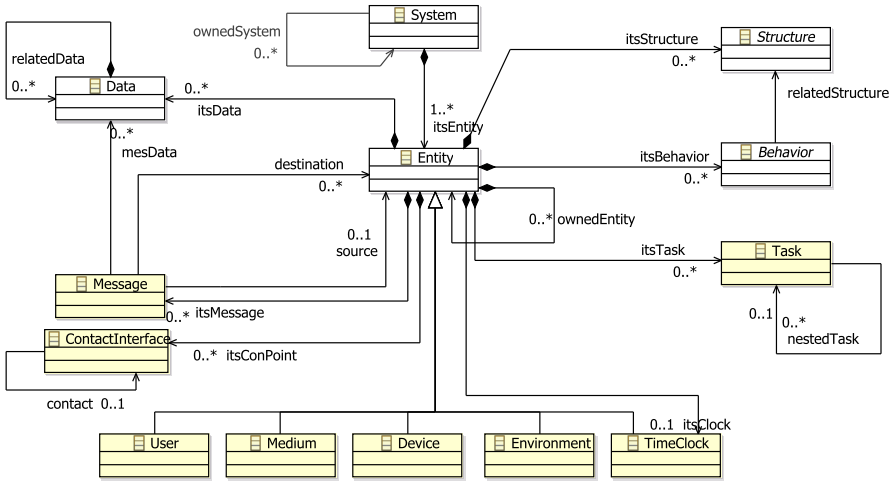


Fig. 1. A view of generic meta-model components (Basic package)

- To each entity, a *Behavior* concept is added, in order to describe the different tasks an entity may realize. From this concept, code generation or simulation may be refined.

As the *Entity* concept is very general, in order to improve our generic meta-model, different sub-entities have been specialized using the concept of heritage:

1) *Device* represents a physical component such as a sensor, a router, a computing unit, etc.
2) *Medium* element represents an entity deployed and used between other entities to specify communication's medium. In classical approaches, medium is often ommitted and considers as perfect links. In our work, medium has its own structure, behavior, contact interfaces, datas, messages and tasks. Different mediums may be described and classified following their own specifications, wired or wireless for example.
3) *User* element is used to model a person interacting with other entities, it is a kind of human avatar. Like other entities *User* can have different presentation depending on model view.
4) The different entities composing the system are under some physical conditions and constraints that affect and influence the global behaviour. In our meta-model we introduce the *Environment* concept to describe such conditions. Because we are dealing with a complex system, the *Environment* is an entity of this system and not just an external element. In our modeling approach, *Environment* acts on internal elements in order to build a global circumstances of evolving system and gives context for ubiquitous entities and systems that are context-aware.
5) The concept of time is crucial for telecontrol systems that is why a *TimeClock* entity is added to our generic meta-model for measuring, recording or indicating time.

The presented generic meta-model intends to specify any kind of systems containing communicating objects. At this abstraction level, we present just a first facet (also named *Basic package*), without getting into all the details of the meta-model.

B. Terra Nova Energy meta-model

In the previous section, a generic meta-model, at a first level of abstraction with some inheriting elements from *entity* such as *Device* has been presented.

According to the concept of inheritance, the Terra Nova Energy meta-model (Figure 2) is built as an extention and a generalization of the generic meta-model. The first element is the *TNESystem* that inherits from the generic element *System*. Other elements have been spread over two sets: generic devices and off the shelf devices that could be grouped in a specialized library.

1) Terra Nova Energy's generic devices: This first set of elements represents generic devices that allow to add a new industrial component to a library or to build a specific system. All these devices inherit from generic *Device*, imported from Basic package as shown in Figure 2. We present them successively as follows:

- *BoxaNova* models the main device of TNE systems. It collects information, sends orders and manages the entire flow of data between different devices and users. A *BoxaNova* device is generally created by some others devices depending on system configuration and deployment.
- *Router* models a device that handles message transfers between different TNE elements. It handles also some other functions like controlling repeaters, sensors, actuators and some other server facilities.
- *Concentrator* is used to model a collecting place of data coming from repeaters, sensors, actuators before sending them to routers depending on requests or preprogrammed sending tasks.
- *Sensor_ActuatorDevice* is used to capture measurement data and to send them or to act pursuant to an order.

 ISBN: 978-1-61208-100-7

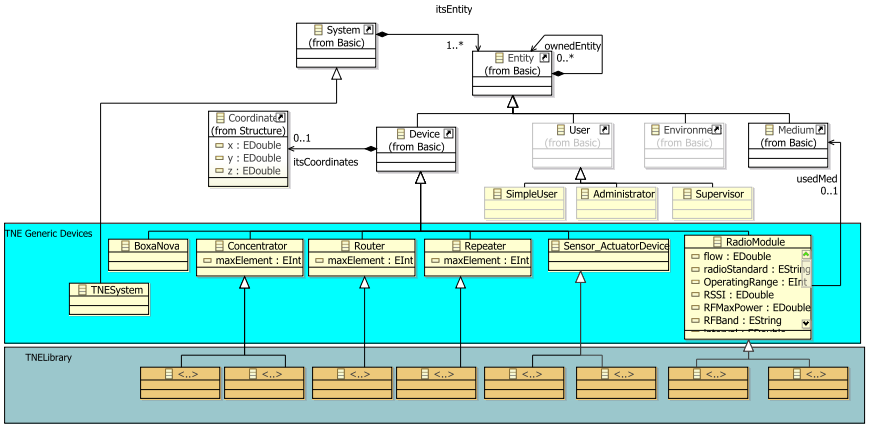


Fig. 2. A view of Terra Nova Energy meta-model

- *Repeater* is used to ensure the link between a *Concentrator* and *Sensor_ActuatorDevice* elements if they are not within reach.
- *RadioModule* models the device that ensures exchanging data messages between all other devices when wireless communication links are used.

Since TNE devices are specialized from the generic element *Devices*, they inherit all properties and references.

2) Terra Nova Energy's off the shelf devices: This second set of elements represents real devices ready to use in the framework. They are another level of specialization of the first set. They may be grouped in a kind of library (blue box in Figure 2), composed with industrial devices coming from different manufacturers and in order to be integrated into TNE systems. In this paper we neither describe this library of devices nor the manner of their built because we focus on some other properties and aspects of model analysis.

In the next section, we propose a methodology of analysis showing the usefulness of these presented meta-models, and their examination and validation on real industrial cases.

IV. MODEL ANALYSIS

In this section, we present how to exploit the MDE approach to analyse placement for a given instance, and automatic placement of repeaters and concentrators for an uncompleted system, where only sensor, actuator positions are known.

The first part describes our methodological approach, based on processing properties and applied constraints for model's elements.

A. Presentation

From the TNE domain specific meta-model, presented in the previous section, an instance that describes a real system of telecontrol may be defined. According to the user needs, this instance captures a particular concern for a system and gives the necessary elements and their relevant properties. Such a model requires some processing stages to meet the final requirements in terms of analysis. These processing stages are done in response to expected system constraints applied to its various elements. Figure 3 shows a model overview of the various elements used to carry out those treatments for analysis methodology purpose.

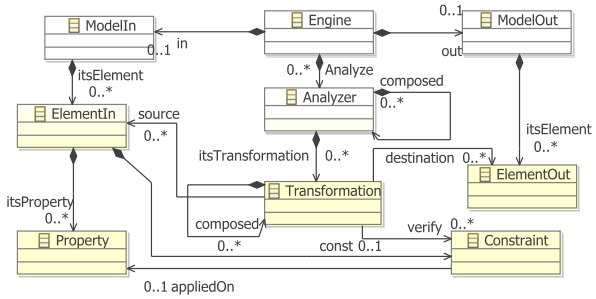


Fig. 3. A view of analyzer's engine model

The first analysis component (*ModelIn*) loads the instance that describes a telecontrol system according to its meta-model. This instance is composed of elements called *ElementIn*. The *Engine* scans iteratively the input model, element by element, to find different properties and constraints. For each property the *Analyzer* checks if it fits well the associated constraint. When properties are verified, the element will be transformed by *Transformation* and treated by *ElementOut*. The *ModelOut* generates automatically output model formed by those transformed elements. Output model should be consistent to its meta-model that may be the input one or another depending on performed transformation.

Generally, input models have a tree shape structure and their components can be composed of other components. Properties and applied constraints can also have this composition criterion. In order to deal with this kind of structure and having a generic analyzer, possible compositions have been taken into account for analyzing and transforming tasks. In Figure 3, these composition aspects are represented by the *Composed* reference. The next part of this section deals with an application of this methodology to a system. First of all, analysis for

 ISBN: 978-1-61208-100-7

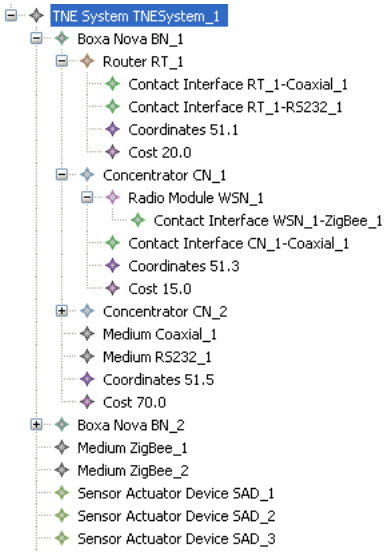


Fig. 4. A screen shot of manual instance creation

a completely described given instance are conducted to verify that elements are well placed and may communicate together properly. Second we tell that the Model Driven Engineering approach used, may also be helpful to build instances, starting with an uncomplete system, where only sensors and actuators are known, to add necessary repeaters and concentrators.

B. First case study with a complete TNE's instance

In this first case study, an instance of a TNE meta-model has been created with a tree editor as shown in Figure 4. It is an abstraction of a *TNESystem* named TelecontrolSystem_1 composed of two *BoxaNova* that manage twenty *Sensor_ActuatorDevices*. Each *BoxaNova* consists of two *Concentrator* elements and a *Router*. Communication between these components uses two *Mediums*; a coaxial cable that connects the first concentrator and the router and a RS232 cable that provides serial communication between the second concentrator and the router.

Sensor and actuator devices use a ZigBee protocol (IEEE802.15.4 standard) [16] to communicate with the *BoxaNova* concentrators. To achieve this task, they use *radioModules*. Concentrators also have their own radio modules. In general, wireless propagation environments in TNE systems may have different attenuations. Thus, each communication between a sensor/actuator and its repeater or concentrator is defined by a medium with relevant properties and necessary (contact) interfaces. In our model, this information is kept by properties such as the attenuation value, speed of transmission, signal length, etc. Connection with the medium and other elements is ensured by the definition of two entities as *ContactInterface*; one for the medium and the other for the associated element.

The model also captures other information like the physical location (x, y coordinates) of the different parts of the system. In this way, *Communication vs. placement* analysis have been conducted. In fact, component's placement in TNE sytems is crucial to ensure good communication and proper information transfer. As the model is fully known, an analyzer can be created to verify that every *Sensor_ActuatorDevice* element according to its coordinates may communicate to its corresponding concentrator. This verification takes the attenuation between communicating elements into account, specified as a property of the medium, which links them. Another analyzer can be defined and used for inspecting the number of sensor/actuator devices for each concentrator and its compliance to specified *maxElement* property.

From information analysis text file obtained (see transformation model to text in [11]), manual corrections may be realized. This task is manageable for small systems but gets difficult and very expensive for large ones with many components. The next part gives an illustration of a methodology that can treat this difficulty.

C. Second case study with a partial TNE's instance

In this second case study, only the number and the position of sensor/actuator devices are given. The instance is partial, and the objective is to automatically find a solution for repeater's and concentrator's placement. This problem is a very complex one and several solutions may be found in literature [14][16][17].

In our case an input model is generated automatically with the required properties. As a first step, an instance of the TNE meta-model is created with a system *TNESystem*, a number of *sensor_ActuatorDevice* and an environment that represents a factory hall. In our case, we suppose that radio measurements were made in the environment and have identified attenuation values.

In a second step, an analysis engine for placement, composed of three analysers, is used:

1) The primary analyzer performs an initial refinement of the input model. First, it divides the environment, specified in the input model, into zones where mediums have the same attenuation so that the range of sensor-actuator radio modules, who are currently in, reach the middle. This range is calculated by applying attenuation value imposed in the input model. Then, depending on the number of sensor/actuator devices, it creates the necessary repeaters with their radio modules, and adds them to the input model with coordinates in the vicinity of the centre of the sub-environment. Finally, it creates connections (*ContactInterface*) between repeaters and sensor/actuator elements and adds them to the model.
2) A second pass of refinement with the same treatment approach is made by the second analyzer. Its objective is to connect repeaters with concentrators. The input model environment is split up into several zones with new radio modules ranges of repeaters. The number of added concentrators will depend on their capacity and also on the number of repeaters within reach.
3) The third analyzer performs an optimization on the obtained model that contains sensor/actuator devices that

 ISBN: 978-1-61208-100-7

can communicate directly with concentrators without the use of a repeater. This analyzer checks for each sensor/actuator device if it reaches any concentrator and changes its connection from a repeater to a concentrator. Repeaters without connected sensor/actuator devices will be removed.

At the end of these three analyzer passes, a new output model is obtained. It is composed of initial sensor/actuator devices and added components (repeaters, concentrators and connections). Figure 5 shows a simple placement layout of TNE system components obtained with the following initial model inputs: 200 sensors/actuator devices (location of these elements has been chosen randomly for the test), an environment that represents a factory with 800 meters length and 600 meters width and an attenuation value of 5 decibels by meter. Red rectangles represent repeaters, blue triangles represent concentrators and green circles represent sensor/actuator devices.

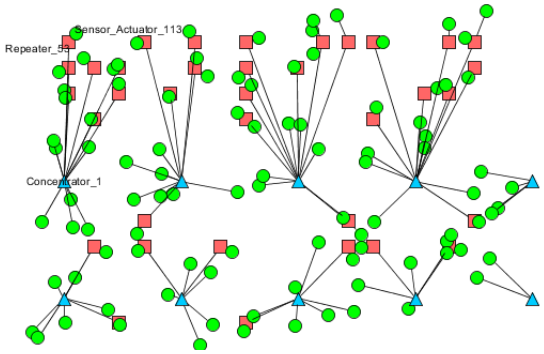


Fig. 5. Placement layout

The generated model allows us to have a first proposal for the positions of the repeaters and of the concentrators in the environment. This obtained placement is of course not optimal, as simple hypothesis have been used. Indeed, the presence of mobile obstacles in the environment and their influence on signal propagation should be considered. To optimize the placement and to ensure the highest data rate optimal algorithms should also be used. The orientation of elements should also be taken into account.

Nevertheless, our objective is to show that using model analysis in ubiquitous systems in MDE approach may support engineers in the design of their systems and verifying some properties in earliest stage without real implementation such as communicating object placement.

V. CONCLUSION AND FUTURE WORKS

In this paper, a generic meta-model for modeling ubiquitous telecontrol systems and, specially telecontrol systems is proposed. A variant of this metamodel is specified to fit the needs of Terra Nova Energy.

The introduction of the Model Driven Engineering approach in this field allows us to exceed the contemplative dimension of models towards productive models. The presented cases studies highlights this approach by proposing a possible model for placement analysis for an example of the TNE system, using transformations.

In the future, we would like to improve our generic meta-model in order to take the behaviour of the different components into account. We would like to move from static analysis to simulation with a complete integration of a system inside its environment.

ACKNOWLEDGMENT

This work is supported by the city of Brest, "Brest Métropole Océane", and performed in partnership with Terra-Nova Energy. We thank them for their help.

REFERENCES

[1] M. Weiser, "The computer for the 21st century," *Scientific American Special Issue on Communications, Computers, and Networks*, 1991.

[2] P. Le Parc, A. Touil, and J. Vareille, "A model-driven approach for building ubiquitous applications," in *The Third International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies - UBICOMM 2009*, Oct. 2009.

[3] S. Kent, "Model driven engineering," *Lecture notes in computer science*, pp. 286–298, 2002.

[4] F. Fleurey, J. Steel, and B. Baudry, "Validation in model-driven engineering: testing model transformations," in *Workshop WS5 at the 7th International Conference on the UML, Lisbon, Portugal*, 2004.

[5] S. Gerard, F. Terrier, and Y. Tanguy, "Using the model paradigm for real-time systems development: Accord/uml," *Lecture notes in computer science*, vol. 2426, pp. 260 – 269, 2002.

[6] D. Moody, "Graphical Entity Relationship Models: Towards a More User Understandable Representation of Data," *Lecture Notes in Computer Science*, vol. 1157, pp. 227–244, 1996.

[7] S. Cranefield and M. Purvis, "UML as an ontology modelling language," in *Proceedings of the Workshop on Intelligent Information Integration, 16th International Joint Conference on Artificial Intelligence (IJCAI-99)*, vol. 212, 1999.

[8] F. Losilla, C. Vecente-Chicote, B. Alvarez, A. Iborra, and P. Sánchez, "Wireless sensor network application development: An architecture-centric mde approach," *Lecture Notes in Computer Science*, vol. 4758, p. 179, 2007.

[9] A. van Deursen, P. Klint, and J. Visser, "Domain-specific languages: an annotated bibliography," *SIGPLAN Notices*, vol. 35, no. 6, pp. 26–36, June 2000.

[10] R. Gronback, "Eclipse Modeling Project: A Domain-Specific Language (DSL) Toolkit," *Addison-Wesley Professional*, 2009.

[11] K. Czarnecki and S. Helsen, "Classification of model transformation approaches," in *Proceedings of the 2nd OOPSLA Workshop on Generative Techniques in the Context of the Model Driven Architecture*, 2003.

[12] T. Alenljung and B. Lennartson, "Formal Verification of PLC Controlled Systems Using Sensor Graphs," in *Proceedings of the fifth annual IEEE international conference on Automation science and engineering*. The Institute of Electrical and Electronics Engineers Inc., 2009, pp. 164–170.

[13] A. Voronov and K. AAkesson, "Verification of process operations using model checking," in *CASE'09: Proceedings of the fifth annual IEEE international conference on Automation science and engineering*. The Institute of Electrical and Electronics Engineers Inc., 2009, pp. 415–420.

[14] S. Ghosh and S. Rao, "Sensor network design for smart highways," in *CASE'09: Proceedings of the fifth annual IEEE international conference on Automation science and engineering*. The Institute of Electrical and Electronics Engineers Inc., 2009, pp. 353–360.

[15] I. C. Committee, "A consensus of the incose fellows," International Council On Systems Engineering, Tech. Rep., 2006. [Online]. Available: http://www.incose.org/practice/fellowsconsensus.aspx

[16] P. Baronti, P. Pillai, V. Chook, S. Chessa, A. Gotta, and Y. Hu, "Wireless sensor networks: A survey on the state of the art and the 802.15. 4 and ZigBee standards," *Computer Communications*, vol. 30, no. 7, pp. 1655–1695, 2007.

[17] X. Cheng, D. Du, L. Wang, and B. Xu, "Relay sensor placement in wireless sensor networks," *Wireless Networks*, vol. 14, no. 3, pp. 347–355, 2008.

 ISBN: 978-1-61208-100-7

Information Dissemination in WSNs Applied to Physical Phenomena Tracking

María Ángeles Serna, Eva María García, Aurelio Bermúdez, Rafael Casado
Instituto de Investigación en Informática de Albacete (I^3A)
Universidad de Castilla-La Mancha
02071 Albacete, Spain
{angeles.serna, evamaria.garcia, aurelio.bermudez, rafael.casado}@uclm.es

***Abstract*— Several works suggest the use of wireless sensor networks to manage crisis situations. An example of such scenarios is forest fire fighting. In these situations, it is necessary to use efficient mechanisms for disseminating the environmental information captured by sensors. This paper analyzes the behavior of some representative broadcasting techniques when they are used to disseminate to the entire network the occurrence of multiple simultaneous events. The performance evaluation carried out shows that a moderate resource consumption in network devices can reduce the operation overhead.**

Keywords-wireless sensor networks; dissemination; performance evaluation

I. INTRODUCTION

One of the most promising applications of *wireless sensor networks* (WSNs) is its application to what is recently being referred to as *situation management* [10]. Situation management deals with dynamic and unpredictable scenarios, where a complex distributed system deployed over a wide area captures real time data from a large number of heterogeneous information sources. The final goal of this system is to provide support for decisions making.

An example of these applications is the EIDOS system (*EquIpment Destined for Orientation and Safety*) [4], in which a large (and dense) WSN is deployed over the area affected by a wildfire by using aerial vehicles. The mission of the network in this case is to directly provide the fire-fighters critical information that can improve their safety and efficiency. Basically, the WSN consists of a set of devices (also called "motes") that are able to capture certain physical magnitudes (temperature, pressure, humidity, etc.) in the environment where they have been deployed, and process and transmit the data acquired through a RF channel.

The most important technical objective in the EIDOS system is the development of a distributed collaborative processing mechanism. Starting from the sensed data, this mechanism allows the network nodes to obtain a simplified representation of the active fire fronts (a fire model), which informs about their localization, shape, speed, etc. This mechanism necessarily relies on an efficient data dissemination technique for propagating the information captured by the sensors.

In this sense, several proposals can be found in the literature to perform the information dissemination in WSNs. In most cases, these techniques have been designed to propagate the occurrence of a single event of interest. However, in phenomena tracking applications (such as the one described), it may occur that, at a given time, several nodes in a dense WSN detect and spread multiple events simultaneously (or almost simultaneously). In our case, every individual event could mean the approach of a fire front to a given location.

In this paper we analyze the behavior of different dissemination mechanisms when they are employed in physical phenomena monitoring tasks. In particular we are interested in evaluating their efficiency in spreading multiple events, the amount of resources required in network devices, and the overhead introduced in the wireless shared medium.

The rest of this paper is organized as follows. First, Section 2 presents some related work in the area of information dissemination in sensor networks. Then, Section 3 details the techniques that we have selected for this study. After that, Section 4 shows some simulation results that allow us to analyze the selected mechanisms. Lastly, Section 5 presents the conclusions of our research and outlines the future work.

II. WSN DISSEMINATION TECHNIQUES

Dissemination (or *diffusion*) refers to the way in which the information is routed from the place where it is obtained (the sensor nodes) to the nodes who are interested in it [11]. In many situations the information flows to a single node, which plays the role of network base station. In this case we use the term *collection* or *gathering*. On the other hand, when all the nodes in the network are interested in receiving the information being disseminated (as in our case) the term *broadcast* is used. In [1] and [19], two classifications for these algorithms can be found. The simplest broadcast mechanism is *blind flooding*. Here, each node which owns (or receives) the information to be disseminated transmits it to the medium, so that it can be heard by all its direct neighbors. Thus, this technique ensures, at least in theory, that the information will reach its destination.

However, it is well known that an uncontrolled flooding can be very inefficient, mainly because a given node can receive the same information from multiple neighbors [15]. This phenomenon is known as the *broadcast storm problem*.

 ISBN: 978-1-61208-100-7

Some *probabilistic-based* techniques, such as *Gossiping* [6], apply probability to control this redundancy. An alternative is that nodes do not rebroadcast the information if they received the same data more than a predefined number of times (*counter-based* techniques) [1]. Another option is to discard a broadcast message if the distance to the transmitter node is less than a threshold, in order that only the furthest nodes from the transmitter retransmits, in turn, the data (*distance-based* techniques). Another possibility is to retransmit if and only if the additional area that would be covered after forwarding the message is large enough (*area-based* techniques) [13].

In many cases, the decision of rebroadcasting the received information is postponed to a later time. This waiting period is usually called RAD (*random assessment delay*) [19]. In this way, if the node receives a new copy of the same message before the RAD ends, the rebroadcasting is canceled, thus avoiding redundant transmissions. The delay can be established randomly, or depending on the distance to the sending node (the greater the distance, the shorter the delay) [20]. The delay can also be calculated according to the area covered by all the copies of the same message received [7], or depending on the perimeter covered by these copies [16]. In the latter two cases, the transmission is canceled when it is no longer necessary.

Other techniques assume a priori knowledge of the geographical location of direct neighbors (*location-based* techniques). Then, information is transmitted if and only if the additional area that will be covered is greater than a certain threshold [15]. Another way to control flooding is to use the hierarchy established by *clustering* algorithms, so that the *cluster head* in each group is the only node that forwards the message (*cluster-based* techniques) [15]. Other proposals based on one- or two-hop neighbor knowledge (*neighbor-knowledge-based* techniques) can be found in [17], [18], and [19]. Many authors have also proposed to use hybrid schemes. Some examples are [9] and [12]. In large networks the techniques introduced in this paragraph involve a considerable control overhead, making them less suitable.

Finally, some dissemination mechanisms assume that only a subset of network nodes is interested in receiving the information. For example, in the *directed diffusion* mechanism [8], an initial phase is executed before the information is sent, in which a node propagates to the rest of the network its interest in a particular event. Then, the node that can provide that information answers with the requested data, using the best possible route.

III. MECHANISMS ANALYZED

The EIDOS system described in the introduction section requires the propagation of a large amount of fire detection events to all nodes in a large (and dense) WSN. Therefore, among the discussed dissemination techniques, we have considered the simplest one (flooding), and three delayed mechanisms (random delay, distance-based delay, and area-based delay). In this section we detail the implementation carried out for each of the chosen mechanisms. Before this, some common assumptions are presented.

We assume that every node in the network knows its location, which can be obtained either through a built-in *Global Positioning System* (GPS) receiver, or through a previous localization process (outside the scope of this paper). This location must be broadcasted to the entire network when the node sensor detects an approaching fire front. In addition, all nodes receiving the broadcast message will store the initiator position in some internal data structure, along with a time reference.

To minimize control overhead, we also assume that network nodes neither maintain any hierarchy nor have information about the amount of neighbors or their location.

Finally, regarding to the radio, we assume the use of omni-directional ideal antennas, resulting in circular coverage areas. In all the cases it is used the same transmission power, so these areas will have the same size.

A. Instantaneous Dissemination

This is a classic blind flooding mechanism [15]. When a node receives a message for the first time, it instantaneously retransmits to all its neighbors and, in turn, they transmit the message immediately. The only restriction that we have imposed is not to allow more than one retransmission of the same message by the same node.

The Algorithm 1 shows the actions carried out by a network node after receiving a message (M).

Algorithm 1. *Instantaneous dissemination*

```
1: if first copy of M then
2:     forward M
3: end if
```

B. Random Delay

This is an improvement over the instantaneous dissemination mechanism, in which the received message is not transmitted immediately, but after a waiting period [19]. The duration of this period (d) is randomly chosen between 0 and a predetermined value (d_{max}). In order to avoid redundant transmissions, if a node receives a new copy of the message before the waiting period ends, the retransmission is canceled. Moreover, as in the previous case, a node only retransmits once the same message.

The pseudo-code shown in Algorithm 2 shows the behavior of a node upon the reception of a message (M).

Algorithm 2. *Dissemination by random delay*

```
1: if first copy of M then
2:    d = random(0, d_max)
3:    forward M in t_current + d
4: else
5:    if transmission of M is still pending then
6:        cancel transmission
7:    end if
8: end if
```

Note that, at a given time, a node may have several pending messages to send, which forces it to maintain a list to store information related to its pending retransmissions.

C. Distance-Based Delay

This mechanism also introduces a waiting period (from 0 to d_{max} seconds) before forwarding the information received. However, in this case the delay is set according to the distance to the sending node [20]. Specifically, the waiting period is inversely proportional to this distance, so that remote nodes will forward the message first, thus covering a larger area.

As before, the reception of a new copy of the message before the waiting period ends results in the cancellation of the retransmission. This is because it is no longer necessary that this node forwards the information due to another node further from the original one has done it before. Furthermore, no node forwards the same message more than once.

In the example of Fig. 1, node A starts the event dissemination. Then nodes B, F, G, H, and I (located on the perimeter of the coverage area) forward the information. The rest of nodes under coverage decide to cancel their own broadcasts after the reception of a second copy of the message.

Note that nodes do not have information about the location of their neighbors, so this technique requires that the sensor nodes incorporate some mechanism that allows them to measure (or estimate) the distance to the message sender. One possibility would be to use the power of the received message, assuming that it will be lower at a greater distance to the sender.

Thus, when a node receives the first copy of a message, it could set the length of the waiting period by using the following expression:

$$d = \left(\frac{P_R - P_{min}}{P_{max} - P_{min}} \right) x\, d_{max} \qquad (1)$$

where P_R is the receive signal power, which varies between P_{max} (transmission power) and P_{min} (minimum reception power).

The Algorithm 3 describes the actions performed by a node after the arrival of a message (M).

Algorithm 3. *Dissemination by distance-based delay*

```
1: if first copy of M then
2:   d = [ (P_r – P_min) / (P_max – P_min) ] × d_max
3:   forward M in t_current + d
4: else
5:   if transmission of M is still pending then
6:       cancel transmission
7:   end if
8: end if
```

In the same that way the algorithm that uses random delays, the implementation of this technique requires managing a list of pending retransmissions.

D. Area-Based Delay

Among the techniques proposed in this category, we have implemented the ABBA algorithm [16]. As in the two previous cases, after the arrival of the first copy of a message, the receiving node establishes a waiting period. However, the copies received before the period ends do not cancel the message broadcast, but serve to adjust its length. In particular, the delay is set according to the coverage area of the receiving node that has already been covered by all the copies of the same message. Thus, the bigger area covered, the longer waiting periods.

Assuming an ideal case where coverage areas are circular, we can compute the area covered by obtaining the covered arc. For example, in Fig. 2, node C has received a message from node A. The perimeter portion of C covered by the transmission of A is given by the intersection of two circles, and can be expressed through the difference between the initial and final angles. Moreover, the fact of using the same radio circles ensures that at most two segments will remain uncovered within that perimeter, which reduces the amount of information to store at each node.

Note that to allow the updating of the perimeter covered at a receiving node, the received message needs to explicitly include the transmitter position. Obviously, this introduces an additional communication overhead.

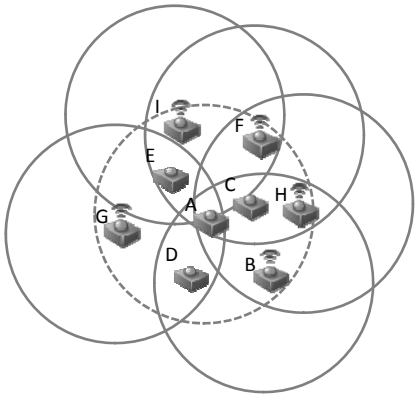


Figure 1. Example of dissemination through the distance-based delay.

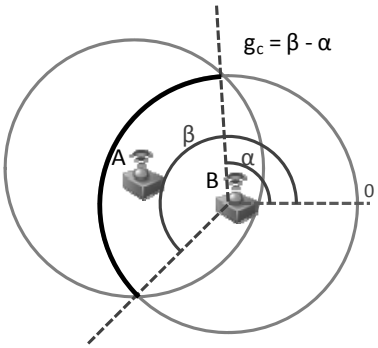


Figure 2. Perimeter of node C covered by a message received from A.

 ISBN: 978-1-61208-100-7

Upon receiving a message, and after updating the perimeter portion which has been covered by previous copies, the node establishes a temporary delay for forwarding the message, according to the following expression:

$$d = \left(\frac{g_c}{360} \right) \times d_{max} \quad (2)$$

where g_c is the angle (in degrees) covered. However, the forwarding of a message is cancelled if, before the expiration of the corresponding delay, the whole perimeter has been covered by other transmissions.

The actions to be carried out after receiving a message (M) are detailed in Algorithm 4.

Algorithm 4. *Dissemination by area-based delay*

```
1:  if M has not been forwarded yet then
2:      compute g_c for M
3:      if g_c has changed then
4:          if g_c = 360º then
5:              cancel transmission
6:          else
7:              d = (g_c / 360) × d_max
8:              forward M in t_current + d
9:          end if
10:     end if
11: end if
```

Obviously, this algorithm requires that each node maintains a list to store the messages waiting to be broadcasted, along with the perimeter not covered yet by previous copies if those messages.

IV. PERFORMANCE EVALUATION

This section analyzes the performance of the four dissemination techniques detailed, from the point of view of their efficiency, resource consumption associated, and overhead introduced.

A. Simulation Environment and Methodology

We have used a simulation environment [5] developed for the EIDOS system. This tool is composed of several independent and interconnected modules, which share information by means of a global MySQL database. In short, first we use FARSITE [3] to simulate a wildfire over a particular area, by using real geographical, environmental and vegetation data. After that, a WSN simulator (developed in Python/TOSSIM [14]) executes the EIDOS application in each network node.

In order to obtain realistic results, the simulator incorporates a noise and interference model and the Friis *free-space* signal propagation model. We have modeled the *Crossbow Iris* radio [2], applying a transmission power of 3 dBm and a minimum reception power of -90 dBm. Under these conditions, we obtain an approximate radio range of 87 meters. The simulated protocol for media access control is basic CSMA [14].

In each simulation run, network nodes were distributed randomly in a square area of 400 × 400 meters. We have considered network sizes of 100, 300, 500, 600, 800 and 1000 nodes, with an associated density or connectivity degree (average number of direct neighbors) of 11.9, 35.77, 62.93, 71.87, 97 805 and 124.09, respectively.

For each experiment, we simulate the spreading of a forest fire in the deployment area, so that the fire reaches all the nodes of the network (without burning them). Every time a node detects the proximity of a fire (by a sudden rise in the sensed temperature), it broadcasts its position to the entire network. For localization purposes, we have assumed that all network nodes are equipped with a GPS receiver. For the execution of random-, distance- and area-based mechanisms the value of d_{max} parameter has been set to 5 seconds.

Finally, in order to increase the representativeness of the results shown, each experiment was repeated several times for each of the dissemination techniques studied, showing here the average values.

B. Simulation Results

First of all, Fig. 3 (a) shows the efficiency of the dissemination mechanisms analyzed, expressed as the average number of events received by each node, in function of the network degree. As the upper bound of this statistic matches the network size, Fig. 3 (b) shows the same results, but normalized according to that bound.

Note that, in general, an increment in network density penalizes all mechanisms. In particular, the instantaneous algorithm is quite efficient at a low network density, but its benefits are quickly reduced as the density increases. The random algorithm has a lower efficiency in a low density network, but it is not as sensitive to an increment in density

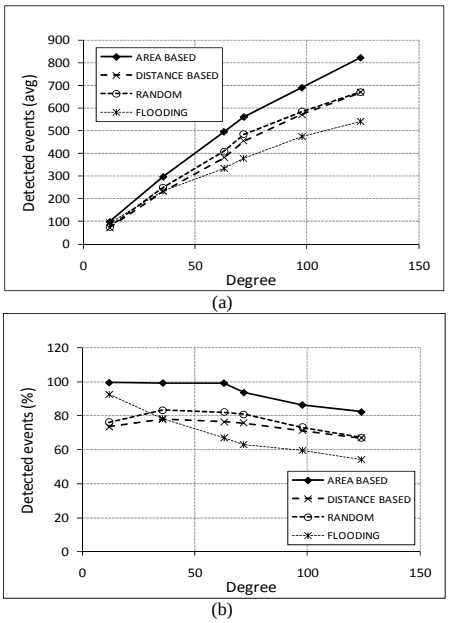


Figure 3. Efficiency of the dissemination mechanisms.

 ISBN: 978-1-61208-100-7

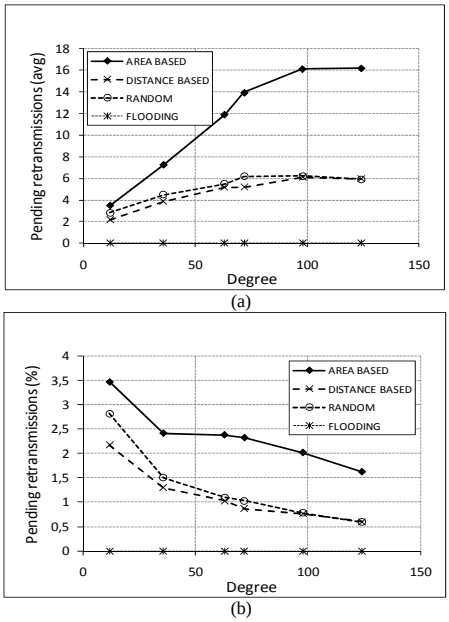


Figure 4. Amount of pending retransmissions at each node.

as the previous algorithm is. The distance-based algorithm behaves like the last one, although slightly worse because it produces more collisions (this issue is discussed later). Finally, the area-based algorithm obtains the best performance, being optimal at low and medium densities.

Next, Fig. 4 shows the amount of events pending to be forwarded at each node, for the different evaluated mechanisms, and in function of the network degree. Note that the storage of these pending retransmissions consumes memory resources at the devices. Fig. 4 (a) shows average values, and Fig. 4 (b) shows the same results, but normalized according to the number of events propagated during the complete simulation run.

Obviously, the instantaneous algorithm does not consume any resource at the nodes. In the remaining algorithms, the total amount of outstanding retransmissions tends to stabilize as the network degree increases (a), thus the relative percentage tends to decrease (b). It may be noted that the random- and distance-based algorithms have a similar behavior, while the area-based has slightly higher requirements.

Finally, Fig. 5 shows the effect of the dissemination mechanisms on the wireless medium, through the amount of sent messages, duplicates and collisions per node, and in function of the network degree. It can be seen that the instantaneous algorithm generates a high overhead in the channel (a), producing a large number of collisions (b). This leads to a very few duplicated messages (c), and a low redundancy in the information transmitted (d). This algorithm forwards (once) all the information received, getting the worst possible efficiency of the performed transmissions (e).

The random algorithm is the one which transmits less information, keeping a low level of collisions and duplicated messages. With a moderate redundancy, it gets the highest efficiency of the information transmitted in low density networks.

The distance-based algorithm behaves like the former one, except that it has a slightly higher number of collisions. This is because the random distribution in time of the broadcasts is uniform, while the spatial distribution tends to increase with distance. The consequence is that the distance-based algorithm gets a slightly lower efficiency.

Finally, in the area-based algorithm, the amount of messages sent is not so sensitive to the network density, being the best starting from certain degree. Moreover, the number of collisions is negligible (b), so that the amount of duplicated messages is higher than the obtained by other algorithms (c), obtaining a redundancy that increases linearly (d). Finally, this algorithm presents the best efficiency in the transmissions made for dense networks (e).

V. CONCLUSIONS AND FUTURE WORK

In this work, we have studied the behavior of some representative dissemination techniques when they are used to broadcast multiple events in a dense WSN during the task of monitoring a physical phenomenon.

From the analysis carried out it is shown that the technique that uses transmission delays based on area is the most efficient, in terms of the amount of events that it spreads. Moreover, from the point of view of the media access level, this mechanism exhibits the best behavior, at the expense of certain resource consumption in network devices. However, we found that the use of a reliable broadcast on a not reliable access level does not guarantee the propagation of all the events.

As future work, we plan to consider using fusion techniques in order to reduce the amount of information to propagate. Our goal is to define some kind of representation which allows grouping a large number of geo-referenced events in a single data structure to spread.

ACKNOWLEDGMENT

This work was supported by the Spanish MEC and MICINN, as well as European Commission FEDER funds, under Grants CSD2006-00046 and TIN2009-14475-C04. It was also partly supported by the JCCM under Grants PREG-07-25, PII1C09-0101-9476, and PII2I09-0067-3628.

REFERENCES

[1] Z.-Y. Cao, Z.-Z. Ji, and M.-Z. Hu, “An energy-aware broadcast scheme for directed diffusion in wireless sensor network,” Journal of Communication and Computer, vol. 4(5), pp. 28–35, 2007.

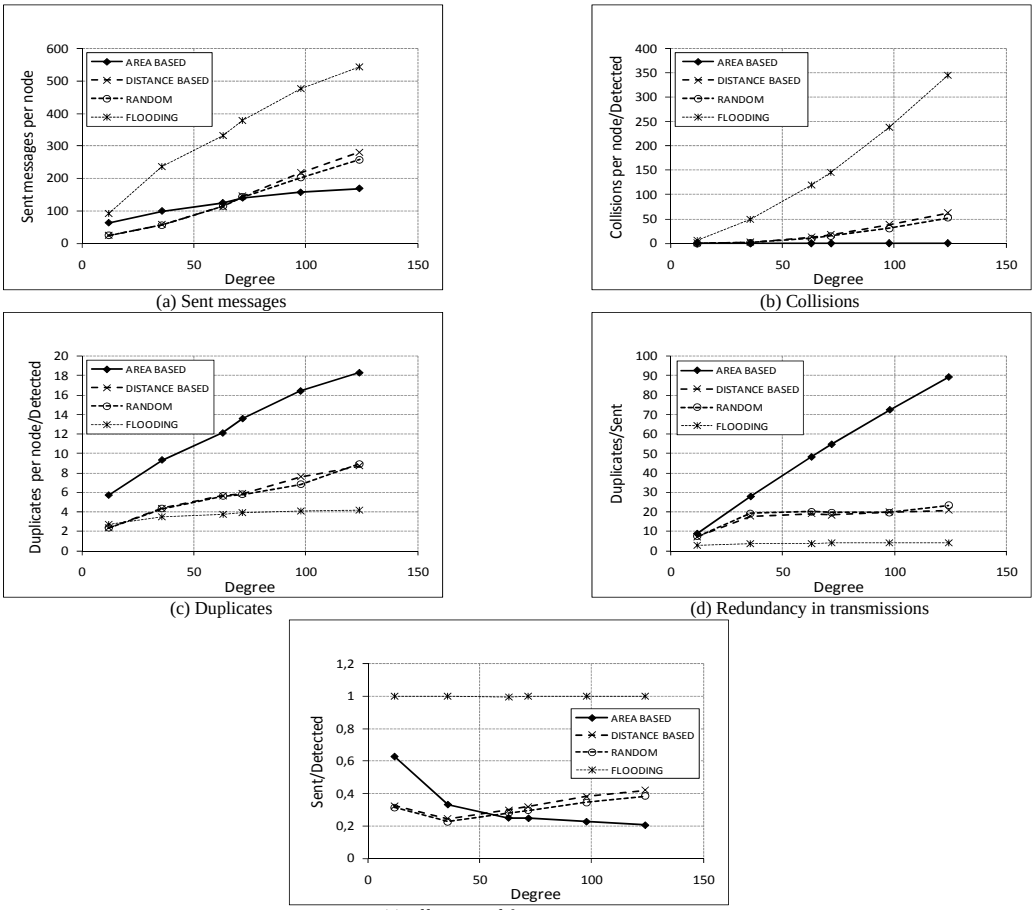
[2] Crossbow Technology, Inc. http://www.xbow.com, 2010.

[3] Fire.org, http://fire.org/, 2010

[4] E. M. García, A. Bermúdez, R. Casado, and F. J. Quiles, “Collaborative data processing for forest fire fighting,” In 4th European Conference on Wireless Sensor Networks, Adjunct poster/demo proceedings, pp. 3–5, 2007.

[5] E. M. García, M. A. Serna, A. Bermúdez, and R. Casado, “Simulating a WSN-based Wildfire Fighting Support System,” In IEEE International Symposium on Parallel and Distributed Processing with Applications, pp. 896–902, 2008.

 ISBN: 978-1-61208-100-7



(a) Sent messages

(b) Collisions

(c) Duplicates

(d) Redundancy in transmissions

(e) Efficiency of the transmissions

Figure 5. Channel overhead caused by the dissemination mechanisms.

[6] Z. J. Haas, J. Y. Halpern, and L. Li, "Gossip-Based Ad Hoc Routing," IEEE/ACM Transactions on Networking, vol. 14(3), pp. 479–491, 2006.

[7] M. Heissenbüttel, T. Braun, M. Wälchli, and T. Bernoulli, "Optimized stateless broadcasting in wireless multi-hop networks," In IEEE Infocom, 2006.

[8] C. Intanagonwiwat, R. Govindan, and D. Estrin, "Directed Diffusion: A Scalable and Robust Communication Paradigm for Sensor Networks," In 6th Annual International Conference on Mobile Computing and Networks, pp. 56–67, 2000.

[9] S. Izumi, T. Matsuda, H. Kawaguchi, C. Ohta, and M. Yoshimoto, "Improvement of Counter-based Broadcasting by Random Assessment Delay Extension for Wireless Sensor Networks," In International Conference on Sensor Technologies and Applications, pp. 76–81, 2007.

[10] G. Jakobson, J. F. Buford, and L. Lewis, "Guest Editorial: Situation Management," IEEE Communications Magazine, vol. 48(3), p. 110, March 2010.

[11] H. Karl and A. Willig, "Protocols and Architectures for Wireless Sensor Networks," Wiley, 2005.

[12] K.-W. Kim, K.-K. Kim, C.-M. Han, M. M.-O. Lee, and Y.-K. Kim, "An Enhanced Broadcasting Algorithm in Wireless Ad-hoc Networks," In International Conference on Information Science and Security, pp. 159–163, 2008.

[13] J. Kim, Q. Zhang, and D. P. Agrawal, "Probabilistic broadcasting based on coverage area and neighbor confirmation in mobile ad hoc networks," In IEEE Global Telecommunications Conference Workshops, pp. 96–101, 2004.

[14] P. Levis, N. Lee, M. Welsh, and D. Culler, "TOSSIM: accurate and scalable simulation of entire TinyOS applications," In 1st ACM Conference on Embedded Networked Sensor Systems, pp. 126–137, 2003.

[15] S.-Y. Ni, Y.-C. Tseng, Y.-S. Chen, and J.-P. Sheu, "The broadcast storm problem in a mobile ad hoc network," In 5th Annual ACM/IEEE International Conference on Mobile Computing and Networking, pp. 151–162, 1999.

[16] F. J. Ovalle-Martínez, A. Nayak, I. Stojmenovic, J. Carle, and D. Simplot-Ryl, "Area-based beaconless reliable broadcasting in sensor networks," International Journal on Sensor Networks, vol. 1(1/2), pp. 20–33, 2006.

[17] A. Qayyum, L. Viennot, and A. Laouiti, "Multipoint Relaying for Flooding Broadcast Messages in Mobile Wireless Networks," In 35th International Conference on System Sciences, pp. 298–307, 2002.

[18] P. Wei and L. Xicheng, "AHBP: An efficient broadcast protocol for mobile ad hoc networks," Journal of Computer Science and Technology, vol. 16(2) pp. 114–125, 2001.

[19] B. Williams and T. Camp, "Comparison of broadcasting techniques for mobile as hoc networks," In 3rd ACM International Symposium on Mobile Ad Hoc Networking & Computing, 194–205, 2002.

[20] C. Zhu, M.J. Lee, and T. Saadawi, "A border-aware broadcast scheme for wireless ad hoc network," In 1st IEEE Consumer Communications and Networking Conference, 134–139, 2004.

 ISBN: 978-1-61208-100-7

Semantic P2P Overlay for Dynamic Context Lookup

Shubhabrata Sen, Hung Keng Pung,Wai Choong Wong
School of Computing, Department of Electrical and Computer Engineering
National University of Singapore, Singapore
{shubhabrata.sen, dcsphk, elewwcl}@nus.edu.sg

Wenwei Xue
Nokia Research Center
Beijing, China
wayne.xue@nokia.com

***Abstract*— Context-aware applications generally need to retrieve various kinds of dynamic context data from a large number of context sources. A middleware managing context sources must provide an efficient context lookup mechanism to ease application development. In this paper, we categorize the context sources as operating spaces and propose semantic peer-to-peer overlays over these spaces to accelerate dynamic context lookup in a context-aware middleware. Our proposed overlay structure is specially designed to deal with dynamic sensory context such as a person's location and temperature that are frequently changing and difficult to be promptly retrieved using traditional peer-to-peer protocols. Our overlay and indexing method has a low maintenance overhead. Measurement results show that the proposed overlay achieves a good response time and accuracy for context lookup as well as requires low maintenance overhead.**

Keywords-context-awareness; context-aware middleware; operating spaces; semantic peer-to-peer overlays

I. INTRODUCTION

The recent advances in pervasive computing have lead to an increased research focus on the development of sophisticated context-aware applications. Initially restricted to user location, *context* [7] is generally defined as any data that can be used to characterize the situation of an entity involved in the user-application interaction. We call such an entity a *context source*.

Context-aware middleware systems [2][14] have been explored to provide effective support for the development of context-aware applications by providing mechanisms to efficiently manage various data retrieved from context sources. We define an *operating space* as a person, object or place in the physical or virtual world having a software module hosted in a computing device through which the affiliated context sources in the space can communicate with and provide data to external consumers. Example classes of operating spaces include persons, homes, offices and shops.

We define a *context attribute* as a kind of context data provided by an operating space. The context attributes can be static or dynamic. *Dynamic attributes* refer to sensory data that change asynchronously and frequently, such as the *location* of a person and the *temperature* in an office. *Static attributes*, in contrast, refer to data that seldom changes, such as the *name* of a person and the *location* of a shop. Most attributes involved in context-aware applications are dynamic and continuously changing over, as the main focus of these applications is to ensure that they can adapt to the context changes in an unattended fashion.

An important component of a context-aware middleware system is the context lookup mechanism. *Context lookup* is defined as the process of searching and identifying the operating spaces that an application that have the required context data, as well as the actual operation of obtaining the data from each of these spaces. The context lookup process is quite challenging since it requires dealing with dynamic attributes and it is important to ensure that the data being retrieved is up to date. For example, a healthcare application monitoring a patient's heartbeat by data from body sensors triggers an alarm upon any abnormality. The application needs to ensure that it always reads the most recent data because stale data might lead to a severe consequence.

In this paper, we propose to overcome the limitations of traditional database methods for dynamic data management by using a semantic *peer-to-peer (P2P)* overlay [18] as a dynamic context lookup mechanism over operating spaces. The intuition is that by classifying the operating spaces into different semantic domains and having a distributed index for context data enables an efficient context lookup over these spaces. The main reason for us to use a P2P technology for context lookup is to utilize the dynamic leaving and joining facilities in P2P to cope with the dynamic changes in context data. We have implemented and evaluated this overlay structure in a research prototype of context-aware middleware infrastructure named C*oalition* [5] under ongoing development in our project.

The remainder of the paper is organized as follows. We discuss the related work in Section II. We provide an overview of our Coalition system and explain its various components in Section III. Section IV describes our detailed design of the semantic peer-to-peer overlay and the maintenance operations associated with the overlay. We present our current experimental results over the proposed overlay structure in Section V. Section VI concludes the paper with future research directions.

II. RELATED WORK

Traditional database indexes suffer from heavy update cost when data values associated with an index are dynamic and frequently change as in context-aware computing. We need the design of specialized distributed "sparse" index

 ISBN: 978-1-61208-100-7

structures that focus on minimizing index maintenance overhead upon the continuously arriving new data values.

Indexing moving objects: A recent research focus in database community is the data management and indexing techniques for moving objects. This is an enabling technology mainly for location-aware applications. The R-tree style index structures [13][16][19][21] designed for moving objects seek to minimize the index update overhead by minimizing the number of update operations that actually need to be applied to an index. This is realized by representing the motion of a moving object as a function and updating the index when the parameters of the function change or by utilizing physical properties of moving objects like trajectories to make indexing decisions. Trajectory information can be used to predict the motion path of objects.

Indexing dynamic web documents: Inverted index structures are often used to index web documents that provide a mapping between words and their positions in the document. Inverted indexes designed for web documents that change frequently [15][17] seek to minimize the index update overhead by reducing the number of memory operations that occur during an index update. This is achieved by selective rebuilding of the index and by reducing the information to be updated via forward indexing techniques.

The indexing techniques for moving object environments or dynamic web documents are not generic in nature since they are designed for a particular application class and largely take special properties of that application class into account when designing the index. Moreover, they utilize centralized indexing strategies, which could prove to be a very critical bottleneck for a context-aware middleware operating at the Internet scale.

Indexes for P2P systems: The indexes built for P2P systems are generally distributed in structure. Since the P2P systems are intended for information sharing, the indexes are built on the file metadata information like names and sizes [3] which are relatively static. P2P index structures like the routing index [6] attempt to minimize the number of peers to be looked up for a query by associating the routing information with links and do selective forwarding based on the usefulness of neighboring peers. These index structures are same as those designed for working with static data as they focus on the nature of the data content rather than the actual content itself.

Distributed hash table (DHT) based data lookup techniques for P2P was described in Chord [20] and a few other systems like CAN, Pastry and Tapestry. DHT uses hash functions to assign data to nodes as well as to perform the lookup operation. Since a change in data will change the associated hash value, static data is used as the key value. DHT based techniques are first designed to support point queries while later there are DHT variants [10] that work with range queries as well, but neither case explicitly handles data items frequently change in nature and might result in a large update overhead if used with dynamic data.

Indexes for sensor networks: Another area where the problem of dynamic data management is encountered is sensor networks [8][9][12]. Sensor networks require distributed index structures that can efficiently manage dynamic sensor data while ensuring that the data management procedure is energy-efficient. Sensor network indexes are query-driven and operate in a proactive or reactive mode based on the query nature and frequency.

A different approach to managing sensor network data is taken by visualizing the sensor network as a database. This involves processing each query issued to the sensor network similar to a database query and directing the query to specific sensors as determined by the query plan. Since the index structures are constructed based on the query issued, the data management aspect does not address the concern of indexing the actual dynamic sensor data. The scope of these index structures is generally limited to a single sensor network.

III. SYSTEM OVERVIEW

The overall architecture of our Coalition [5][24] context-aware middleware is illustrated in Figure 1. The operating spaces in Coalition are categorized and clustered into multiple domain classes such as persons and shops. Each class is called a *context domain* and all spaces in a domain provide a similar set of context attributes [24]. The system introduces the concept of an *operating space gateway (OSG)*, which is a software program that provides a single point of interaction between an operating space and the "outside" applications or other operating spaces. The OSGs of all operating spaces having a common attribute in a domain are organized into a P2P network. This P2P network having OSG peers is termed as a *semantic cluster*. The semantic clusters for different attributes in a domain are connected with a ring topology to form an overall semantic P2P overlay.

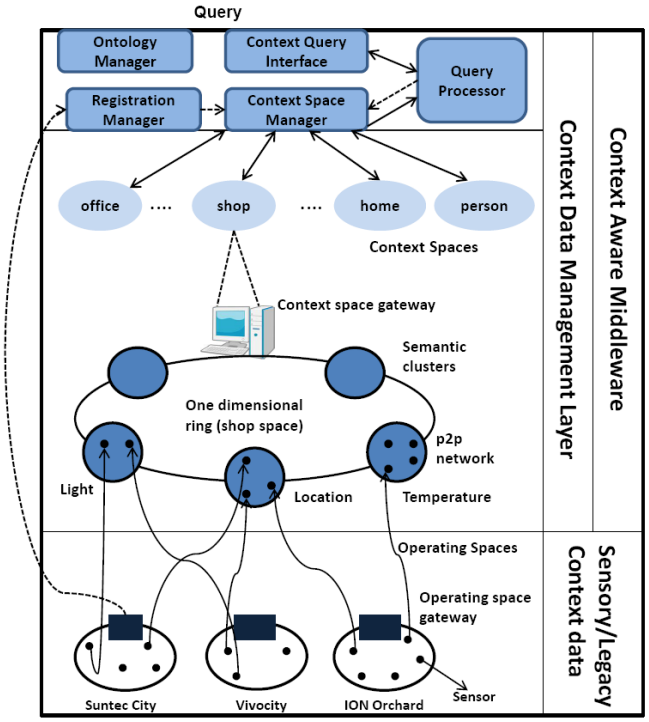


Figure 1. Overview of Coalition System Architecture.

The context data at all operating spaces is modeled using a simple key-value model to make the design generic and extensible. The clustering process of operating spaces into different context domains is automatic. There exist a set of global context schemas at the system server, each of which corresponds to one of the context domains and is incrementally updated from the local schemas submitted by individual operating spaces during registration. The integration of local space schemas into global domain schemas is handled by a context schema matcher that we developed. The technical details of the schema matcher are beyond the scope of this paper and available in our prior paper on schema matching [23].

The context domain manager at the system server manages the data structures of individual context domains and the "pointers" to their corresponding semantic P2P overlays: each overlay of a domain is associated with a *context domain gateway (CSG)* that acts as the single entry of query injection from the server into the overlay. The overlay for the SHOP context domain is amplified in Figure 1 as a first-hand example. More details of the overlay structure and operations are presented in Section IV.

Coalition supports an SQL-based declarative interface [24] for context-aware applications to acquire data from operating spaces or to subscribe to the events of interest that occur in these spaces. The context query engine identifies the proper context domains involved in a query and forwards the query to be executed onto OSG peers in the desirable semantic clusters of the domain's P2P overlay. Because a semantic cluster may contain numerous OSGs, context lookup in a cluster can generate large overhead if a simple flooding protocol is used to route the query throughout the P2P network wherein no distributed indexing is implemented for acceleration. The semantic overlay based lookup mechanism we propose in this paper aims to address this inefficiency issue by reducing the query routing cost via the introduction of attribute value-based sub-clustering and ordering of OSGs within a semantic cluster.

IV. SEMANTIC PEER-TO-PEER OVERLAY

A. Overall Overlay Structure

The structure of our proposed semantic P2P overlay for dynamic context lookup is exemplified in Figure 2.

As we have discussed earlier, the global schema maintained by the Coalition system is integrated from all the local schemas of operating spaces in the domain. The global schema of a domain is then mapped into a semantic P2P overlay with all attributes in the schema mapped to the semantic clusters of P2P networks in the overlay.

A semantic cluster groups gateways for operating spaces that have a common attribute. For example, all OSGs of the shops providing attribute temperature will join as peers in the semantic cluster "temperature", as shown in Figure 2. In order to facilitate a better context lookup operation, a semantic cluster is further partitioned into a number of disjoint range clusters. Each range cluster corresponds to a specific range of values for the attribute (See Figure 2). Each range cluster forms a sub-cluster of smaller P2P network that is contained in the overall P2P network of a semantic cluster. The peer membership of operating spaces in a semantic overlay is flexibly maintained in a distributed manner.

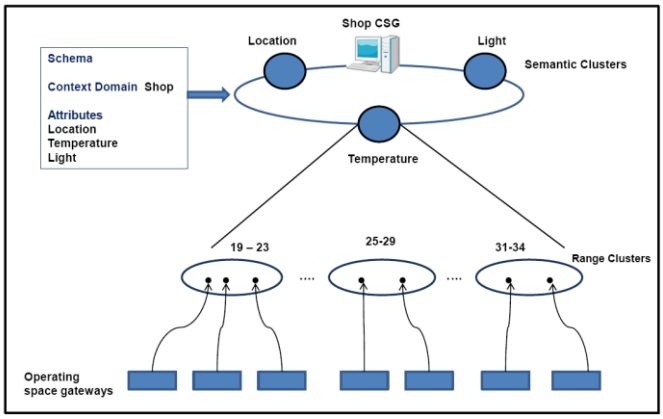


Figure 2. Illustration of the Overlay Structure.

An operating space joins a semantic cluster when it provides the attribute of the cluster, and it leaves the cluster when it no longer provides the attribute. An operating space joins a range cluster of a semantic cluster when its current attribute value falls into the particular value range of the cluster, and it leaves the cluster when its attribute value falls outside the range.

B. Context Lookup in semantic and range clusters

The notion of semantic cluster is introduced in our system to accelerate the context lookup process by grouping operating spaces according to the context data semantics they provide. A semantic cluster for an attribute of a domain is created when the first operating space registers the attribute to Coalition. This 'on-demand' approach bypasses the need to create the semantic clusters beforehand. A semantic cluster is removed when there is no space left in the cluster due to the OSG de-registrations.

An OSG providing an attribute of a domain must be assigned to a proper range cluster in the corresponding semantic cluster of the domain's P2P overlay. The assignment is based on the real-time value of the attribute at the OSG when it registers. Later, each OSG must monitor its attribute value periodically and move itself to a new range cluster when the new attribute value moves out of the bounds of the current range. The operations of an OSG joining a range cluster and switching to a different one upon attribute value change are shown in Figure 3.

Each range cluster maintains a range of numeric values in our approach. Attributes having symbolic string values are handled by hashing the strings to a numeric representation using a hash function. We set two system-defined parameters to restrict the maximum and minimum numbers of OSGs that a range cluster can contain. These values are used to ensure that a cluster is not too heavily or lightly loaded.

The context lookup in Coalition is carried out by an application issuing a query that specifies the required data acquisition from operating spaces satisfying certain range-

 ISBN: 978-1-61208-100-7

search conditions. In the absence of range clusters, the query must be flooded to all OSGs in a semantic cluster and is not scalable with large cluster size. The usage of range clusters addresses this problem by minimizing the search space. The query processing operation is depicted in detail in Figure 4.

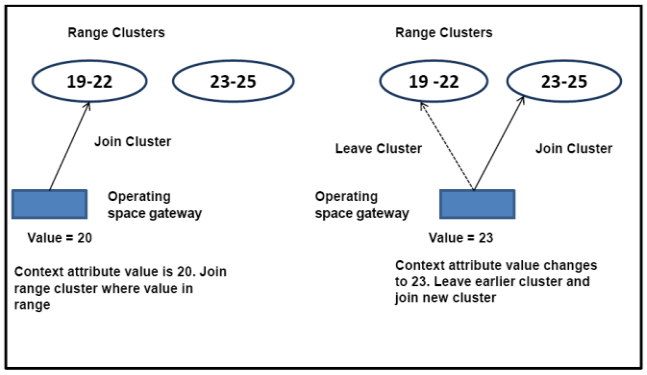


Figure 3. Joining and leaving of range clusters for OSG.

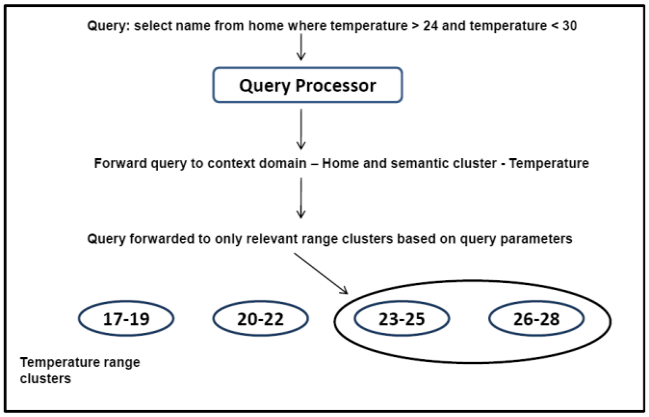


Figure 4. Query processing with range clusters.

Since the query is only flooded in the relevant range clusters rather than the whole semantic cluster now, the context lookup process is more efficient than the flooding based approach.

C. *Maintenance of range clusters*

The basic operations required for the range clusters include the provision for allowing an OSG to join/leave a particular range cluster. Moreover, in order to ensure that the load on a range cluster is not too heavy or too light, a range cluster can either be merged with a neighboring cluster or split into two clusters.

The joining and leaving of OSGs in a range cluster can trigger the splitting and merging of the cluster: if the operation causes the cluster size to fall above/below our system-defined values for maximum/minimum cluster sizes, the cluster is adjusted. The functional pseudo-codes realizing the joining and leaving of an OSG in a range cluster are shown Functions 1-2.

The cluster splitting operation requires OSGs in a range cluster to be redistributed among the two split clusters, as in Figure 5. The operation is carried out by sorting all OSGs in the cluster to be split in an ascending order of their current attribute values. The cluster is then divided into two smaller clusters. The first half of the sorted list is assigned to the first cluster and the second half the second cluster. The cluster bounds of the new clusters are updated accordingly. The cluster splitting operation only affects the cluster being split and does not affect other range clusters. The number of range clusters is increased by one due to this operation.

Function 1: Join_Range_Cluster
Begin
  Locate required context domain and semantic cluster
  If no range cluster present
    Create a new range cluster
  Else
    Locate range cluster by comparing attribute value against cluster bounds
    Assign operating space to the range cluster
    If (range cluster size > Maximum Cluster Size Threshold)
      Split Cluster
End

Function 2: Leave_Range_Cluster
Begin
  Locate range cluster by checking attribute value against cluster bounds
  Remove the operating space from range cluster
  If (range cluster size < Minimum Cluster Size Threshold)
    Merge the cluster with an adjacent one
  Else if (range cluster size == 0)
    Delete corresponding semantic cluster

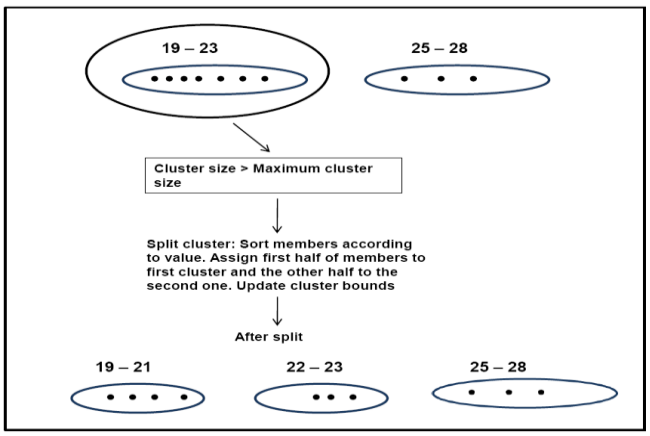


Figure 5. Splitting of a range cluster

In the range cluster merging operation, the cluster is merged with its adjacent cluster to form a bigger cluster. The operation is depicted in Figure 6.

If the cluster to be merged is the first or last range cluster in the sequence, there is only one option for choosing the adjacent cluster to be merged. In case a cluster having two adjacent clusters needs to be merged, the cluster is merged with the adjacent cluster having the lower size among the two. Also, in case the merging process creates a new cluster with a size greater than the maximum cluster size, a

 ISBN: 978-1-61208-100-7

subsequent splitting operation is required. The new cluster bounds are obtained by merging the cluster bounds of the clusters being merged. This operation decreases the number of range clusters by one.

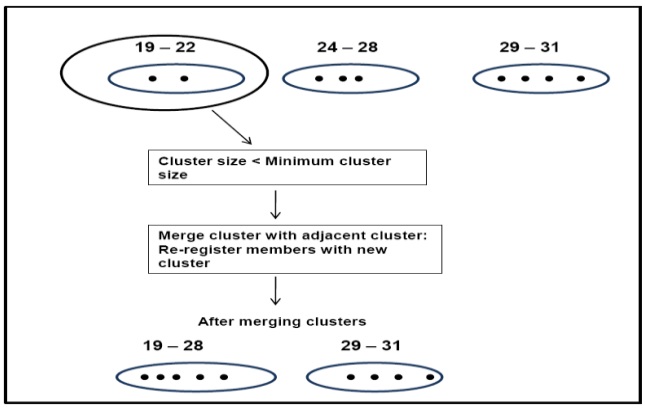


Figure 6. Merging of a range cluster

D. *Further discussions on range clusters*

Our current implementation of range clusters defines the range bounds of clusters as a numerical value range. Simple string attributes can also be partitioned using this scheme by hashing a string to get a hashed numeric value. The range clusters can then be constructed on these hashed values. However, such hashing method is only useful for answering queries that look for exact string matching but not for wildcard based queries.

To handle composite attributes composed of one or more sub-attributes (e.g. *location* composed of *city* and *street*), a simple solution is to hash the composite representation of the attribute value to a real number and generate the range clusters based on the hash values. This approach does not prove to be useful for queries interested in a sub-attribute. The current range cluster approach needs to be extended to support more sophisticated queries with regard to string attributes and composite attributes. The changes to be made to the range cluster structure to handle different types of context in practice also need to be studied.

The proposed current range cluster design generates clusters based on attribute values. Since the values of context attributes tend to be dynamic, the range cluster bounds need to be carefully chosen to minimize the number of cluster update operations. An alternative to this technique would be using the statistical properties of data, e.g. mean and variance, to generate the range clusters. This idea has been explored in [22] where the statistical properties of data are used to build an R-tree like index structure.

V. EXPERIMENTAL RESULTS

We present our current performance evaluation results of the proposed semantic P2P overlay in this section.

A. *Experimental Setup*

We implemented the semantic P2P overlay structure on top of Gnutella [11] in our Coalition prototype. We used a desktop PC as the middleware server and other four PCs to host the OSGs in the experiments. Each PC has an Intel Core 2 Duo 2.83 GHz CPU and runs the Windows XP OS. As our OSG is a software module, we uniformly installed and ran multiple OSG instances onto the four PCs to emulate a number of "operating spaces" in different experimental rounds. Each value in a figure or table shown in the following is the average of tens of independent experimental runs.

B. *Query Response Time*

We first studied the query response time over the semantic P2P overlay. The *response time* of a query is measured as the time interval between the issuing of a query from an application and the reception of query result by the application. In each round of this experiment, a random percentage of OSGs in a few range clusters of a semantic cluster had required results for the query. The total number of OSGs emulated in each run of experiment is the network size for that run.

We compared the query response time achieved by three P2P schemes: the original Gnutella flooding implemented using two different underlying protocols – TCP and UDP, as well as our proposed overlay structure using UDP. The results are shown in Figure 7 with the total number of OSGs in the semantic cluster where the query is routed and executed inside varying.

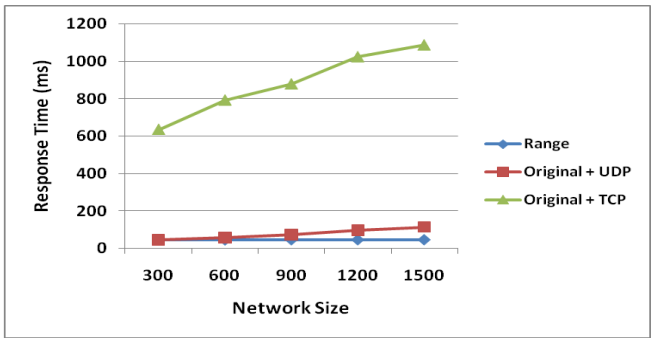


Figure 7. Query response time with different network sizes

The results show that the query response time increased with network size for all three schemes. This is intuitive as network size increase effectively increases the size of search space for a query. The increase curve of our semantic overlay with range clustering (denoted as "Range" in Figure 7) is flatter than the other two schemes. The response time increase of our approach is almost negligible in the experiment, because our distributed indexing based on range clusters largely reduce the search space of query flooding.

We further examined the response time of different schemes by varying the percentage of OSGs with the required results for a query in a semantic cluster. The network size of the semantic cluster was fixed to be of 100 OSGs in this experiment. The results are in Figure 8. Since

 ISBN: 978-1-61208-100-7

the response time of a query depends on the number of OSGs reporting a valid answer, the time intuitively increases with the growing percentage of valid answer sources.

This increase magnitude is quite steep for the original Gnutella scheme, either using TCP or UDP, compared to our approach whose increase curve is smooth after the initial sharp rise. This indicates our semantic overlay could achieve an overall better response time due to the ordering of OSGs according to attribute values, which enables the faster localization of OSGs having valid answers.

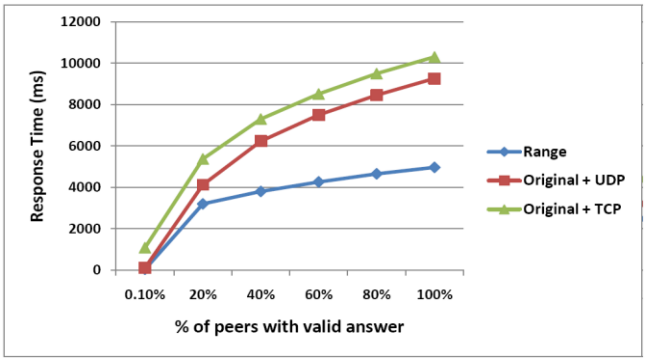


Figure 8. Query response time with different number of qualifying OSGs

C. *Time breakdown for query processing and maintenance operations*

We next analyzed the time breakdown for query processing to identify the time taken in different stages of the operation. Table I shows the breakdown of query response time in our approach and UDP-based Gnutella flooding. The dominant operation in the time breakdown is observed to be the P2P lookup operation since it involves the query flooding within the underlying P2P network. The results clearly illustrate that our approach reduces the P2P lookup time compared to the original UDP version of Gnutella, and achieves an overall faster query response time. Since our approach also requires maintenance like splitting and merging of range clusters, we analyzed the time breakdown of these maintenance operations in Tables II-III, respectively. We also measured the time breakdown of an OSG joining/switching a new range cluster in Table IV.

It is observed that the cluster splitting and merging operations require a similar maintenance overhead with the merging requiring slightly more. The regrouping costs of OSGs in the newly merged/split clusters dominate the total time taken for these operations, as expected. Unless the dynamism of the context data and the memberships of OSGs are high, the frequency of merging/splitting will be low and hence the latency of the operation will not affect the overall system performance. The operation of an OSG joining a new range cluster is comparatively faster than merging or splitting.

TABLE I. TIME BREAKDOWN FOR QUERY PROCESSING

Operation	Time Taken (Range) (ms)	Time Taken (UDP) (ms)
Query parsing	1	1
Context Domain Lookup	2	2
Semantic Cluster Lookup	7	9
Range Cluster Lookup	11.2	
P2P Lookup	26.06	95.4
Total	47.26	105.4

TABLE II. TIME BREAKDOWN FOR CLUSTER SPLITTING

Operation	Time Taken (ms)
Sorting	120
Cluster update	6.3
PSG reclustering	757.9
Total	884.2

TABLE III. TIME BREAKDOWN FOR CLUSTER MERGING

Operation	Time Taken (ms)
Cluster update	0
Re-Grouping of affected PSGs	647.4
Total	647.4

TABLE IV. TIME BREAKDOWN FOR JOINING RANGE CLUSTER

Operation	Time Taken (ms)
Range request	15.9
Neighbour notification	26.5
Neighbour reconnection	18.7
Total Time	61.1

VI. CONCLUSION

We have presented a semantic P2P overlay structure to support dynamic context lookup in a context-aware middleware. The P2P overlay is created for each context domain and represents the set of attributes in the domain. Each semantic cluster of P2P network for an attribute in the domain is further partitioned into a number of range clusters to support the efficient lookup of dynamic context data via simple SQL-based queries. Our simulation results demonstrate the proposed overlay structure achieves a good response time for context lookup. The results also suggest the time required for the range cluster maintenance operations is reasonably small. Our future work includes more extensive testing of the range cluster-based semantic P2P overlay, extending the structure to support symbolic or composite context attributes and the impact of dynamism of OSGs on the system performance.

ACKNOWLEDGMENT

This work is partially supported by National Research Foundation grant NRFIDM-IDM002-069 on "Life Spaces (POEM)" from the IDM Project Office, Media Development Authority of Singapore.

 ISBN: 978-1-61208-100-7

REFERENCES

[1] M. Ali and K. Langendoen, "A case for peer-to-peer network overlays in sensor networks," Proc. WWSNA, 2007, pp. 56-61.

[2] M. Baldauf, S. Dustdar, and F. Rosenberg, "A survey on context-aware systems," International Journal of Ad-Hoc and Ubiquitous Computing 2 (4) (2007), pp. 263-277.

[3] R. Blanco, N. Ahmed, D. Hadaller, L. Sung, H. Li, and M. Soliman, "A survey of data management in peer-to-peer systems," Tech. rep., University of Waterloo (2006), pp. 1-51 .

[4] G. Chen, M. Li, and D. Kotz, "Data-centric middleware for context-aware pervasive computing," Pervasive Mob. Computing. 4 (2) (2008), pp. 216-253.

[5] Context-Aware Middleware Services and Programming Support for Sentient Computing, http://lucan.ddns.comp.nus.edu.sg:8080/PublicNSS /researchContextAware.aspx (last accessed 03/06/2010)

[6] A. Crespo and H. Garcia-Molina, "Routing indices for peer-to-peer systems," Proc. ICDCS, 2002, pp. 23.

[7] A.K. Dey, "Understanding and using context," Personal and Ubiquitous Computing 5 (1) (2001), pp. 4-7.

[8] Y. Diao, D. Ganesan, G. Mathur, and P.J. Shenoy, "Rethinking data management for storage-centric sensor networks," Proc. CIDR, 2007, pp. 22-31.

[9] V. Dyo and C. Mascolo, "Adaptive distributed indexing for spatial queries in sensor networks," Proc. DEXA, 2005, pp. 1103-1107.

[10] J. Gao and P. Steenkiste, "An adaptive protocol for efficient support of range queries in DHT-based systems," Proc. ICNP, 2004, pp. 239-250.

[11] Gnutella Protocol Development, http://rfc-gnutella.sourceforge.net (last accessed 10/07/2010)

[12] B. Greenstein, D. Estrin, R. Govindan, S. Ratnasamy, and S. Shgenker, "DIFS: A distributred index for features in sensor networks," Ad Hoc Networks 1 (2-3) (2003) , pp. 333-349.

[13] A. Guttman, "R-trees: A dynamic index structure for spatial searching", Proc. SIGMOD, 1984, pp. 47-57.

[14] K.E. Kjær, "A survey of context-aware middleware," Proc. SE, 2007, pp. 148-155.

[15] N. Lester, J. Zobel, and H. Williams, "Efficient online index maintenance for contiguous inverted lists", Information Processing & Management 42 (4) (2006), pp. 916-933.

[16] W. Liao, G. Tang, N. Jing, and Z. Zhong, "VTPR-tree: An efficient indexing method for moving objects with frequent updates", Proc. CoMoGIS, 2006, pp. 120-129 .

[17] L. Lim, M. Wang, S. Padmanabhan, J. Vitter, and R. Agarwal, "Efficient update of indexes for dynamically changing web documents", World Wide Web 10 (1) (2007), pp. 37-69.

[18] E.K. Lua, J. Crowcroft, M. Pias, R. Sharma, and S. Lim, "A survey and comparison of peer-to-peer overlay network schemes," IEEE Communications Surveys & Tutorials 7 (2) (2004), pp. 72-93.

[19] S. Šaltenis, C.S. Jensen, S.T. Leutenegger, and M.A. Lopez, "Indexing the positions of continuously moving objects", Proc. SIGMOD, 2000, pp. 331-342.

[20] I. Stoica, R. Morris, D. Liben-Nowell, D. Karger, M. Kaashoek, F. Dabek, and H. Balakrishnan, "Chord: A scalable peer-to-peer lookup protocol for internet applications," IEEE/ACM Transactions on Networking 11 (1) (2003), pp. 17-32.

[21] Y. Tao, D. Papadias, and J. Sun, "The TPR*-tree: An optimized spatio-temporal access method for predictive queries", Proc. VLDB, 2003, pp. 790-801.

[22] Y. Xia, S. Prabhakar, S. Lei, R. Cheng, and R. Shah, "Indexing continuously changing data with mean-variance tree", Proc. SAC, 2005, pp. 263-272.

[23] W. Xue, H.K. Pung, P.P. Palmes, and T. Gu, "Schema matching for context-aware computing," Proc. Ubicomp, 2008, pp. 292-301.

[24] W. Xue H.K. Pung W.L. Ng, and T. Gu,"Data management for context-aware computing", Proc. EUC, 2008, pp. 492-498.

[25] W. Xue, H.K. Pung, W.L. Ng, C.W. Tang, and T. Gu, "Gateways of physical spaces in context-aware computing," Proc. ISSNIP, 2008, pp. 441-446 .

 ISBN: 978-1-61208-100-7

Exploring Techniques for Monitoring Electric Power Consumption in Households

Manyazewal Fitta, Solomon Biza, Matti Lehtonen, Tatu Nieminen
Department of Electrical Engineering
Aalto University, School of Science & Technology
Espoo, Finland
manyazewal.fitta@tkk.fi, solomon.biza@tkk.fi,
matti.lehtonen@tkk.fi

Giulio Jacucci
Helsinki Institute for Information Technology HIIT and
Department of Computer Science
University of Helsinki
Helsinki, Finland
giulio.jacucci@hiit.fi

***Abstract*—Recent works in ubiquitous computing have addressed analysis of electric power for energy conservation by detailing and studying consumption of electrical appliances. We contribute with an approach to develop techniques for fingerprinting and monitoring consumption of electric power in households. The approach builds on previous works and employs three phases: feature extraction of attributes such as real power and current harmonic contents, event detection and pattern recognition. A load library is foreseen that stores appliance characteristics as corpus data for training and recognition. We report early findings achieved using a high definition sensor directly applied to the loads showing promising results but also challenges in event detection (smaller state transitions, challenges in detecting and pairing switch on and off events). These studies are important in order to be able to address opportunities of identifying and monitoring directly at the appliance level or sensing the total load of the network. Future applications include monitoring and detailing of loads in a "balance sheet" and context aware service with advice tips for energy users.**

Keywords- ubiquitous computing; load monitoring; fingerprinting; pattern recognition; energy awareness

I. INTRODUCTION

In ubiquitous computing, measuring electric power consumption has been pursued for energy conservation and also as a means to recognize user activities [1]. Power measurement and analysis can be done at whole-house level as demonstrated in [2] and [3], on selected plugs as shown in [4] or by placing sensors near electrical devices [5]. The current state is that several electrical appliances can be identified either by using measurement from the main power entry point or with sensors placed near the appliance. While researches successfully showed a way to finger print appliances switching on and switching off, less explored activities are beyond these simple operations such as the analysis of specific uses of the device. Appliances can have nominally the same instantaneous power usage and yet perform quite different operations. Such variety of operations can be possibly tracked by utilizing different attributes with a state of the art sensor that monitors electrical parameters such as active power, power factor, harmonic distortion, crest factor and energy consumption.

In this paper, we discuss techniques for appliance identification and monitoring system that uses measurement data received from sensors installed in customer premises for applications in energy awareness and power quality monitoring. Two techniques, namely, device fingerprinting and balance sheet application are being developed, which are foreseen to provide detailed information about residential electric power consumption in a near-real-time processing environment.

This paper is organized as follows: a review of related work is presented in Section II; the two techniques being developed are introduced in Section III; Section IV discusses device fingerprinting in detail and Section V presents preliminary findings; Section VI explains the balance sheet application. Finally, Section VII concludes the paper, pointing out major problems and further improvements.

II. RELATED WORK

Earlier researches conducted on nonintrusive appliance load monitoring system (NIALMS) were based on the principle that the waveform of a total site load changes in predictable manner as appliances are individually turned on and off [7].

One of the pioneering researches on load identification was made in the US at Massachusetts Institute of Technology [6]. An algorithm was developed for two-state appliances and a recorder was installed next to the existing energy meter, with sampling rate of 2 kHz, to record step changes in real and reactive power (ΔP, ΔQ). The time stamped edge data are then stored for semi-automatic identification by employing a library of load models and finding a possible match. The main limitations of this work were inability to detect edges (events) for appliances that are always in continuous operation and to distinguish between appliances with similar power consumption profiles.

Another study, in Finland, at VTT Technical Research Center implemented a 3-phase power quality monitoring energy meter and it developed its own load identification algorithm that required a prior manual set up for the naming of appliances and building a signature library (the manual set up is a one-time intrusive activity in which signatures are observed and named as appliances are manually turned on and off) [7]. The system utilized fundamental frequency signatures (ΔP, ΔQ) for edge detection and load identification.

 ISBN: 978-1-61208-100-7

Subsequent studies put into practice supplementary methods such as higher-order current harmonics and transients to be used as additional parameters in load identification [8]. Higher harmonics studies showed that evaluating power consumption data at higher frequencies helps to disaggregate certain loads that have overlapping consumption profiles at fundamental frequency. Another advanced approach presented in [9], in addition to the parameters at higher harmonics, used phase shift between the fundamental input current and the source voltage for load identification.

The study on transients was based on the principle that appliances can be identified by their unique load transient shapes [8]. Events detected from the aggregate power stream are matched to previously recorded and defined transient signatures. However, the test results for the sampled devices indicate that similar transient patterns can be achieved only for switch-on transitions.

On the pattern recognition aspect, an algorithm for identifying the type of domestic appliances based on fuzzy logic theory is proposed in [10]. It identifies an electric load by comparing its transient with a database of known transients and selecting the closest match. The algorithm was tested on a few sample waveforms and it is stated that promising results were obtained.

Other researches of interest in this field include - load monitoring system developed in Germany using existing energy meters fitted with optical sensors, appliance identification algorithm based on Dynamic Time Warping for micro grids and NIALMS based on Integer Programming, which are presented in [11], [12] and [13] respectively.

III. FINGERPRINTING AND CONSUMPTION BALANCE SHEET CALCULATION

The device fingerprinting concept is based on the fact that different appliances exhibit different operational electrical characteristics. Fingerprinting adds the capability to monitor appliances individually or at least in categories, which in turn provides the possibility for near-real-time monitoring of appliances' energy consumption and thus the corresponding energy cost.

The BeAware project fingerprinting research is aimed at exploring various electrical characteristics of household appliances that have the potential to be used in load identification techniques. On the other hand, the balance sheet feature provides the aggregate energy consumption report and more importantly, the respective share of categories of appliances that are not directly monitored by the sensor.

Before commencing the development of the fingerprinting and balance sheet applications, an extensive series of measurements on various household appliances was conducted. The primary objectives were to obtain a technical load classification and to examine the operating behaviors of different loads using the measurement results.

It was observed that most appliances fall in one of the four major classes: resistive load, electronic load, motor driven load or general inductive load. From an operation point of view, appliances can also be divided as:

- Short-cyclic, long-cyclic or continuously-operating depending on their ON cycle time
- Mono-mode or multi-mode depending on the number of sequences of tasks they perform
- Two-state or multi-state depending on the number of distinct switch-on states they possess

In addition, closer studies were performed on switch-on, switch-off and steady-state characteristics, duty cycle and current harmonic contents of individual appliances.

IV. DEVICE FINGERPRINTING APPLICATION

The device fingerprinting process is depicted in Fig. 1. It consists of three steps - feature extraction, event detection and pattern recognition [14]. The load library stores appliance signatures that are used for matching purpose during pattern recognition stage.

A. *Feature extraction*

Different attributes such as real power, current harmonic contents and power factor are extracted from voltage and current waveforms. Feature extraction is performed by a sensor installed in the customer premise.

For this particular work, TOPAS 1000 power quality analyzer [20], a product of LEM NORMA GmbH, Austria, was used for feature extraction. It is capable of measuring harmonics up to 50^{th} order. The unit has 8 electrically isolated analogue inputs that can be used for current and voltage measurement. Each channel is equipped with a 16-bit analogue-to-digital converter. Sampling of all channels is synchronous based on a common clock signal. The sampling rate is synchronized with the line frequency and is typically 6400 Hz on a 50 Hz line.

B. *Event detection*

Changes in the extracted features are detected and classified as events based on static or dynamic thresholds. The information obtained from the event detection stage is required for:

- Triggering the appliance identification (pattern recognition) stage
- Computing the appliance's duration of operation (length of time it was connected)
- Estimating the energy consumed in that period

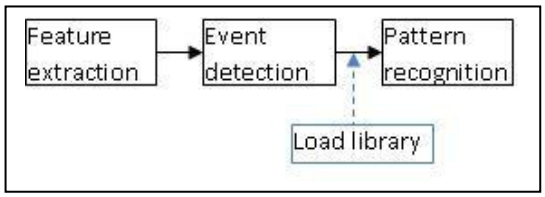


Figure 1. Fingerprinting process.

 ISBN: 978-1-61208-100-7

There are several ways in which events can be defined. 'For example, events can be changes in the average real power exceeding a certain threshold, the appearance of a known start up transient shape or any other suitable condition' [14]. It is possible to select a single method or to have a combination of different methods for detecting events.

After selecting a suitable method, it is necessary to define an appropriate threshold. Previous works such as [3], [6], [7] and [15] utilized changes in average real and/or reactive power for detecting events based on static thresholds. For instance, in [15] changes are classified as events if they exceeded 200 W. Opting for a static threshold presents its own challenges: choosing a large setting means it will not be possible to detect the operation of small appliances and on the contrary, a small setting means the system becomes too sensitive for appliances with large power consumptions.

The event detection technique in our work will employ a dynamic threshold, which allows the detection of events for different appliances over a wide range of real power consumption. The advantage is that the threshold adjusts itself according to the power consumption level of the appliance connected to the sensor, which avoids the need for equipment-specific settings.

The other challenge in event detection is the need to distinguish between genuine events and fluctuations that normally occur during the operation of majority of appliances. Fig. 2 shows the load curve of a sample desktop computer during a half hour operation. In this particular example, only the switch-on and switch-off transitions (marked with bold arrows) shall be classified as events.

C. *Pattern recognition*

At this stage, a set of averaged steady-state load features, captured after the occurrence of a corresponding event, is processed to find its match from a load library using pattern recognition. A study conducted on load monitoring techniques compared the performance of four different training classifiers in noise free (laboratory) as well as real world situations [16]. The tested classifiers were – Gaussian Naïve Bias, 1-Nearest Neighbor (1-NN), AdaBoost and Decision Tree. One of the conclusions of the study indicates that 1-NN algorithm, which is a specific instance of the *k*-NN algorithm (*k*=1), has a decent performance in classification tasks when applied in load monitoring. For the sake of its simplicity and adequate performance, it was studied in further detail so as to explore its applicability in the BeAware fingerprinting process.

The nearest neighbor algorithm was originally suggested in [17] and nowadays it is among the most applied classification methods. Its operation is based on comparing a new record with a set of training records in order to find the ones that are similar to it [18]. The training phase of this algorithm consists of storing the training records and their class labels.

Every record with *n* attributes represents a point in an *n*-dimensional space. When given a new record, the *k*-NN algorithm searches the space for the *k* training records that are nearest to the new record and then predicts the label of the new record using the class labels of these nearest neighbors.

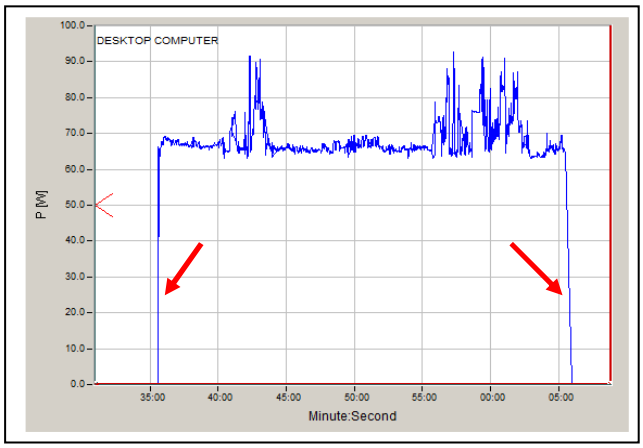


Figure 2. Events and fluctuations during an appliance operation.

In this algorithm, nearness is defined in terms of a distance metric such as Euclidean distance. For any two records consisting of *n* continuous attributes or two points in an *n*-dimensional space, $\mathbf{p} = (p_1, p_2 \ldots p_n)$ and $\mathbf{q} = (q_1, q_2 \ldots q_n)$, the Euclidean distance d(**p,q**) is defined as –

$$\mathrm{d}(\mathbf{p}, \mathbf{q}) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2} \quad (1)$$

After calculating the distances between the new record and the respective training records, the resulting values are sorted and *k* nearest neighbors are selected. A training record will qualify as nearest neighbor if its distance from the new record is less than or equal to the k^{th} smallest distance.

The next step is to provide a classification decision for the new record based on the class labels of the selected *k* nearest neighbors. Generally, two methods exist: *unweighted voting*, in which the class label most frequent among the neighbors is simply selected as the class label of the new record without considering the preference of the neighbors, and *weighted voting*, in which more weight is given to the neighbors that are closer to the new record.

A common weighting scheme is to give each neighbor a weight of ***1/d***, where ***d*** is the distance to the neighbor. Thus, the nearer neighbors will have more influence on determining the class label than the more distant ones. After weighting the neighbors, the sum of weights of neighbors with the same class label is calculated. Finally, the class label corresponding to the neighbors with the largest sum of weights is selected as the class label (identity) of the new record.

If the features of records, for instance $x_1, x_2 \ldots x_n$ of **x**, are on different scales then it is necessary to remove scale effects. A common way to do this is to transform the records by applying *Z-score standardization*, in which a raw feature

 ISBN: 978-1-61208-100-7

value x_{ij} is transformed using the mean and standard deviation of all feature values, given by the relationship:

$$\frac{x_{ij} - \mu_j}{\sigma_j} \tag{2}$$

where x_{ij} is the j^{th} feature of the i^{th} record, μ_j is the arithmetic mean and σ_j is the standard deviation of all the j^{th} features. The distance calculated using these transformed values is called *standardized Euclidean distance*. For the i^{th} record $\mathbf{x_i}$ and new record $\mathbf{y}$, the standardized Euclidean distance can be written as:

$$d(x_{ij}, y_j) = \sqrt{\sum_{j=1}^{n} (1/\sigma_j^2) * (x_{ij} - y_j)^2} \tag{3}$$

Since the j^{th} mean cancels out when computing the differences between the corresponding feature values, the standardized Euclidean distance is in effect the normal Euclidean distance with a weight attached to the respective squared differences, which is the inverse of the j^{th} variance ($1/\sigma_j^2$).

In order to apply this pattern matching technique in the fingerprinting process, all measured appliances are modeled as records using their steady-state attributes and stored in a load library. Another important aspect is the accuracy of pattern recognition; validation of the classification accuracy is usually performed by using *k-fold cross validation* (also possible with *leave-one-out cross validation* at k=1). The result obtained will be indicative of the classification accuracy for the particular dataset, which is characterized by its number of attributes, size of instances (records) and number of class label types.

D. Load library

The basis for the pattern recognition phase is the preparation of a load library that shall consist of known appliance signatures. Taking into account the parameters that will be measured and transmitted by the sensor, a given appliance can be characterized by the following average values - active power (P), power factor (PF), crest factor (CF), total harmonic distortion (THD_I), fundamental power factor (FPF) and 3rd harmonic, 5th harmonic and 7th harmonic currents.

It is planned to create a large load library using appliance samples, which will be collected during the project trial phase. The information that will be stored in the load library also provides the opportunity to monitor the performance of an individual appliance; for instance, to track any degradation in its efficiency through time.

V. PRELIMINARY FINDINGS

Preliminary tests were done for the event detection and pattern recognition stages using measurement data collected from a number of household appliances during the measurement period, which are listed in Table I.

For the event detection part, an algorithm with dynamic threshold was developed, which detects events using criteria that are defined based on standard deviation of a stream of active power measurement. For test purpose, accumulated active power values (recorded at one second interval) were simulated as incoming data.

The tests done on the available two-state appliances resulted in the correct detection of 87.50% of the total possible *switch-on events* and 90.63% of the total *switch-off events*.

The following points were observed from the event detection tests –

- Switch-on events belonging to appliances with large starting current (such as refrigerators) were not detected. Besides, the falling edge of such an initial spike can be wrongly detected as a switch-off event. Hence, an improvement is needed to search for the first stable transition point.
- In the case of multi-state appliances (such as a fan operating at different speeds), improvement is needed to detect smaller state transitions (other than the two large switch-on and switch-off events, intermediate transitions can possibly occur due to user selection or automatic settings).
- By observing the number of switch-on and switch-off event pairs (especially if they occur within a short period of time as in the case of microwave oven), it is possible to substantiate the output of the pattern recognition algorithm.

TABLE I. APPLIANCES LIST WITH PARTIAL PARAMETER VALUES

	Appliance name	P [Watt]	PF	THD_I
1	Coffee maker	2079.16	0.99	1.24
2	Space heater	1196.09	0.99	0.91
3	LCD PC monitor	33.93	0.51	156.92
4	LED lamp	16.33	0.47	178.93
5	Desktop computer	110.35	0.77	81.60
6	CRT television	84.14	0.76	79.42
7	CRT PC monitor	88.49	0.76	83.19
8	Vacuum cleaner	1119.10	0.96	24.56
9	CFL	14.37	0.57	110.75
10	Table fan	26.80	0.61	12.04
11	Freezer	334.27	0.68	10.23
12	Incandescent lamp	60.83	0.99	1.35
13	Refrigerator	99.50	0.73	8.16
14	Laptop	23.36	0.37	233.61
15	Dishwasher - wash cycle	76.81	0.89	6.17
16	Fluorescent lamp	53.78	0.52	9.42
17	Cloth washer- spin cycle	420.85	0.79	47.04
18	Toaster	754.30	0.99	2.03
19	Halogen lamp	53.88	0.99	1.43
20	Microwave oven	796.87	0.73	27.92

 ISBN: 978-1-61208-100-7

The pattern recognition stage was also tested with the available appliance samples using the nearest neighbor approach explained in the previous section. By removing one appliance record from the load library at a time, the classification capacity was checked for the available records, which yielded an *overall accuracy* of 62.30%.

The respective values for active power (P), power factor (PF) and total harmonic distortion (THD_I) are provided for selected samples from each appliance type in Table I. These three attributes are among the eight parameters used for defining a given sample during recognition phase.

This test is equivalent to a first-time classification, i.e., without prior knowledge about the appliance nature except the availability of other records from the same family of appliances in the load library. For instance, if there are 3 coffee maker records in the load library, one will be removed temporarily and a possible match is searched for this record from the library. It is expected that subsequent re-classifications of the same set of appliances using data acquired from new measurements will have a much improved accuracy owing to the fact that the data obtained from these new measurements will closely match with the respective steady-state features already available in the load library. This is assuming that newer measurements are conducted while the appliance is operating under *normal condition* and also assuming that there is *no degradation* in the performance of the appliance through time.

The first-time classification itself can be improved by increasing the size of the load library via the collection of more appliance samples during the project trial phase. Furthermore, it is planned to improve the overall reliability of the system by modeling the operation of more complex appliances, such as those with multiple states and those performing sequences of tasks, using other techniques such as Hidden Markov Model in conjunction with the nearest neighbor algorithm.

VI. BALANCE SHEET FOR LOAD DISAGGREGATION

The device fingerprinting application relies on information received from the sensor, which at present is designed for precise measurement at socket outlet level. This calls for the development of a supplementary technique to keep track of the consumption that is not directly monitored. The balance sheet application is designed to serve this purpose, i.e., to compute the power consumption of the loads that are not monitored individually and thus disaggregate their consumption and provide their respective energy cost.

Such inaccessible loads are directly connected to the electric supply network, which include significant domestic loads such as electric Heating, Ventilation and Air Conditioning (HVAC). The parameters proposed to estimate these permanently connected loads are electrical consumption and external factors (temperature, season and time of day).

A. *Electrical consumption*

Electrical consumption share of an unidentified load is calculated by the balance sheet from the aggregate consumption data and individual consumption shares of plugged-in devices. For plugged-in devices, the sensor records and transfers consumption readings of each appliance. The calculated consumption share of an unidentified load is cross-checked with the offline range of consumption profile for common significant household appliances. The reference offline consumption profile is developed from the data gathered during the project trial phase.

When a match is obtained, the system will keep on checking for external factors or indicators to ensure the reliability of the estimation. The proposed external, non-electrical parameters used for load identification are temperature, season and time of day data during operation. Evaluating as many applicable parameters as possible will allow the system to provide a more accurate estimate for load identification. Even if a match is not found from the offline reference, the system still checks for the availability of external parameters indicating the load type. If no related external factor exists, then the load is labeled as anonymous load together with its calculated consumption share from the balance sheet.

B. *External factors*

The proposed external factors, temperature, season and time of day, are essential parameters to consider for implementing a probabilistic approach in order to estimate significant loads. However, experimental study results are required on each parameter to confirm the relation with electrical consumption. In this project, an indoor temperature reading device will be installed at the base station, which is a measurement gateway installed at each household, and the temperature data will be analyzed with electrical consumption data of significant loads. In addition, outdoor temperature measurement will also be provided as an input to the system.

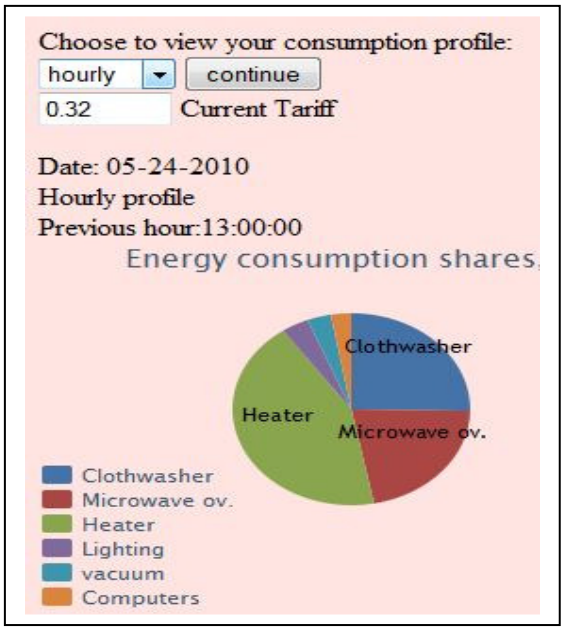


Figure 3. Snapshot of part of the balance sheet model.

Even if the applicability of the parameters is not verified yet using probabilistic approach, practical assumptions were considered for the model balance sheet so as to simulate the load identification. For instance, temperature reading is expected to have inverse relationship with electrical heating system consumption. Air conditioners are likely to operate during the summer season and there is a common time of the day for certain residential appliances to be used.

The model balance sheet developed is a dynamic web page presenting graphically (with a pie chart) consumption shares and cost shares of plugged-in as well as directly connected loads. After estimating or identifying the permanent loads as well as receiving the outputs of the device fingerprinting algorithm for plugged-in devices, their individual consumption is used in the balance sheet calculation. The percentage shares are finally calculated from the individual shares and the total household energy consumption. These percentages are given as inputs to the pie chart, which accordingly displays the consumption shares as shown in Fig. 3.

VII. CONCLUSION

Given the rapid developments in sensor and communication technologies, systems for analysis of electrical power are of great interest to ubiquitous computing for application in energy conservation or recognition of user activities. Different approaches have been documented in the literature about monitoring the total load and trying to fingerprint individual loads by utilizing different parameters from simple instantaneous power through optical interfaces to more sophisticated sensing. We proposed an approach that opens the possibility to integrate sensors also at the plug level as a new commercial solution and also utilizing advanced analysis based on various electrical attributes.

The approach comprises three phases: feature extraction, event detection and pattern recognition. In feature extraction, detailed characteristics are extracted such as active power, power factor and harmonic contents. In event detection, changes in extracted features are detected for triggering appliance identification, computing duration of operation and the energy consumption. The pattern recognition phase starts with a given set of averaged steady-state load features, captured after the occurrence of a corresponding event, and is then processed to find a match from a load library using pattern matching techniques.

The approach has been experimented using a high definition sensor (power quality analyzer) and has resulted in promising findings as well as certain challenges. Some issues were encountered in identifying state transitions of appliances with complex operations and detecting switch-on and switch-off events of certain types of appliances. The pattern recognition phase in this work is implemented using the nearest neighbor approach and it can also be enhanced by employing other techniques such as HMM.

The preliminary results are promising showing satisfactory success rates even with small number of samples in the load library. Such a sensor-based system is foreseen to be applied in real time analysis and monitoring of loads at socket outlet level as well as at whole-house level. This allows the accurate tracking of different appliances followed by the identification of other significant loads from the remaining unknown consumption, thereby enabling the creation of a "balance sheet" of electric power consumption for a specific household.

A fully developed working system needs to overcome a number of problems that include production of hardware and refinement of techniques presented here including collection of a database of load signatures. In addition to appliance monitoring and detailing of loads in a "balance sheet", future applications in the BeAware project [19] will include the development of context aware services with advice tips for energy consumers that are triggered when specific situation arises, for example, when users open too many times or for too long the refrigerator door. Research will have to explore which situations can be exploited in order to provide such services through the utilization of advanced monitoring techniques.

ACKNOWLEDGMENT

This work has been co-funded by the European Union through the 7th Framework ICT Programme in the BeAware Project 224557.

REFERENCES

[1] J. Fogarty, C. Au, and S.E. Hudson, "Sensing from the Basement: A Feasibility Study of Unobtrusive and Low-cost Home Activity Recognition," Proc. 19th Annual ACM Symposium on User Interface Software and Technology (UIST 06), ACM Press, Oct. 2006, pp. 91–100, doi: 10.1145/1166253.1166279.

[2] S.N. Patel, T. Robertson, J. A. Kientz, M.S. Reynolds, and G.D. Abowd, "At the Flick of a Switch: Detecting and Classifying Unique Electrical Events on the Residential Power Line," Proc. 9th International Conference on Ubiquitous Computing (UbiComp 07), Springer-Verlag, Sept. 2007, pp. 271-288, ISBN ~ ISSN:0302-9743, 978-3-540-74852-6.

[3] G.W Hart, "Nonintrusive Appliance Load Monitoring," Proceedings of the IEEE, vol. 80, no. 12, Dec. 1992, pp. 1870-1891, doi: 10.1109/5.192069.

[4] X. Jiang, S.D. Haggerty, P. Dutta, and D. Culler, "Design and Implementation of a High-Fidelity AC Metering Network," Proc. 2009 International Conference on Information Processing in Sensor Networks (IPSN 09), IEEE Computer Society, Apr. 2009, pp. 253-264, ISBN:978-1-4244-5108-1.

[5] Y. Kim, T. Schmid, Z.M. Charbiwala, and M.B. Srivastava, "ViridiScope: Design and Implementation of a Fine Grained Power Monitoring System for Homes," Proc. 11th International Conference on Ubiquitous Computing (UbiComp 09), ACM Press, Oct. 2009, pp. 245-254, ISBN: 978-1-60558-431-7.

[6] Electric Power Research Institute (EPRI), "Nonintrusive Appliance Load Monitoring System (NIALMS) Beta-test Results," Technical Report, Sept. 1997, TR-108419.

[7] H. Pihala, " Non-intrusive Appliance Load Monitoring System Based on a Modern kWh-meter," VTT publications, no. 356, May 1998.

[8] C. Laughman et al., "Power Signature Analysis," Power and Energy Magazine, vol. 1, no. 2, Mar-Apr. 2003, pp. 56-63, doi: 10.1109/MPAE.2003.1192027.

[9] Y. Nakano et al., "Non-Intrusive Electric Appliances Load Efficient Monitoring System Using Harmonic Pattern Recognition-Performance Test Results at Real Households," The Fourth International Conference on Energy Efficiency in Domestic Appliances and Lighting (EEDAL), June 2006, London.

 ISBN: 978-1-61208-100-7

[10] S.R. Kamat, "Fuzzy Logic Based Pattern Recognition Technique for Nonintrusive Load Monitoring," Proc. IEEE Regional 10 Conference (TENCON 2004), IEEE Press, Nov. 2004, vol.3, pp. 528-530, doi: 10.1109/TENCON.2004.1414824.

[11] M. Baranski and J. Voss, "Nonintrusive Appliance Load Monitoring based on an Optical Sensor," Proc. IEEE Power Tech Conference 2003 IEEE Bologna, June 2003, IEEE Press, vol. 4, 8 pp., doi: 10.1109/PTC.2003.1304732.

[12] F. Kupzog, T. Zia, and A.A. Zaidi, "Automatic Electric Load Identification in Self-Configuring Microgrids," Proc. AFRICON 2009 (AFRICON 09), IEEE Press, Sept. 2009, pp. 1-5, doi: 10.1109/AFRCON.2009.5308129.

[13] K. Suzuki, S. Inagaki , T. Suzuki, H. Nakamura, and K. Ito, "Nonintrusive Appliance Load Monitoring based on Integer Programming," Proc. SICE Annual Conference, IEEE Press, Aug. 2008, pp. 2742 – 2747, doi: 10.1109/SICE.2008.4655131.

[14] H.S. Matthews, L. Soibelman, M. Berges, and E. Goldman, "Automatically Disaggregating the Total Electrical Load in Residential Buildings: a Profile of the Required Solution," Proc. Intelligent Computing in Engineering (ICE08), July 2008, pp. 381-389, ISBN: 978-1-84102-191-1

[15] M.L. Marceau and R. Zmeureanu, "Nonintrusive Load Disaggregation Computer Program to Estimate the Energy Consumption of Major End Uses in Residential Buildings," Energy Conversion and Management, vol. 41, October 1999, pp. 1389-1403.

[16] L. Soibelman, H.S. Matthews, M. Berges, and E. Goldman, "Automatic Disaggregation of Total Electrical Load from Non-intrusive Appliance Load Monitoring," Carnegie Mellon University, Feb.2009, <dodfuelcell.cecer.army.mil/rd/NZE_Workshop/7d_Soibelman.pdf> [Accessed 11.03.2010]

[17] T.M. Cover and P.E. Hart, "Nearest Neighbor Pattern Classification," IEEE Transactions on Information Theory, vol. 13, no. 1, Jan. 1967, pp. 21-27, ISSN : 0018-9448.

[18] M. Moradian and A. Baraani, "K-Nearest-Neighbor-Based-Association-Algorithm," Journal of Theoretical and Appllied Information Technology, vol. 6, Dec. 2009, pp 123-129.

[19] G. Jacucci et al., "Designing Effective Feedback of Electricity Consumption for Mobile User Interfaces", PsychNology Journal, vol. 7, no. 3, 2009, pp. 265-289.

[20] LEM Norma GmbH, "TOPAS 1000 Power Quality Analyzer", Operating Instructions, 2003, A 5505 1 GA 3 E Rev. A.

 ISBN: 978-1-61208-100-7

Extending a Middleware for Pervasive Computing to Programmable Task Management in an Environment of Personalized Clinical Activities

Giuliano Ferreira, Iara Augustin,
Giovani Rubert Librelotto,
Fábio L. da Silva, Alencar Machado
Federal University of Santa Maria
Technology Center
Avenida Roraima 1000, 97105-900
Santa Maria, RS, Brazil
Email: {august,librelotto}@inf.ufsm.br,
giuliano@cpd.ufsm.br, alencar.ufsm@gmail.com

Adenauer Correa Yamin
Catholic University of Pelotas
Rua Felix da Cunha, 412, 96010-100
Pelotas,RS, Brazil
Email: adenauer@ucpel.tche.br

***Abstract*—Currently, Pervasive Computing has focused on the development of programmable and interactive environments, which are intended to help the user in daily activities. The health system of the future envisages the use of Pervasive Computing as a way of optimizing and automating clinical activities. Under such perspective, the present study has tried to adapt a middleware for pervasive environment management to support and manage the accomplishment of clinical tasks (pervasive applications that help physicians perform their activities), fulfilling some requirements of activities-oriented computing, and creating a tool that will help physicians in their daily tasks. So, ClinicSpace can be seen as a system context aware pervasive oriented clinical tasks.**

***Keywords*-Ubiquitous Computing; middleware; daily activities oriented computing; end-user programming; clinical activities.**

I. INTRODUCTION

The first systems of pervasive computing have concentrated on the creation of middleware for management and availability of the ubiquitous environment, aiming at understanding its requirements and need [6][3][14]. Systems available at that time made possible to create some concepts inherent to the nature of the pervasive space: communication, mobility, context-awareness and daily activities. Communication and mobility have been approached for a long time; context is currently being handled; but daily activities require further efforts, because they are in an early phase, where what is desirable has still to be defined [10].

The current focus of Pervasive Computing is turned towards processing daily human activities in the most integrated way as possible to the real environment known by the final user (user-centric computing). Under such perspective, one of the major application areas is Health Care, it already faces situations where information from the physical world is proactively acquired and automatically integrated to applications (virtual world) [15].

Then, we can consider that Pervasive Health (or Ubi-Health) is in its first generation, trying to understand the needs, features and technologies required to design systems that will create the hospital of the future [9]. The project Hospitals of the Future envisages the use of technologies that will make an intelligent space, reactive and proactive, where information management systems will take decisions and will adapt to the situations they detect [11].

However, one argues that the system proactivity should not be too strict, once it is designed in a generalized way and not customized. If the physician wants to make things his way, he must be able to do so and interact, command and interfere in the tasks/activities managed by the pervasive system. Therefore, the pervasive system is required to focus on the end-user (physicians) and provide support to his daily tasks, balancing between proactivity (act in the place of the user) and customization (individual way of performing a task).

Based on such premises, the proposal of the project "ClinicSpace: Support to clinical tasks in the hospital environment of the future based on Ubiquitous/Pervasive Computing technologies" is to develop a pilot-tool that makes it possible to physicians to customize his tasks, which are managed and performed by a middleware in a pervasive environment. The main goal in the customization of tasks is to reduce the impact of interference of the automated system in a clinical environment and, therefore, minimize the high rejection to computational systems that such interference may cause. So, the middleware must provide mechanisms for users to run, stop, resume and schedule their tasks. Moreover, it should also control the trigger of tasks in response to context changes.

This paper is organized in the following sections: Section 2 discusses how clinical activities/tasks were modeled; Section 3 discusses the architecture for programming and running tasks, briefly discussing target middleware for

 ISBN: 978-1-61208-100-7

pervasive environments upon which the current study was based, as well as changes made to adapt the middleware to the activities-oriented computing and clinical environment requirements; Section 4 introduces a use case about the tasks management system; Section 5 discusses the prototyping and evaluation of developed architecture; Section 6 describes some related works and eventually Section 7 makes final considerations about the study.

II. Task-Oriented Computing

Clinical activities, such as outpatient care, are processes that take place collaboratively, in a coordinated and distributed way, in a given space and can be aided by computational applications [1]. According to the Theory of Human Activity [13], human activities can be modeled with a subject (who practices an activity), an objective (what to do), an action (process to accomplish the objective) and operations (how the action is performed). From that theory we derive the concepts adopted to model tasks for the construction of the ClinicSpaces software architecture.

Thus, clinical activities were modeled in tasks (actions). For example, the activity "outpatient care" can be modeled as a mixed task that identifies the patient automatically, looks for his electronic medical records, reviews pending tests, etc., displays them to the physician, and makes it possible that the physician adds information during the visit. Thus, a task that will proactively help the physician during patient's visits is created. This task can also be linked to other tasks, so that the activity can be modeled in the most complete way as possible. For example, it could be connected to a task of tests request and/or treatment prescription [12].

Tasks can be customized according to how the user (physician) can perform his activities. The goal is to create abstractions so that the computational system can fit the real features of the clinical environment. Thus, in the same way as actions are composed of operations, tasks are composed of sub-tasks. Sub-tasks are pervasive applications that maintain a direct relation with computational artifacts (applications, services and resources) and make available the computational support to customized tasks, which together model the clinical activity to be performed. Sub-tasks are concerned to the applications of the pervasive health care electronic system (pEHS) [8], such as finding the electronic records, viewing tests and management systems of the pervasive environment, extended to tasks handling (Section 3) Both tasks and subtasks were modeled through specific ontologies [5], containing information regarding identification (user that created the task, clinical expertise, description of functionalities) and management (status, required resources, contexts supported).

So that, through the Tasks Edition Interface (a programming tool oriented to the end-user, see Figure 1), a task is programmed intuitively by the physician through clustering of sub-tasks and tasks, in a sequence that reflects the way

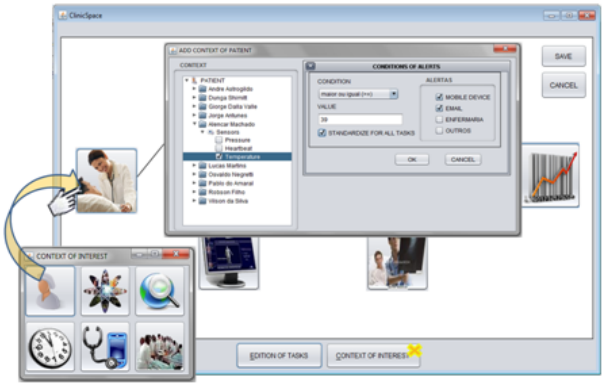


Figure 1. Task Edition Interface

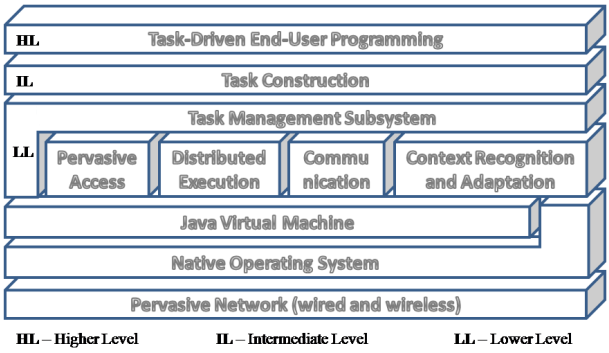


Figure 2. System architecture for tasks managing and programming

how the user performs his activity (customization). At the same time, customization can also be made at the sub-task level, with the configuration of specific parameters, which change the way how the sub-task processes its functionality.

III. Task Management

As we see, the ubiquitous systems bring new challenges to different areas. Among them is the development of a new middleware that, besides managing the pervasive environment, must manage users' daily tasks. Middleware must allow users to customize (program), run, stop, resume, schedule their tasks and control how they respond to context changes. Our challenge in this work, within the scope of the ClinicSpaces project, was to adapt a pervasive environment management middleware to manage the users' tasks as well (physicians), meeting the requirements of user-centric and daily activities support in the clinical environment [9].

The architecture designed for programming and management of tasks (see Figure 2), is organized in levels that reflect the system views:

- Higher level, composed of end-user (physician) who interacts with the tool to build and edit his tasks that would be triggered by changes in context (reactive). The ClinicSpaces interface is based on principles of the

 ISBN: 978-1-61208-100-7

Visual Programming paradigm, with a hybrid solution that uses graphic elements (icons and diagrams).
- Intermediate level, composed of mapping and conversion of tasks defined by the user in pervasive applications and by the management of tasks programmed by the middleware;
- Lower level, composed of the set of the middleware services for pervasive environment management and supports the execution of pervasive applications.

A. EXEHDA - middleware to pervasive environment management

Middleware EXEHDA - Execution Environment for Highly Distributed Applications [2] is used to manage the pervasive environment where tasks will be performed. This middleware aims at creating and managing a pervasive environment providing a virtual environment to the user, where applications have the style "follow-me" [7]. Thus, EXEHDA makes it possible to the user to access its computational environment regardless of place and time.

EXEHDA was designed to easily add new components. It is structured in a minimum kernel (required to boot) and services loaded on demand, which are organized in subsystems that manage: (a) distributed execution; (b) communication; (c) context recognition and adaptation; (d) pervasive access to resources and services. EXEHDA sub-systems are formed by services, and each service defines an interface and may be implemented in different ways, suitable to the types of devices that will be supported. Services provided by EXEHDA are customizable at the host level, and are determined by the execution profile, which defines the set of services to be activated and the parameters for their execution, besides associating each service to a specific implementation. This way, services can be "plugged" with no need to have the middleware's kernel modified [4].

A pervasive scenario, as a hospital, defines a personal virtual environment that should follow the user in his movement. This movement can involve both logical mobility (data, code and computation) and physical mobility (resources and devices in use). We refer to this feature of environment and applications as follow-me semantics.

The Pervasive Access Subsystem of EXEHDA is responsible for implementing the resources availability at anywhere, at anytime. To manage these functionalities, the middleware maintains information about user, applications, resources and services. Applications are not installed in the traditional way, meaning that the application's executable code is neither stored in nor managed by the user's virtual environment service. In fact, application installation consists only of copying the application's launching descriptor to the users virtual environment. The executable code for the application is still provided on-demand by the Application DataBase (ABD) service by the time the application starts to execute in a given device. Application profiles (resource descriptors and shared data) are stored in the Application Virtual Environment (AVE), which disappears when the application has finished its execution.

As a user physically moves (by carrying its current device - user mobility - or changing the device being used - terminal mobility), his currently running applications, in addition to the user's virtual environment, need to be continuously available to him, following the user's movement in the pervasive space.It is desirable that, when the system state changes, the middleware dynamically reallocate, reschedule and restructure the (logical and physical) resources available for the application. Application is managed by Distribution Execution subsystem which is responsible by lauching, migration and controling of execution through interaction with other subsystems. Application is on-demand installed by the ABD service in the target device. The executable code for the application is continually provided on-demand according to the Executor service.

The assembling of the context state information, which guides many of the middleware operations and also the application's adaptive behavior, is accomplished by the Context Recognition and Adaptation subsystem, through the cooperative operation of the services Monitor, Collector and Context Manager. The produced context state information feeds both functional (that modifies the code being executed) and non-functional (related to scheduling and resource allocation) adaptation processes, which are managed by the AdaptEngine and Scheduler services respectively. The adaptation model adopted is collaborative and it is reached by two forms: (i) adaptation commands, by explicit calls to some of the middleware's services, and (ii) adaptation policies implicitly guide middleware's operations. Adaptation policies are in the form of XML documents, deployed together with the application's code when it is installed in the BDA pervasive repository. Typically, adaptation policies are defined at development-time by the application designer.

With respect to communications, EXEHDA currently provides, through the services Dispatcher, WORB, and CCManager, three types of communication primitives, each addressing a distinct abstraction level. The Dispatcher service corresponds to the lower abstraction level, providing message-based communications. Message delivering is done through per-application channels, which may be configured to ensure several levels of protection for the data being transmitted. Protection levels range from data integrity, using digital signatures, to privacy through encryption. Additionally, the Dispatcher uses a checkpointing/recovery mechanism for the channels, which is activated when a planned disconnection is in course. This feature may or may not be activated by the upper communication layers depending on its particular demands. In order to make easier the development of distributed services, EXEHDA's also provides an intermediary solution for communications, based on remote method invocations, through the WORB service.

 ISBN: 978-1-61208-100-7

The programming model supported by WORB is similar to Java RMI, though it is turned to the pervasive environment while RMI is not. Specifically, WORB remote method invocations, differently from Java RMI, do not require the device to keep connected during the entire execution of the method in the remote node. At a higher level, the EXEHDA's CCManager provides tuples-space based communications. It builds on the WORB service, which already handles planned disconnections, providing applications with an anonymous and asynchronous communication model.

B. Adapting the EXEHDA to manage the personalized clinical task

EXEHDA was projected in modular way to easy the future adaptation. Due to the flexible features of EXEHDA, as the integration of new services, the adaptation of the middleware to the new tasks management was modeled as a new subsystem, named Sub-system of distributed task management (SDTM). The SDTM services shown in Figure 2 are: (i) Task Management Service, responsible for managing tasks execution; (ii) Tasks Access Service, responsible for the pervasive access to the user's task repository; (iii) Tasks Context Service, responsible to make available the context information relevant to tasks; (iv) Inference Service, responsible for the identification and activation of tasks based on context or schedule; (v) Active Tasks Service, responsible for the pervasive access to user's active tasks; (vi) Interception Service, responsible for connecting SDTM to a system of electronic records, which makes the access available to clinical applications.

The *Tasks Access Service* (TAS) role is to provide pervasive access to the tasks repository of each user, as well as to the repository of sub-tasks, which is unique for all the system, once they are not changeable. Through this service, the Task Management Service searches the list of tasks available to the user when his session starts, and the customizable ontological description [5] of tasks to be instantiated in pervasive applications and start to run. Besides, this service makes available access to information about tasks execution, which support the user's preference processing. The Tasks Access Service use the Pervasive Access Sub-system from EXEHDA to access the user's virtual environment.

The *Tasks Context Service* encapsulates, within objects used by tasks and sub-tasks, context information obtained by the Context Recognition Sub-system from EXEHDA. In our perspective, context is related to users, location, time and resources. Thus, context can be handled in a simpler way by the sub-tasks programmer and becomes more understandable from the user's point of view, who will access contexts that he considers useful for the accomplishment of tasks. This is necessary because the context provided by EXEHDA is in form of raw data. The goal of this service is to gather this information in form of objects with a simple API.

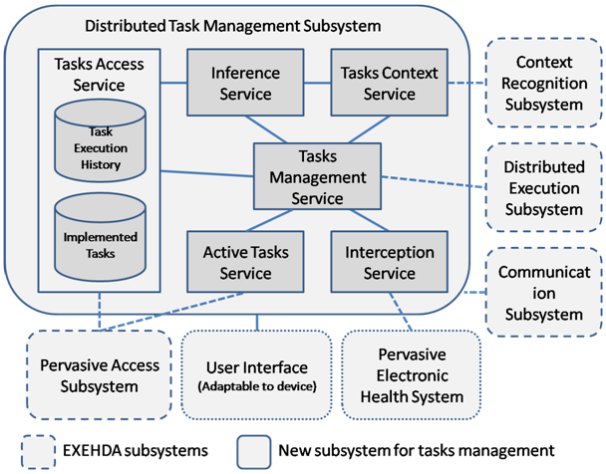


Figure 3. Architecture of the Tasks Distributed Management Sub-system

The *Inference service* processes historic information about the execution of tasks to infer upon the tasks activation, based on context change, in the user's preference in some contexts and scheduling. This way, the system proactivity is improved. This service is still under definition, and was not included in the tool pilot prototype.

The *Tasks Management Service* (TMS) role is to control the execution of user's tasks, enabling him to (i) trigger or schedule the execution of a given task; (ii) stop and later resume a task that was under execution; (iii) cancel a task that does not need to be continued. For that end, TMS makes an API available so that the user can control his tasks. Such API is encapsulated in the graphic interface adequate to the type of device and activity that is being modelled. Besides, TMS manages the migration of tasks, so that they can "follow" the user when he changes device.

The *Active Tasks Service* was created with the goal of keeping the active tasks of each user under a centralized control. For if such control was distributed (carried out by each TMS instance), it would require too much communication and processing to maintain tasks synchronized, making management complex and even unfeasible for some portable devices. Besides, a host is required where to the tasks should be migrated if the user is not using any device.

As a consequence, the Active Tasks Service works in an architecture client/server through HTTP requirements. A host (server) keeps information about tasks active (for each user) and clients make requests to obtain, add, exclude and update tasks. The Active Tasks Service was based on HTTP requests because there is already an infrastructure in middleware for such type of communication. Thus, the Active Tasks Service use the HTTP Service functionality from EXEHDA in the same way as services of the Pervasive Access Sub-system. A specific handler is used to convert service requests into HTTP requests. Therefore, the use of

 ISBN: 978-1-61208-100-7

HTTP communication is transparent from the perspective of other services that use the Active Tasks Service, which makes possible to change the implementation without changing coding of clients from such service.

The *Interception Service* is responsible for handling events generated by the Pervasive Health Care Electronic Register (under definition). In order to fulfill the objective of the project ClinicSpaces, i.e., to include the computational system in the clinical environment reducing its interference in the work of professionals, this service is aimed at intercepting the events of pEHS and process them, triggering applications that will help the physician handle the event, requiring the least possible interaction with the computational system. For example, if an event requires the immediate attention from the professional, the Interception Service may trigger the interruption of running tasks and start a task that help him to meet the event, for example, searching the patient's records (if this is the reason for the event).

The Interception Service can implement abstractions to make the use of pEHS applications easier, which makes possible to easily implement sub-tasks, considering that (i) the ClinicSpaces architecture can be seen as a management agent that facilitates the input of data in the healthcare system used; (ii) each sub-task is associated to at least one application of pEHS, which can be called to interact with the user or can be used only in second plan; and (iii) the system, through passing of parameters, can inform previously data for the application, minimizing the data input from the user.

IV. USE CASE

This case study comprised patients seen at ophthalmological emergency service of the University Hospital of Santa Maria (HUSM). Santa Maria is a tertiary and most important general hospital of the center region of the state of Rio Grande do Sul (Brazil) providing medical care to a population of 1.1 millions inhabitants from 47 cities. It is the biggest hospital, which provides ophthalmological emergency in the region.

In the same way, their patients attend the ophthalmological emergency service mainly being referred by general physicians, nurses or an ophthalmologist, or also by their own decision. There they are examined by an ophthalmologist, treated if necessary and sometimes referred for other center.

Data about patients were collected during interview and ophthalmological examination. When a patient had more than one diagnosis, only the most serious one was listed. Veracity of the emergency was categorized as *true* or *not-true* emergency. A true ocular emergency was considered if there were risk of decreasing or loss vision, as well as cases requiring immediate (same day) evaluation in either an emergency department or an ophthalmology outpatient department due the intensity of symptoms.

In this case we applied the middleware EXEHDA to control daily tasks of an ophthalmologist. This way, he will be able to create (personalize) a task that will help him in this activity, through of Task Edition Interface (see Figure 1). For example, if the ophthalmologist usually first check the latest information about the patient's medical records, he starts the task's construction with a sub-task to search the patient's medical records. The tasks construction tool adds, automatically, a sub-task to identify the patient using RFID for example, since it is necessary to identify which medical records should be accessed. Then, the physician adds a sub-task to view the information in his preferable form.

Continuing the task's composition, the ophthalmologist may add a sub-task to register the information about the patient's care. In addition, a task to request examinations could be added. It will register the request in patient's medical records and forward the request to the laboratory.

Finally, the ophthalmologist could add a task to prescribe the medication or treatment. This task will register the prescription in patient's medical records and will print it. When finished the task construction (personalization), it will be stored in the User Virtual Environment.

A. Task Execution

When the ophthalmologist examines a patient (human activity modeled as a set of tasks), the system interface is accessible and the tasks are available [2]. The patient's examination task is programmed (customized) in advance and triggered by the physician or by changes in context (patient's arrival detection). The Tasks Management Service (TMS) accesses the Tasks Context Service (TCS) to obtain the physician's identification, device configuration and other information required to instantiate the task. Processing the ontological description of the task, TMS finds sub-tasks that compose it and finds their code.

The first sub-task to be performed is the patient's identification. For that end, TCS uses sensors managed by the EXEHDA middleware, through of the context recognition subsystem [2][6], to identify the patient that will be examined. The information that returns is sent to other sub-tasks. Then, the sub-task that searches for the patient's information is triggered in the pEHS linked to the ClinicSpaces architecture. The information is then used as a parameter for the next sub-task, which displays the data from the records in the format the user selected. It is only from this moment that the interaction between ophthalmologist and task starts, because the execution of previous sub-tasks is transparent.

The system's interface checks which tasks are available to the user. These tasks – composed of other tasks and sub-tasks – were validated at the time of creation and customization of these tasks using TMS.

After the doctor reviews the information, he terminates the application (sub-task) and TMS triggers the next sub-task, which will store the ophthalmologist's notes in the electronic

 ISBN: 978-1-61208-100-7

records (pEHS). This sub-task calls a specific application of pEHS, and parameterizes it with the ophthalmologist's and patient's identification. At the end of this sub-task, the next – tests request – is triggered. This will be a task created by the user or obtained from the system. This task is independent, and can be performed in isolation.

The tests request task stores the ophthalmologist's request in the patient's records and send it to the lab, which is integrated to pEHS. At the end of this task, TMS triggers the treatment prescription task, which is stored in the patient's records and sends it to the closest printer, obtained through TCS.

The module decomposes and recomposes the tasks flow at the execution point. The TMS decomposes a task into sub-tasks. A task is represented by an XML file that links to the XML files of the sub-task. The decomposition is done when the system reads the XML files, instantiates objects (sub-tasks), sets the parameters, and executes them.

V. Prototyping and Evaluation

In order to validate the concepts discussed in this paper and evaluate the impact of task management in middleware and in execution of pervasive applications (tasks), a pilot prototype of the Sub-system of Distributed Task Management was developed. This prototype is composed by the Task Management Service, Tasks Access Service, and Active Tasks Service. All these services were were developed in J2EE, as well EXEHDA middleware and its applications. These services were integrated with the EXEHDA middleware and instantiated in two nodes: one base node, that is responsible for the cell's management; and one common node (client), representing a user's device.

The user's application, through API of the SDTM, searches the tasks available for user and the tasks that were initiated and do not terminated in previous session. As this information is located in base node, the local instances of the Tasks Access Service and Active Tasks Service generate requests for their remote instances. Therefore, the operations of these two services were monitored in experiments to determine the impact of them on the system, in terms of startup time of tasks and number of remote requests. On the other hand, the tasks execution management is done, locally, by the Tasks Management Service. Thus, this service was monitored in terms of extra processing to control the tasks.

The experimentations showed that the impact of new services, necessary for the tasks management, was minimal, both in terms of middleware as user's applications. Moreover, the evaluation of services indicated points in the prototype that can be improved, as the number of remote requests, that although they are acceptable, they can be reduced.

Therefore, in general, the evaluation of the experimentations showed that the proposal to extend a middleware for pervasive computing to manage the daily tasks of health

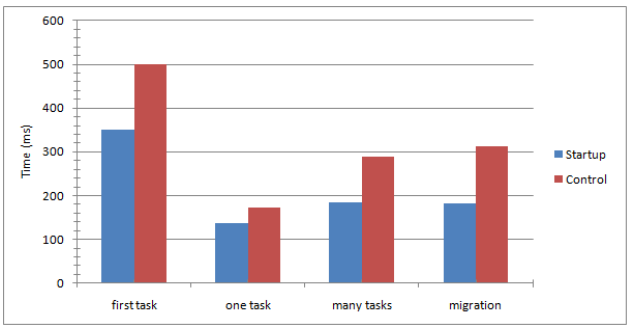


Figure 4. Startup and control time for different situations in tasks management

professionals is feasible and promising. The results were very satisfactory, since the prototype worked together with EXEHDA, performing their functions without generating overhead in the middleware.

As can be seen in Figure 4, the average time for initialization and control of tasks are below a millisecond. In the graph can be noted that the migration of tasks do not have significant influence on time for management. However, in the execution of the first task (after system boot), the management time is influenced for the startup of the middleware services. Moreover, the graph shown that even in the management of concurrent tasks (5 to 10 tasks), the overload of the system is minimal.

VI. Related Works

The ISAM [6] project, as well as Gaia [14], originated a middleware for pervasive environment management that integrates the premises of grid computing, mobile computing and context-aware computing. EXEHDA [2][7] middleware creates and manages a virtual environment for the user, in which applications are executed in a distributed and context-adaptable way. However, it does not approach the concepts of daily activities.

The concept of Task-based Computing was introduced by the Aura project [3] so that the middleware can manage the pervasive environment proactively in a way that the user can keep the continuity of his activities, at the same time as he moves from one place to the other. In this project, tasks are modeled as collections of services; the service description is used to find the necessary resources or reconfigure the middleware to run the task. The Aura system is automated and proactive, and it urges the user to perform his tasks according to predefined and preprogrammed settings, that is, it does not allow for the customization of tasks (services), which increases the interference of the system in the environment.

The Activity-based Computing [10] project brings a proposal for the use of Task Based Computing in health care environments. A framework was developed in this project that provides the required infrastructure to perform services

 ISBN: 978-1-61208-100-7

that will support the features inherent to the health care professionals activities. Thus, services can be initialized, suspended, stored, resumed in any device and at any time, as well as they can be sent to other users or shared among different users. The project aims also to allow developers of clinical applications to incorporate in its programs support to mobility, interruptions, concurrent activities and cooperation.

Some ideas taken from such projects had some influence in our work. We highlight the use of EXEHDA middleware and the Activity-based Computing project, which guided the definition of concepts related to tasks in the health care area. However, as one can observe in Figure 5, none of the projects focuses on the final user as the ClinicSpaces does. This project differential is that it enables ophthalmologist to customize tasks and balance the proactive execution with the control over the execution of tasks.

	Pervasive Computing		Task-based Computing		
	Mobility	Context	Automation	Customization	Control
Gaia	X	X			
ISAM	X	X			
Aura	X	X	X		X
ClinicSpaces	X	X	X	X	X

Figure 5. A comparison among ClinicSpaces' related works

VII. Conclusion

Pervasive computing promises ubiquitously support to users in accomplishing their tasks. Hardware and networking infrastructure to make the pervasive computing come true are increasingly becoming a reality. While this scenario represents an attractive computational environment, it poses three major challenges in the form of heterogeneity, dynamism, and execution context changes, all these are intensified because of user's mobility. Our project, named ClinicSpaces develops a pilot tool that will allow the physician to configure/program computational tasks that will help him in daily professional activities. Thus, we intend to minimize the avoidance of these professionals with relation to the use of computational systems which have a pre-established behavior.

ClinicSpaces' architecture is innovative in that the system can be customized according to the user's characteristics, in a way that professionals can use the computational system as if it were an assistant that would help him in his activities, instead of imposing the way how he should perform his work. In order to make such ideas real, the present study aimed at adapting a pervasive space management middleware to support tasks execution using EXEHDA, because it employs a lot of strategies in its services to allow the adaptation to the current state of the execution context, such as on-demand adaptive service loading and dynamic discovery and configuration.

The project is under development, it was implemented but not yet tested by physicians. The actual utility will be verified in a next field research. One of the contributions of this work is the service architecture required to adapt a pervasive environment management middleware to the management of tasks in this environment, fulfilling some requirements of activities-oriented computing to the final user. Another outcome from this work is the pilot prototype of an assistant (implemented as a middleware for pervasive environment management) that helps physicians in tasks management (daily activities). The next step of our research is to assess the usability of this solution in a real environment, after the insertion of the Pervasive Electronic Health Care System (pEHS) and this architecture.

Acknowledgements – This work has been partially supported by Brazilians Agencies FINEP/CNPq/MCT.

References

[1] A. Ranganathan and R.H. Campbell, *Supporting Tasks in a Programmable Smart Home*. In From Smart Homes to Smart Care, v. 15. Amsterdam: IOS Press, 3-10, 2005.

[2] A. Yamin, I. Augustin, L.C. Silva, R.A. Real, and C.F.R. Geyer, *EXEHDA: adaptive middleware for building a pervasive grid environment*. In Frontiers in Artificial Intelligence and Applications: Self-Organization and Autonomic Informatics (I), vol. 135, pp. 203-219. IOS Press, 2005.

[3] D. Garlan, P. Steenkiste, and B. Schmerl, *Project Aura: Toward Distraction-free Pervasive Computing*. In IEEE Pervasive Computing. New York, NY, 2002. pp. 22-31.

[4] G. Ferreira, I. Augustin, G.R. Librelotto, F.L. Silva, and A.C. Yamin, *Middleware for management of end-user programming of clinical activities in a pervasive environment*. In Proceedings of the 2009 Workshop on Middleware for Ubiquitous and Pervasive Systems. ACM New York, USA. 2009. pp 7-12.

[5] G.R. Librelotto, J.B. Gassen, M.C. Silveira, and L.O. Freitas, *OntoHealth - Um framework para o gerenciamento de ontologias em ambientes hospitalares pervasivos*. In: II Workshop on Pervasive and Ubiquitous Computing, 2008. pp. 31-42.

[6] I. Augustin, A.C. Yamin, L.C. Silva, R. Real, G. Frainer, G. Cavalheiro, and C. Geyer *ISAM: joing context-awareness and mobility to building pervasive applications*. In Mobile Computing Handbook. I. CRC Press, New York, NY, 2004.

[7] I. Augustin, A.C. Yamin, and L.C. Silva, *Building a Smart Environment at Large-Scale with a Pervasive Grid Middleware*. In Grid Computing Research Progress. Nova Science Publishers, Inc, 2008. pp. 182-186.

[8] J.B. Jorgensen and C. Bossen, *Executable use cases: requirements for a pervasive health care system*. IEEE Software. 21(2):34-41, 2004.

[9] J.E. Bardram, *Hospitals of the Future: Ubiquitous Computing support for Medical Work in Hospitals*. In Proceedings of 5th International Conference on Ubiquitous Computing, pp. 1-7. 2003.

 ISBN: 978-1-61208-100-7

[10] J.E. Bardram and H.B. Christensen, *Pervasive Computing Support for Hospitals: An overview of the Activity-Based Computing Project*. In IEEE Pervasive Computing, v. 6, issue 1, 44-51, 2007.

[11] J.E. Bardram and T.R. Hansen, *Context-Based Workplace Awareness*. In Computer Supported Cooperative Work. v. 19, issue 2, 105-138. Kluwer Academic Publishers. 2010.

[12] K.T. Unruh, M. Skeels, A. Civan-Hartzler, and W. Pratt, *Transforming clinic environments into information workspaces for patients*. In Conference on Human Factors in Computing Systems. SIGCHI: ACM Special Interest Group on Computer-Human Interaction. 2010. pp. 183-192.

[13] M. Kaenampornpan and E. O'Neill, *Integrating History and Activity Theory in Context Aware System Design*. In Proceedings of 1st International Workshop on Exploiting Context Histories in Smart Environments, Munich, 2005.

[14] M. Roman, M. Román, C. Hess, R. Cerqueira, R.H. Campbell, and K. Nahrstedt, *Gaia: a Middleware Infrastructure to Enable Active Spaces*. In IEEE Pervasive Computing. New York. 2002. pp. 74-83.

[15] U. Varshney, *Pervasive Healthcare*. IEEE Computer, 36(12): 138-140, 2003.

 ISBN: 978-1-61208-100-7

The Economic Impact of IPTV Deployment in the European Countries:

An Input-Output Approach

Ibrahim Kholilul Rohman, Erik Bohlin
Division of Technology and Society
Department of Technology, Management and Economics,
Chalmers University of Technology, Gothenburg, Sweden
ibrahim.rohman@chalmers.se, erik.bohlin@chalmers.se

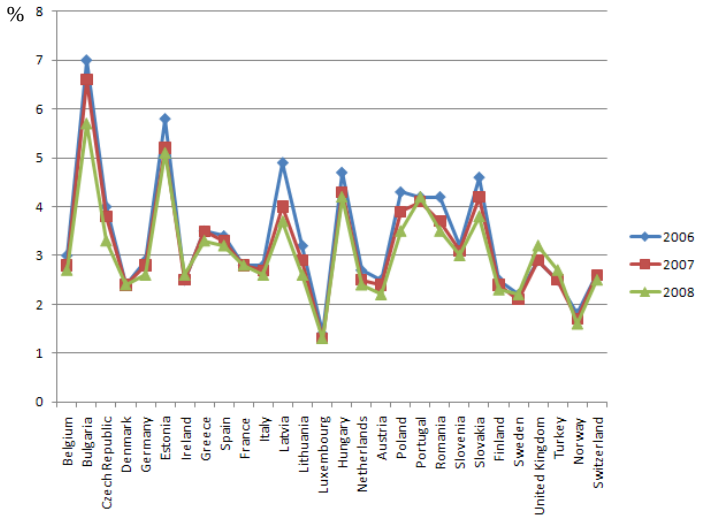
Abstract*-- A double-edged sword phenomenon is illustrated by the consequences of the rapid penetration rate of information and communication technology (ICT) devices. Besides the success story of these devices becoming ubiquitous, there is also a visible decline in performance of the ICT industry in financial respects. Due to more intense competition and market saturation, the players in this industry are now facing limited revenue sources. Among other things, traditional TV industry offering broadcasting services is suffering a significant drop in viewers and advertising revenue because of the massive substitution effect of the vast penetration rate of Internet and broadband. It is then important to see how the innovation of ICT devices can create possible alternative sources of revenue. Vibrant ICT devices, which combine television (video), telecommunication (audio) and data (Internet) in so-called triple play may enable operators to obtain additional revenue. IPTV (the TV which is transmitted to the Internet protocol) is seen as a strong new triple-play device which can support ICT as well as being a precursor of further economic impact, especially in driving output multipliers and Gross Domestic Product (GDP). This study is aimed at investigating the economic impact of IPTV deployment in the European countries, using the Input-Output table. The method enables us to estimate the economic multiplier for each Euro investment in IPTV deployment as well as to estimate the contribution to the GDP from two main sources: the production phase, when the deployment is implemented by installing fiber and network to the household, and the diffusion phase, where the consumption of IPTV services increases after the completion of the investment project. Among fourteen European countries investigated, the study reveals that Sweden is the country which enjoys the highest level of impact due to the construction activities, while Austria gets the larger portion of the multiplier from the diffusion side.

Keywords- IPTV; investment; input-output

I. INTRODUCTION

There is evidence that the ICT industries have been facing very difficult conditions in the last couple of years, especially in plain old traditional services. For instance, British Telecom (BT) retail revenue from traditional services decreased by 3% in 2008. In cellular industry, the Belgian incumbent operator, Belgacom, suffered similarly when its revenue from voice subscription dropped by 2% in 2007, with another loss of about 11% from voice traffic. In most Organization for Economic Co-operation and Development (OECD) countries, the market maturity and fall in retail price are seen as the drivers for this situation [1].

The gradual decline of communication consumption also indicates why this recession afflicts most of the ICT industry in OECD countries. The annual data on the ratio of expenditure on communication to GDP shows that the ratio has been declining gradually. Even though the ratio dropped only slightly from 3% (2006) to 2.9% (2008) throughout the EU (15 countries) as shown in the following Fig. 1, it has dropped substantially for major and leading ICT countries like Germany, the Netherlands, Italy, Norway and Finland. For instance, Germany dropped from 2.9% (2006) to 2.6% (2008), and the Netherlands has continued declining from 2.7% (2006) to 2.4% (2008).

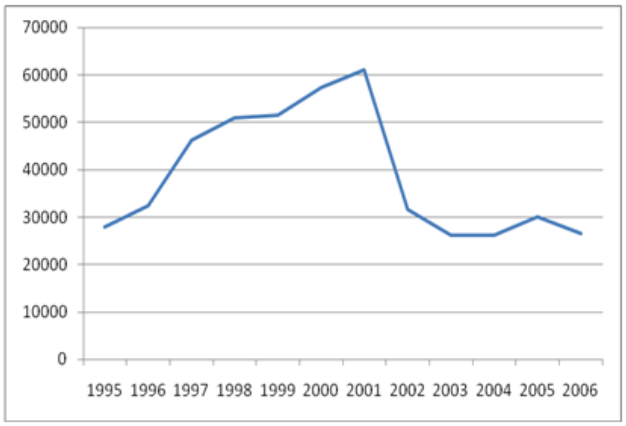


Source: Eurostat

Figure 1. Communication expenditure as a percentage of GDP (%)

As a result, it is not surprising that investment in this sector is also affected by the recession. From the data comprising 33 European countries, the average annual growth of telecommunication investment during 2000-2006 was -6% compared to 16.2% during the period 1995-2000. Fig. 2 shows the decline of telecommunication investment which began massively in 2001.

 ISBN: 978-1-61208-100-7



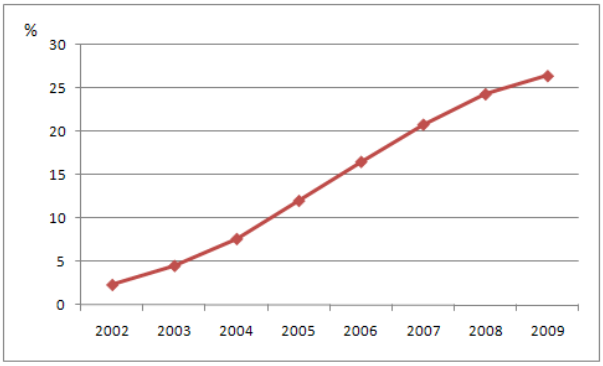
Source: Eurostat

Figure 2. Telecommunication investment in European countries (MEUR)

This situation has to be anticipated by the European countries, since the slowdown in investment in the information and communication sector may be followed by the falling in the productivity level. This is, moreover, supported by the reason that the increase of investment in ICT capital (and growth of human capital) contributed significantly to labor productivity growth in market services across all European countries and the US [2]. Therefore, a decline of investment in the ICT sector might affect the rate of productivity in the EU economy. It has been proven by the fact that since the mid-1990s, the European Union has experienced a slower growth of productivity where, at the same time, the United States is significantly boosted. The steady growth of productivity in the United States is indicated to be due to a combination of high levels of investment in a rapidly progressing ICT sector, especially during the second half of the 1990s. It is also followed by a rapid productivity growth in the market services sector during the first half of the 2000s [3].

This phenomenon, especially in the broadcasting industry, has actually been predicted before. The significant drops in advertising revenue together with the delay in the rollout of digital broadcasting are among the reasons why the industry has faced obstacles recently [4]. In the US, as a comparison, broadcasters are asking the audience to view each show "live" to ensure the audience to the advertisers and consequently get them to finance subsequent episodes of the show [5]. From the surveys conducted in some countries, it is discovered that the largest channels in each country are suffering a decline in their ratings. Although the aggregate revenue still increases across Europe during 2006-2008, the public broadcasting sector has seen a drop of more than 4 percentage points in its total market share, while the commercial sector (both radio and TV financed by advertising) has grown modestly [6]. It is also predicted that the advertising revenues of traditional channels are not likely to grow significantly over the next decade. Using two different econometric models, it is estimated that the gradual decline is around 0.2–0.5% [7].

Thus, there is a need to grasp any possibility for industry players to create new devices which synergize ICT function, especially utilizing the fast growth of current ICT penetration rate. Compared to other devices (e.g. cellular, internet), broadband is still relatively in the stage of further diffusion process in European countries where the broadcasting industry can be supported. Fig. 3 shows the impressive growth of broadband penetration in the European countries.



Source: Eurostat

Figure 3. Average European countries' broadband penetration rate per 100 inhabitants

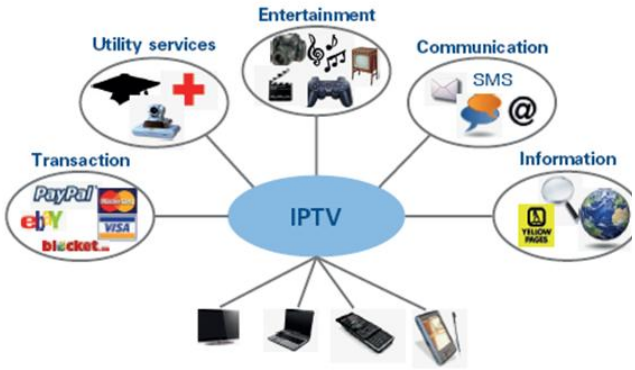
To operationalize this strategy, there are seven pillars with which broadband creates a significant impact in the economy: E-health, E-government, E-environment, E-business, E-employment, E-science and E-agriculture [8]. Thus, combining the broadband with television and telecommunication attached to these functions might generate an additional niche market in each European country. To address this issue, IPTV is seen as part of triple-play devices that might be an alternative product to provide a variety of services.

This study aims at investigating the economic impact of IPTV investment in the selected European countries. To answer the research question on how much will the IPTV investment contributes on the European economy, the study is presented in the following sections: Section 2 presents the nature of IPTV and variety of services offered to the customer. The mechanics of the Input-Output (IO) table as the main methodology is discussed in Section 3. Some previous analysis is presented in Section 4. The data used in this study is elaborated in Section 5. The results of the study are shown in Section 6. Section 7 concludes the paper.

II. THE IPTV

The Internet Protocol TV (IPTV) is viewed over a fixed broadband connection (DSL or Fiber to the home, FTTH) with a standard telecommunication set. The services are offered over a closed content distribution network whose common services cover TV broadcasting and stored video on demand (VOD), and the personal video recorder [9]. Additionally, not only can IPTV platform support a range of digital utility services, such as e-health, e-learning, e-working and home security – in contrast to the traditional

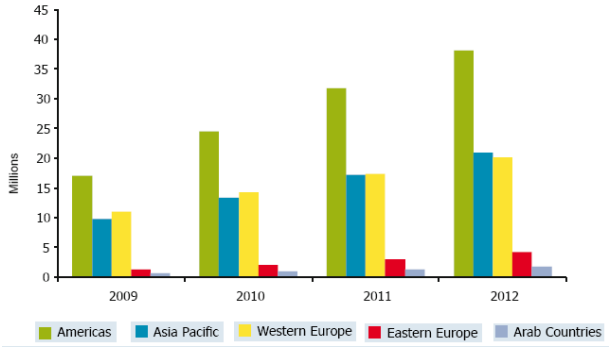
television, it also provides more control and choice for the customer [10]. Fig. 4 explains how the IPTV comprises a variety of services for the end users.



Source: Arthur D. Little (2009)

Figure 4. Variety of IPTV services

The development of IPTV in Europe shows positive growth. By the end of 2008, the total of IPTV subscribers had reached 21.7 million, which is an increase of 63% compared to the end of 2007 [11]. Additionally, it is predicted that the number of subscribers will continuously increase until the end of 2012. Even though the domination will be still in America, Western European countries are seen as a substantial market which will developed even further [12]. Fig. 5 shows the forecasted numbers of subscribers up to 2012 on major continents.



Source: ITU (2009)

Figure 5. Forecast number of IPTV subscribers

Fig. 6 forecasts there will be more than 64 million IPTV subscribers worldwide by the end of 2012, with the European market representing almost 38% of total subscribers.

III. The Input-Output Model

The input-output table depicts the transaction flow across sectors, where each sector produces a certain output and consumes input from another sector at the same time. The table consists of three main quadrants. The first quadrant describes the inter-linkage between sectors in a so-called intermediate transaction, while the quadrants II and III are the final demand and primary input respectively [13]. The table is shown in the following Fig. 6.

Intermediate transaction Intermediate demand/ Intermediate inputs I	Final Demand II	Total Output
III Primary Input Value Added		
Total Input		

Figure 6. Input-Output (IO) table

The flow of transaction in the table can be explained in the following equation (1). Assumed the economy consist of 4 sectors.

$$\begin{aligned} x_{11} + x_{12} + x_{13} + x_{14} + C_1 &= x_1 \\ x_{21} + x_{22} + x_{23} + x_{24} + C_2 &= x_2 \\ x_{31} + x_{32} + x_{33} + x_{34} + C_3 &= x_3 \\ x_{41} + x_{42} + x_{43} + x_{44} + C_4 &= x_4 \end{aligned} \quad (1)$$

where x_{ij} denotes the output from sector i which is used by sector j as an intermediate input (the input from sector i which is used for further production process in sector j). Moreover, c_i refers to total final demand of sector i and x_i refers to total output of sector i.

Introducing matrix notation, we can modify equation (1) to obtain the following matrix:

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_4 \end{pmatrix}; c = \begin{pmatrix} c_1 \\ \vdots \\ c_4 \end{pmatrix} \quad (2)$$

Equation (2) is the matrix form of equation (1), where x denotes the column matrix for output and c is the column matrix for the final demand.

$$I = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}; A = \begin{bmatrix} a_{11} & \cdots & a_{14} \\ \vdots & \ddots & \vdots \\ a_{41} & \cdots & a_{44} \end{bmatrix} \quad (3)$$

Matrix (3) consists of two parts; the left-hand side is the identity matrix, a diagonal matrix whose off-diagonals are zero. Whereas A is the technology matrix which consists of the ratio of the intermediate demand to the total output, $\frac{x_{ij}}{x}$.

By combining (1), (2) and (3), the equation can be modified as follows:

$$\boldsymbol{Ax} + \boldsymbol{c} = \boldsymbol{x} \quad (4)$$

$$(\boldsymbol{I} - \boldsymbol{A}) = \boldsymbol{c}$$

$$\boldsymbol{x} = (\boldsymbol{I} - \boldsymbol{A})^{-1}\boldsymbol{c}$$

From (4), the multiplier is defined as the inverse Leontief matrix, $(I - A)^{-1}$[13]. It measures the ratio of output changes in the equilibrium as the result of the change in the final demand. Therefore, the output multiplier measures total change throughout the economy from a unit change in final demand. The final demand itself might come from private consumption, government expenditure, investment and export. As the consequence of production linkages, a change of output will be larger than a change in the final demand. For instance, if the final demand of the ICT sector (e.g. additional purchasing of personal computers) increases by 10 EUR, the output in the economy will grow by more than 10 EUR or as much as the multiplier coefficient of this sectors. (Additional purchasing of personal computers will induce packaging services and transportation, for instance).

There is a reason that the firm level of analysis is a more appropriate approach to investigate the productivity impact from such an ICT investment. However, relying on meso studies can also lead one into the trap of a productivity paradox where the contribution of ICT is not absorbed in statistical reports or a so called Solow computer paradox [14]. Therefore, applying the Input-Output (IO) analysis in this study enables investigation at sector level, which has a direct link to firm level as well as to macro level. The intermediate transaction in quadrant I consists of the data gathered from an industry survey [13], [15], [16]. Moreover, the relation between the IO and the macro variable is very much straightforward. The primary inputs in quadrant II reflect the measurement of the Gross Domestic Product (GDP) by the income approach, comprising wages, salary, profit, etc. In addition, the final demand in quadrant II shows the calculation of the GDP as the sum of consumption, investment, government spending, and export and (minus) import [14].

Apart from that, the reason for using the IO method is also supported given the difficulties when measuring the indirect impact of ICT investment [17]. In this regards, IO has strong ability for capturing the direct and indirect impact from such an investment outlays [16]. Consequently, employing the IO method enables the estimation from both production and diffusion process of ICT sectors investment [18].

IV. PREVIOUS STUDIES

There have been vast investigations of the impact of ICT on the economy using econometrics or the IO method. Applying the panel data of US municipalities, it shows that the broadband will generate economic impact through creating employment, increasing housing rent, industrial mix and stimulating business establishment [19]. It is also estimated that the contribution of broadband to employment and productivity is visible in Latin America. Employing the IO method, it is suggested that to fulfill the increasing demand for broadband requires an additional 41% lines, which will contribute 378,000 more new jobs [20]. In the United States, the contribution of broadband to job creation is about 128,000 new jobs whereby each job will cost 50,000 US$. This result is determined from both network construction and network externalities, though the latter is less consistent due to uncertainty surrounding that impact [21]

A study in Indian economy concludes that the economics of information technology in the country is mainly supported by domestic demand. Therefore, to increase the performance of the Indian economy to be able to contribute on the GDP, it is important to boost the contribution from export; thus it requires stronger linkage between telecommunication and infrastructure facilities (power, water supply, and transportation). It is also found that private and government expenditure contributes respectively 32% and 26% while the investment contributes only 13%. In addition, among the export drivers, communication equipment and electronics equipment are the most dominant sector [22]. Similar study found a significant contribution of the ICT sector to the Singapore economy. It is estimated that each 10% increase in information input price during 1995-2000 will generate 0.84% increase in GDP growth which is twice the figure for the 5-year period before. Moreover, there was a shift from the export driver in the first half of 1990 to the domestic demand during the second half of the period [18].

Among the previous analyses using the IO method, this study is the first to investigate a comparison between several countries. It is explained that the productive use of ICTs requires organizational and working practice changes, and depends on contextual factors such as transport infrastructure, cultural values, and the routines organizing everyday life [23]. Therefore, a comparative study is a more appropriate way to see different characteristics of ICT utilization among countries.

V. THE DATA

This study will use the data of 59-Input Output Table from selected European countries. To attain the goal of the study, the analysis uses the domestic transaction of the IO table. The available years for each country are shown in the following Table 1.

 ISBN: 978-1-61208-100-7

TABLE I. IO TABLE FOR SELECTED EUROPEAN COUNTRIES

No	Country	IO availability
1	Austria	1995, 2000, 2005
2	Belgium	1995, 2000, 2001
3	Denmark	2001, 2002, 2003, 2004, 2005
4	France	2001, 2002, 2003, 2004, 2005, 2006
5	Germany	2001, 2002, 2003, 2004, 2005, 2006
6	Ireland	1998, 2000, 2005
7	Italy	2000, 2005
8	Netherland	2000, 2001, 2002, 2003, 2004, 2006
9	Poland	1995, 2002, 2003, 2004, 2005, 2006
10	Portugal	1995, 1999, 2005
11	Spain	1995, 2000, 2005
12	Sweden	1995, 2000, 2005
13	United Kingdom	1995

Source of data: Eurostat

As has been discussed, the investigation of the impact of the ICT sector should be made with two sources of growth: the production and diffusion [17], [24]. The production phase refers to the investment and infrastructure development of the IPTV whereas the diffsion is investigated in relation to induce income as the consumption increases from households. Thus, the calculation of the multiplier is done as shown in the following steps in Fig. 7.

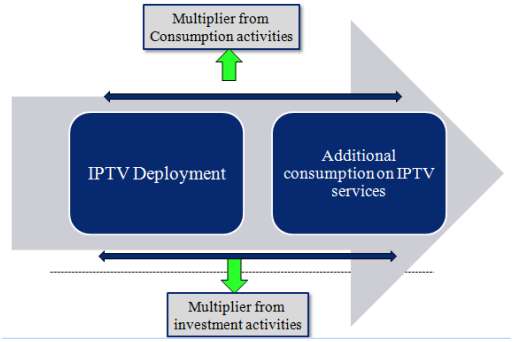


Figure 7. Two sources of output multiplier

The next important question to be addressed is how to define and match the IPTV deployment with the existing IO table. The following flow chart explains how this study is conducted.

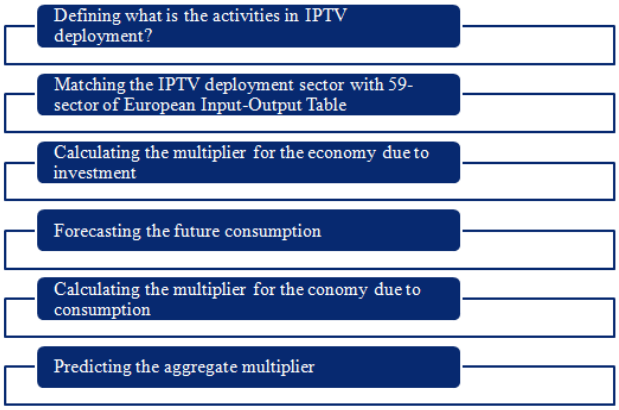


Figure 8. Flow of studies

The IPTV investment is established especially to enable consumers to use HDTV and ultimately 3DTV; hence it requires an advanced penetration of infrastructure. A high-speed large-scale rollout of fiber networks is to be connected to the households [9]. Therefore, it is assumed that assessing this criterion with the template of the IO table, the investment activities are grouped into sector number 34 (Construction). In addition, based on the Statistical Classification of Economics Activities (SCEA), the appropriate sub-sector for the investment activities is the installation of electrical wiring and fitting (SCEA code 45.31). Moreover, the diffusion impact which is measured from consumption after the finishing of investment activities is grouped into the telecommunication sector (sector number 43).

Furthermore, the value of investment in this study follows the assumptions as follows [9]:

(i) Investment is implemented throughout European countries to enable people to access the IPTV services.

(ii) The annual investment is 10 BEUR yearly for all European countries. The budget for each household is around 1150-1700 EUR/household.

(iii) Investment cost is assumed to decrease by 2% per year due to more efficient fiber rollout techniques.

(iv) Investment for each country is proportionally distributed, based on the number of households.

Applying these scenarios, the value of investment for each country can be shown in the following Fig. 9.

 ISBN: 978-1-61208-100-7

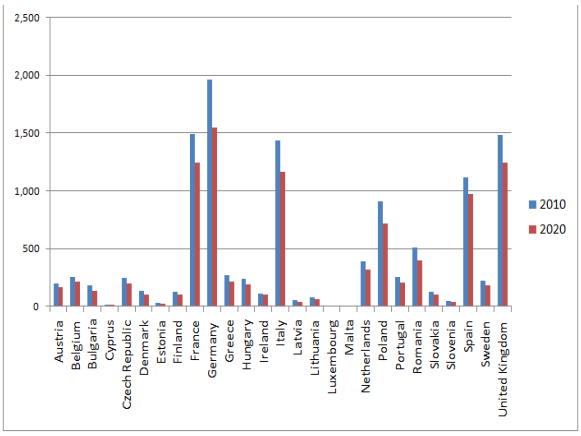


Figure 9. Value of investment in the IPTV deployment (MEUR)

Weighted by the number of households, the four countries which have the largest investment in the IPTV deployment are Germany, France, the UK and Italy.

VI. RESULTS

The first analysis in this study is explaining the multiplier effect of IPTV investment on economy. Using equation (4), the results of multiplier calculation are presented in the following Table 2.

TABLE II. OUTPUT MULTIPLIER FROM INVESTMENT AND FRACTION TO GDP

No	Country	Multiplier	Fraction to GDP
1	Austria	2.840	0.437
2	Belgium	1.529	0.312
3	Denmark	1.854	0.349
4	Finland	1.846	0.417
5	France	1.898	0.428
6	Germany	1.887	0.366
7	Ireland	2.421	0.269
8	Italy	2.551	0.505
9	Netherland	1.977	0.388
10	Poland	2.023	0.669
11	Portugal	2.175	0.303
12	Spain	1.552	0.269
13	Sweden	3.010	0.510
14	United Kingdom	1.898	0.385

Table 2 shows that Sweden, Austria and Italy are the European countries which have the highest output multiplier. Additionally, the fraction of GDP is measured by taking the ratio of Net Value Added to Total Output. In this regard, Sweden and Poland are the two countries which have the highest fraction of GDP over output.

Furthermore, the multiplier from the diffusion (consumption) of IPTV industry is shown below:

TABLE III. OUTPUT MULTIPLIER FROM CONSUMPTION AND FRACTION TO GDP

No	Country	Multiplier	Fraction to GDP
1	Austria	2.743	0.522
2	Belgium	2.533	0.554
3	Denmark	1.791	0.388
4	Finland	1.427	0.636
5	France	1.651	0.579
6	Germany	1.505	0.582
7	Ireland	1.889	0.312
8	Italy	1.744	0.567
9	Netherland	1.762	0.484
10	Poland	1.451	0.509
11	Portugal	2.014	0.379
12	Spain	1.889	0.349
13	Sweden	2.177	0.340
14	United Kingdom	1.651	0.618

Unlike the results from investment activities, the consumption activities on table 3 show that Austria and Belgium are the countries which enjoy the highest multiplier among European countries during the diffusion process.

The next analysis is the aggregation of multiplier on the European level. It is assumed that each European country can be analyzed separately and faces the market independently. Thus, the aggregation at regional level is done by applying the scenario that the multiplier is proportional to the number of households and value of the GDP. The result can be seen in the following Fig. 10.

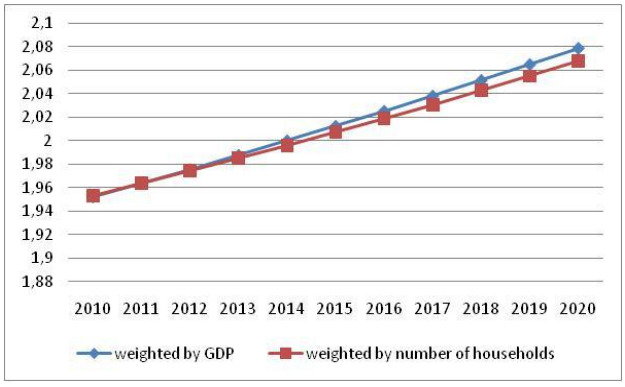


Figure 10. Aggregation of multiplier

From Fig. 10, it is quite interesting to see that both approaches come up with a similar estimate. The multiplier for European regions ranges around 1.95-2.08. It means that

 ISBN: 978-1-61208-100-7

every 1 EUR of additional investments in IPTV deployment will yield 1.95-2.08 EUR for the next 10 years.

The last step in this analysis estimates the impact of output creation from the investment cost projected in Fig. 10 and the diffusion. Having found that the multiplier ranges around 2, the impact of additional investment during the project period is forecasted in the following Fig. 12. In this study, scenario 1 is measured by using the GDP as the weighted index, while scenario 2 uses the number of households.

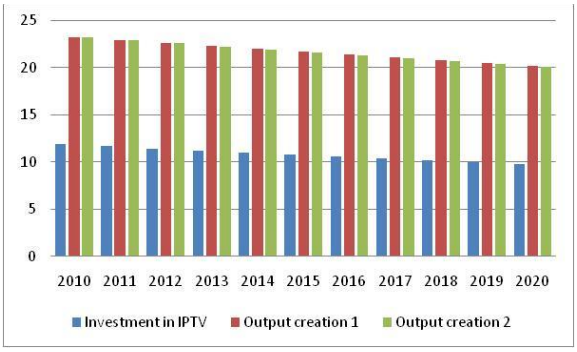


Figure 11. The output impact of IPTV deployment investment (BEUR)

From Fig. 11, it is estimated that the output impact will be smaller as a result of the decreasing value of investment over time. The impact on the GDP is shown in the following Fig. 12.

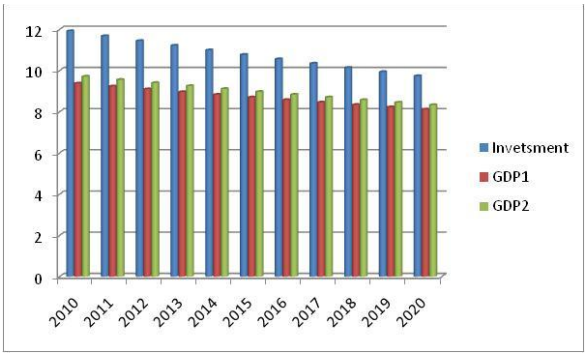


Figure 12. The impact of IPTV deployment investment on the GDP (BEUR)

As indicated by the fraction of the GDP to output, Fig. 12 shows that the impact of investment to output will be smaller. This is due to the fact that most OECD countries have an open economy, which is indicated by the large volume of trade (export and import). As a result, the fraction of GDP will be around 70% of the total investment.

TABLE IV. OUTPUT MULTIPLIER FROM CONSUMPTION TO GDP

Year	by GDP	by households
2010	1.712	1.699
2011	1.715	1.702
2012	1.718	1.705
2013	1.721	1.709
2014	1.725	1.712
2015	1.729	1.716
2016	1.733	1.720
2017	1.737	1.725
2018	1.742	1.729
2019	1.747	1.734
2020	1.752	1.739

Table 4 depicts the output multiplier from the consumption side. It is projected that each 1 EUR additional consumption of IPTV services will generate additional output around 1.7 EUR. Moreover, the contribution to GDP is explained in the following Fig. 13.

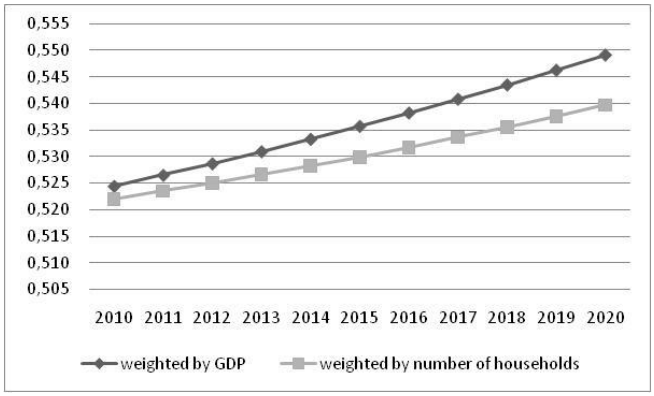


Figure 13. The fraction of IPTV consmption on the GDP (BEUR)

From Fig. 13 it can be seen that the contribution of IPTV consumption to GDP is around 0.52-0.55 from output. If the consumption of IPTV services increases by 100 EUR, it will generate output as large as the multiplier 170 EUR. Additionally, from the amount of output, 55% (about 93.5 EUR) will be directly formed as value added or GDP.

VII. CONCLUSION

The study draws the following important conclusions:

(1) The impact of IPTV deployment contributes to both investment activities and additional consumption activities (as the project ended).

(2) The contribution in generating output multiplier varies across countries. Sweden has to be considered as the largest contributor in terms of investment activities, while

 ISBN: 978-1-61208-100-7

Austria and Belgium are among the largest in terms of the output from consumption.

(3) In aggregation, the IPTV project has a multiplier of 1.9-2.1 in all the European countries. At the same time, the GDP contribution will be around 0.40-0.41. A lower or higher fraction of GDP is found in consumption activities.

It is then recommended that IPTV development has to be supported by the collaboration between the broadcasting and manufacturing industry player where both are capable of providing services to customer in the type of supply side network externalities (hardware-software relationship) [20].

The limitation of this study is related to the aggregation problem in the IO table. Two sources determine the level of aggregation: (1) the problem being considered, of whether or not it is important to distinguish the sector until the detail level; (2) computational expense and availability of the data. Consequently, when deciding that the IPTV rollout is classified as the construction sector, it might cause crude multiplier estimation since conventional construction, for instance, needs stone and cement in a higher proportion, which is not the case in fiber installation. Thus, the aggregation bias might weaken this result [13].

ACKNOWLEDGMENT

The authors are very grateful to Ericsson and Arthur D. Little for scenario data and the discussion. Needless to say that the authors are responsible for all remaining errors

References

[1] OECD, "OECD Communication Outlook 2009". OECD, 2009 Available at : www.oecd.org/sti/telecom/outlook [Accessed: July 10, 2010]

[2] R. Inklaar and M.P. Timmer, "Market services productivity across Europe and the US". Economic Policy, January 2008. Printed in Great Britain, CEPR, CES, MSH, 2008.

[3] B. van Ark, M. O'Mahonny, and M.P. Timmer, "The Productivity Gap between Europe and the United States: Trends and Causes**".** Journal of Economic Perspectives, Volume 22, Number 1, Winter 2008, pp. 25–44. Available at: http://pages.stern.nyu.edu/~cedmond/ge08/van%20Ark%20et%20al%20JEP.pdf [Accessed: July 10, 2010]

[4] L. van der Meulen, "Impending crisis in advert funded television calls for change in European Media Policy", European Platform of Regulatory Authorities. Available at: http://www.epra.org/content/english/news/speechcvdm.doc [Accessed: July 12, 2010]

[5] International Telecommunication Union (ITU), "Information Society: Statistical Profile Europe", 2009. Available at: http://www.itu.int/dms_pub/itu-d/opb/ind/D-IND-RPM.EUR-2009-R1-PDF-E.pdf [Accessed: July 10, 2010]

[6] Open Society Institute, "Television across Europe: more channels, less independence, Overview". Available at: http://www.soros.org/initiatives/media/articles_publications/publications/television_20090313/2overview_20080429.pdf [Accessed: July 12, 2010]

[7] OFCOM, "Economic analysis of the TV Advertising Market". Available at: http://www.ofcom.org.uk/research/tv/reports/tvadvmarket.pdf [Accessed: July 14, 2010]

[8] S. Marine, "Broadband for Economic and Social Development", ITU Regional Workshop on "Broadband Access Technologies". Damascus, 19-20 November 2008 Available at : http://www.itu.int/ITU-D/arb/ARO_2008_work/Broadband/Documents/Doc3-roadband%20for%20Economic%20and%20Social%20Development.pdf [Accessed: July 9, 2010]

[9] Arthur D. Little, "IPTV-Socioeconomics accelerator held back: Assessing the legislative and regulatory environment for IPTV", 2009. Available at: http://www.ericsson.com/campaign/televisionary/#/regulation/contributions/b58416f2_42ae_466e_bf38_c93ab5b421a5 [Accessed: July 5, 2010]

[10] E. Prosperetti, G.Tripaldi, and V.V.Comandini, "IPTV missed expectation. Can regulation do the trick?" III Annual Conference Bocconi University Milan, November 2007. Available: http://guidotripaldi.typepad.com/documents/IPTV_missed_expectations_SIDE07.pdf [Accessed: July 13, 2010]

[11] The Broadband Forum, "IPTV subscribers top ten million", 2009. Available at: http://www.broadband-forum.org/news/download/pressreleeases/2009/YE08_Europe.pdf [Accessed: July 14, 2010]

[12] International Telecommunication Union (ITU), "IPTV-Market, Regulatory trends and policy options in Europe", October 2006. Available at: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4451697&isnumber=4451683 [Accessed: July 7, 2010]

[13] R.E. Miller and P.D. Blair, "Input Output Analysis: Foundation and Extensions", 2nd edition, Cambridge University Press, 2009, p.160.

[14] E. Brynjolfsson, "The Contribution of Information Technology to Consumer Welfare", Information Systems Research, Vol. 7(3), 1996, pp. 281-300

[15] United Nations, "Handbook of Input-Output table Compilation and Analysis", New York: United Nation, 1999.

[16] C-S. Yan, "Introduction to Input-Output Economics.", New York : Holt, Reinhart and Winston, 1968.

[17] H.C. Lucas, "Information technology and the productivity paradox", New York : Oxford University Press, 1999.

[18] T.M. Heng and S.M.Thangavelu, "Singapore Information sector: A study using Input-Output table", Singapore Center for Applied and Policy Economics Working paper Series, No. 2006, 15 November 2006. Available: http://www.fas.nus.edu.sg/ecs/pub/wp-scape/0615.pdf [Accessed: July 10, 2010]

[19] S.E. Gillet, M.A. Sirbu, "Measuring Economic Impact of Broadband Deployment", National Technical Assistance, Training, Research, and Evaluation Project #99-07-13829. http://www.eda.gov/ImageCache/EDAPublic/documents/pdfdocs2006/mitcmubbimpactreport_2epdf/v1/mitcmubbimpactreport.pdf [Accessed: July 6, 2010]

[20] R.L. Katz, "Estimating broadband demand and its economic impact in Latin America", Proceedings of the 3rd ACORN-REDECOM Conference, Mexico City, May 22-23, 2009. Available: http://www.acorn-redecom.org/papers/RaulKatz.pdf [Accessed: July 8, 2010]

[21] R.L. Katz and S. Suter, "Estimating the economic impact of the broadband stimulus plan".
Available:
http://www.elinoam.com/raulkatz/Dr_Raul_Katz_-_BB_Stimulus_Working_Paper.pdf [Accessed: July 4, 2010]

[22] S. Roy, T. Das, and D. Chakraborty, "A study on the Indian Information sector: an experiment with Input-Output techniques", Economic System Research, Vol. 14, No. 2, 2002.

[23] I.Tumomi, "Realising the productivity potential of ICTs". Institute of Perspective Technology Studies (IPTS) Report, Issue 85.

[24] H.R. Varian, J. Farrel., and C. Saphiro, "The economics of information technology: an introduction". Cambridge University Press, 2004.

 ISBN: 978-1-61208-100-7

Design and Implementation of Edge Detection and Contrast Enhancement Algorithms Using Pulse-Domain Techniques

Fatemeh Taherian† Davud Asemani‡ Elham Kermani‡

† Dep. Of Electronic Eng., Islamic Azad University, Tehran Central Branch, Tehran, Iran, F.taherian@iauctb.ac.ir

‡ Electrical Eng. Faculty, K.N. Toosi Univ. of Technology, Tehran, Iran
Asemani@eetd.kntu.ac.ir, Kermani@ee.kntu.ac.ir

***Abstract*—Image pulse sensors provide pixels information with a series of pulse train considering Pulse Frequency Modulation (PFM). PFM sensors are often used in vision chips considering related advantages. In this paper, edge detection and contrast enhancement algorithms are analyzed and simulated using pulse domain techniques of suppression and promotion. Comparing with classical methods, pulse-domain-based algorithms show a better performance as well as simpler implementation in circuit level. Designing and implementing the pulse-domain algorithms on FPGA, a simple and fast realization of image processing methods for edge detection and contrast enhancement is presented. The integration of proposed design in cameras is discussed and evaluated in terms of circuit complexity as well as computational load.**

Keywords- pulse frequency modulation (PFM); suppression; promotion; digital image processing.

I. INTRODUCTION

The demand of solid-state image sensors has grown up largely due to the increasing requests for digital still and video cameras. Charge-Coupled Devices (CCD) were earlier dominant in the image sensor market. However, recent progresses in the design of CMOS technologies have led to appearing CMOS sensors in most imaging products [1]. Acquired pictures of image sensors often encounter quality limitations because of noise, low contrast and blurring effects. Accordingly, image processors are generally used to improve the related image quality after the image sensors. In traditional systems, two chips were allocated for separately receiving and processing the images [2]. This method includes drawbacks such as large chip area, higher costs, higher consumption power and more complexity. In vision chips, this idea has been realized in an integrated way so that necessary processors are implemented along with photosensors to achieve better performance as well as more compact system. The idea of integrating preprocessing and sensor circuits at pixel level before transmitting image pixels has resulted in smart vision sensors and being advantageous over traditional image sensors [2]. A substantial reduction in noise and interferences is obtained as well as one of two chips is realized in the sensor level [3].

With the progress in VLSI technology, digital processing techniques have substituted analog one at the sensor level to improve performance though more transistors are utilized. One of best candidates for image processing at pixel level is pulse domain techniques where the photodiode's output signal is converted to a pulse train considering modulation techniques such as Pulse Frequency Modulation (PFM) [4].

In this paper, the focus is concentrated on basic image processing namely preprocessing techniques in pulse domain considering PFM. PFM photosensor represents the pixel value by the frequency of a series of digital pulses. Pulse modulation exhibit advantages in the image processing applications because of lower power consumption, higher speed and ability to integrate image pre-processing functions at sensor level. Besides, it may be automatically designed and implemented using Electronic Design Automation (EDA) tools. In this paper, image processing techniques including edge detection and contrast enhancement algorithms are designed and simulated in pulse domain as well as implemented on FPGA level. The organization of the paper is as following. Section II provides a review on the classic algorithms of image processing. Section III describes the principles of processing methods in pulse domain. In Section IV, image processing techniques are simulated in pulse domain and compared to classic ones. The results of circuit simulation and implementation are presented in Section V. Finally, Section VI includes the conclusion.

II. CONVENTIONAL APPROACHES

Edge detection and contrast enhancement represent the fundamental operations for improving image quality. In this section, a brief review of main conventional algorithms is described in this regard.

A. Edge detection

Edge enhancement and Edge detection are important methods in image processing. Also, these are basic operations at higher-level visual processing such as image segmentation, recognition and image compression [5]. Changes in some physical properties, such as intensity of illumination, color and reflectance appear as edges in the scene. Various algorithms have been so far proposed for edge enhancement such as Gradient operators (Robert [6], Sobel [7]) using maximum and minimum of first derivative. Gradient methods are more appropriate when the variation of luminance intensity is large in presence of a low noise [8]. Laplacian operator is another approach for edge detection. However, it exhibits a large sensitivity to noise and is unable to detect the orientation of edges. LOG operator is used as another candidate particularly when image is blurring or image noise is large. It is a combination of Gaussian low-

 ISBN: 978-1-61208-100-7

pass filter and Laplacian operator. As another possibility, Canny [9] employs Gaussian smoothing to decrease noise and the Gaussian first derivative to detect edges. It nevertheless takes longer computation time and is naturally more complicated. However, the characteristics of human vision system have not been considered in these approaches. Marr and Hildreth [10] proposed another way based on the visual human performance using mathematical models. An improvement of this manner was developed by Peli [11] in which some filtering channels are used for edge enhancement stage and the threshold is the contrast sensitivity of human eye, and then, outputs of various frequency bands (spatial scales) are combined for producing the final edge detection. Another method for edge detection is the Logical/Linear operators [12] which combines linear operator's theory and Boolean algebra. Also, new algorithms such as fuzzy theory [13], natural network [14], mathematic morphologic theory [15], wavelet transformation [16] and rough sets [17] have been utilized for edge detection. They are often much more complex.

B. Contrast enhancement

One of the main parameters for measuring the quality of image is contrast. The contrast of image stands for a dynamic range representing the ratio of the brightest to the darkest pixel intensities [18] as follows:

$$Contrast = \frac{I_{max} - I_{min}}{I_{max} + I_{min}} \quad (1)$$

That I_{max} and I_{min} represent the highest and lowest intensity of illumination in image respectively. Contrast enhancement procedures leads to an improvement in visual quality for low contrast images. Contrast level is often studied using histograms. The histogram of a digital image with gray levels shows the distribution of pixel intensities [18]. For example, histogram of an image with the pixel intensities in the range $[0, L-1]$ is a discrete function $h(r_k) = n_k$ that r_k is kth gray level and n_k represents the number of image pixels having a gray level of r_k. A normalized histogram is defined as:

$$A(r_k) = \frac{n_k}{n} \qquad k = 0,1,2,\dots,L-1 \quad (2)$$

Contrast enhancement methods include histogram processing methods such as histogram equalization and contrast stretching algorithms such as Negative Transform, Log Transform, gamma correction and gray slicing [5]. The Histogram Equalization (HE) distributes uniformly the pixels over global range of gray levels with a predefined transformation function $s = T(r_k)$. HE method nevertheless suffers from contrast over-enhancement in bright sections.

III. PULSE DOMAIN IMAGE PROCESSING

In pulse-domain, the output signal of photodiode is converted to a pulse train using a pulse frequency modulation (PFM) scheme. In this case, the luminance of pixels is represented by the frequency of related output pulse stream. PFM sensor includes a self-reset feedback loop that consists of a photodiode (PD), a reset transistor and an A/D converter's circuit to obtain the pulse train (Figure 1) [19].

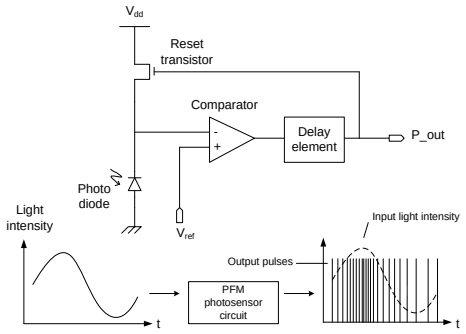


Figure 1. Architecture of PFM photosensor

Various techniques of image processing such as edge enhancement, edge detection and contrast enhancement may be simply realized in pulse domain using only two preliminary operations of suppression and promotion. These basic operations deal with the desired pixel as well as neighboring pixels. In suppression mode, a pulse in the pixel of interest is omitted if it occurs simultaneously with a pulse of neighboring pixel (Figure 2(a)). This cancellation procedure results in a decrease in the intensity of pixel of interest. In second mode, a pulse is inserted in the pulse stream of desired pixel considering promotion method (Figure 2(b)). Promotion algorithm leads to a larger brightness of image since the average number of output pulses increases. To formulate suppression and promotion algorithms, output pulse train IP_OUT may be obtained as follows if assuming that the pixel of interest and neighboring ones are associated with the pulse trains being in phase [19]:

$$IP\_OUT = \begin{cases} P\_OUT.\overline{P\_NBR} & (suppression) \\ P\_OUT + P\_NBR & (promotion) \end{cases} \quad (3)$$

Where P_OUT, P_NBR and IP_OUT stand for pixel of interest, neighboring pixel and the output pulses of pixel of

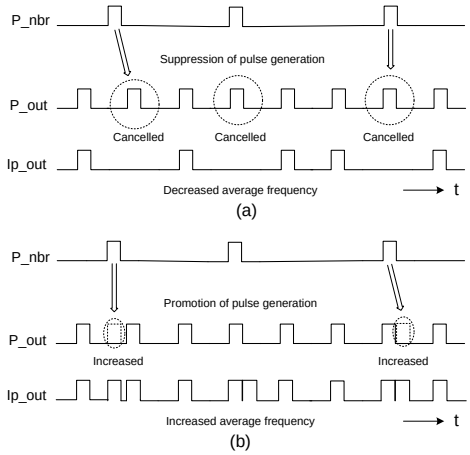


Figure 2. Processing operations in pulse domain: (a) suppression and, (b) promotion algorithms.

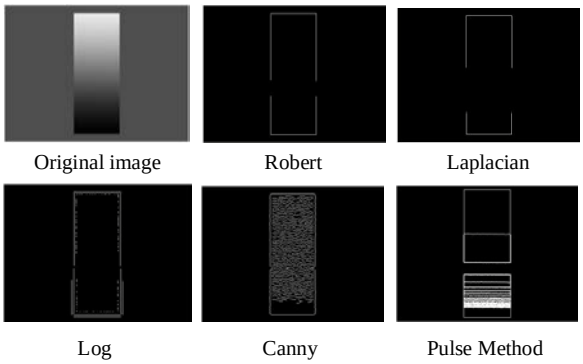


Figure 3. Results of edge detection methods compared to the result of pulse domain technique (right-bottom one).

interest respectively. According to (3), it may be easily shown that suppression operation leads to an edge enhancement and detection (compare with gradient), and pulse promotion operation is associated with a contrast enhancement algorithm.

IV. Simulation and Comparison

In this section, image processing algorithms using suppression and promotion operations are simulated with Matlab and the results are compared with the ones due to conventional algorithms. For this purpose, each image pixel has been represented by a train of 256 pulses as the output of PFM photosensor. Then, the basic pulse operators (suppression and promotion) are applied considering four neighboring pixels. Two codes have been designed as well to simulate PFM modulation and demodulations having applied

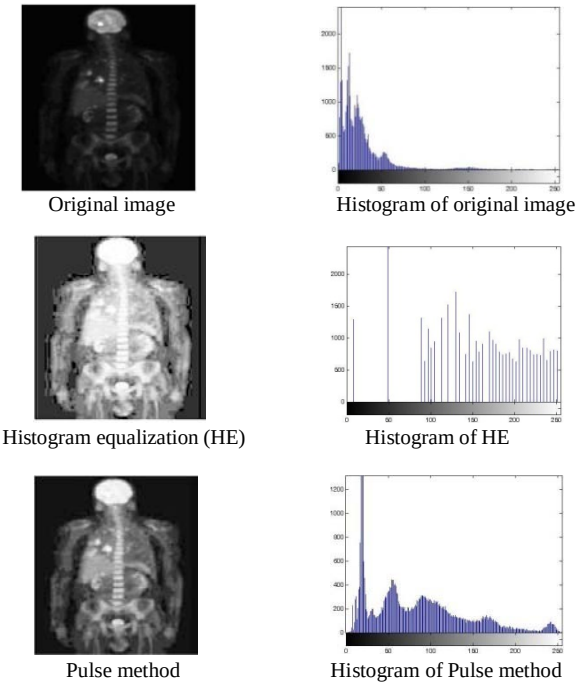


Figure 4. Contrast enhancement: original image (top) vs. the result of HE algorithm (middle) and result of pulse domain technique (promotion).

TABLE I. Comparison of Pulse Method with Conventional Algorithms

Parameter	Conventional Methods	Pulse Method
Data Processing Format	Byte	Bit
Hardware	Adder & multiplier	Bit logic circuits
Complexity	High	Low
Consumption Power	High	Low
Computation Manner	Batch	Real- time
Circuit Cost	High	Low
Speed	Low	High

before and after pulse processing techniques respectively. Then, both input and output of simulations are imagined as image data files. In the suppression algorithm, the pulse of the pixel of interest disappears when any of neighboring pixels send at the same time a pulse. In the promotion algorithm, both pulses of the pixel of interest and the neighboring one are considered in the output separately. Simulations show that pulse domain techniques lead to a better quality of edge detection and contrast enhancement than conventional methods. For example, one can find some edge layers at the output of pulse domain technique that have not been detected through conventional algorithms. Figure 3 shows the results for edge detection algorithms of Robert, Laplacian, Log and Canny compared to the pulse domain result. Besides, it is possible to control the accuracy, orientation and intensity of detected images in pulse domain technique using different neighbor pixel groups. Comparing computation loads, pulse domain technique exhibits faster performance since Laplacian and gradient algorithms are associated with larger computations. For example, a sample simulation time for canny and pulse domain edge detection methods are 0.25s and 0.013s respectively. In Figure 4, the results of contrast enhancement operation using HE algorithm and pulse domain method (promotion) have been shown. The results may be compared using histograms presented in Figure 4 as well. From computational point of view, pulse domain technique shows again a substantial improvement. HE method of contrast enhancement employs a transfer function applying to image that is associated with a large load of computations. Conventional methods such as HE algorithm use a nonlinear function which increases complexity in contrast to pulse domain technique realized by simple logic operations. Accordingly, pulse domain technique provides an algorithm being faster and simpler for contrast enhancement. The comparison of performances has been summarized in Table 1 for pulse domain and conventional methods.

V. Simulation and Implementation on FPGA level

Image processing techniques may be realized at the sensor level by considering two methodologies: inter-pixel [4] and intra-pixel [20] methods. Intra-pixel methods are generally used for improving sensor performance. On the other hand, Inter-pixel implementation leads to realization of programmable pixel sensors. For example, an inter-pixel realization of pulse domain techniques at sensor level has

 ISBN: 978-1-61208-100-7

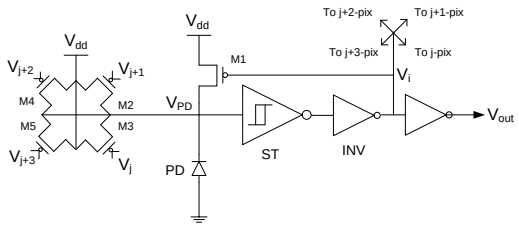


Figure 5. Inter-pixel realization of promotion method at analog sensor level

been shown in Figure 5. It is associated with pulse promotion operation with four neighboring pixels [4]. According to Figure 5, it has been implemented at analog circuit level (sensors). In this paper, image processing algorithms are implemented using an inter-pixel scheme using only digital circuits (not analog sensor level). PFM sensors have been employed for this purpose. A block of 7×7 pixels has been selected as basic block for implementation on FPGA. ISE software from Xilinx has been used along with XST and ISM as integrated synthesis and simulation tools. Considering PFM scheme, a digital basic circuit has been implemented per each pixel. Figure 6 shows the block diagram of related circuit designed for each pixel. In this general configuration, P_nbr represents the output of digital processing on four (up, down, left and, right) neighboring pixel pulses. This architecture realizes both suppression and promotion algorithms depending on control signal of SC. The pixel of interest is represented by p_out and Ip_out signals before and after processing respectively. P2nbr signal is used for next pixels processing as the neighboring pixel. In processing circuits, p_out and p_nbr signals are then utilized. At final stage, a digital circuit has been designed to provide the output in bit format. The output pulses of image processor unit are converted to *n* bits. Considering *M* clock cycles applied in any *n*-bit pixel cycle, the maximum rate of input pulses is supposed being equal to the clock frequency f_{CLK}. In this paper, it has been supposed that $M = 2^n$. Signal rate at each signal wire of pulse train (*P_out*) and output pixel samples (*Q*) would be f_{CLK} and f_{CLK}/M. The output of pixel is read out as an *n*-bit parallel digital signal. The results of timing simulation for one pixel have been demonstrated in Figure 7. In this simulation, it is supposed that n=5 for better illustrating timing diagrams. S and P signals are the outputs of image processing block for suppression and promotion respectively. Meanwhile, the outputs of pulse-to-bit converter block are represented by Q1 and Q2 for suppression and promotion respectively. Figure 8 illustrates

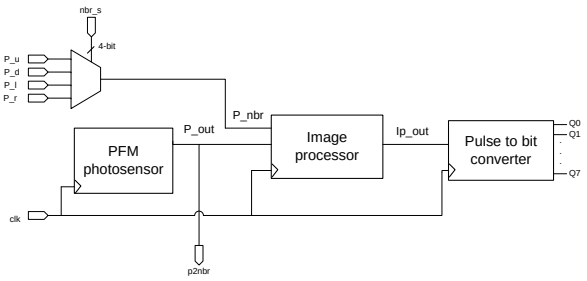


Figure 6. Block diagram of digital circuit per pixel.

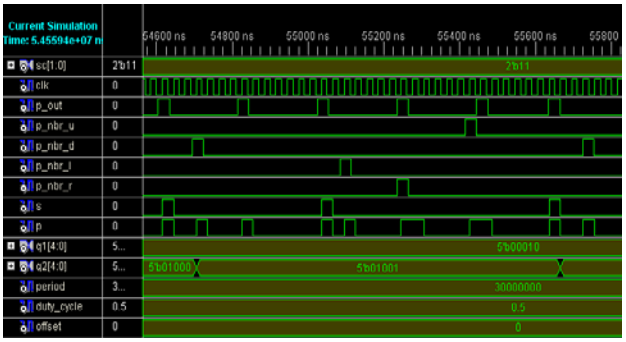


Figure 7.The results of timing simulation for one pixel

the schematic at Register Transfer Level (RTL) for the image processing part of this circuit with n=8. The design summary for image processing and pulse-to-bit converter blocks is demonstrated after implementation in Table 2 considering device counts per pixel. Total number of necessary devices per pixel has been summarized in Table 2 as well. Figure 9 shows the global architecture of an array (block of 7×7 pixels). The output values from any pixel are read out as XY address format.

VI. CONCLUSION

In this paper, parallel image processing algorithms have been discussed and simulated for contrast enhancement and edge detection at the pixel level using pulse domain basic techniques of promotion and suppression respectively. Simulation show that pulse domain techniques lead to a better quality as well as higher performance in comparison to conventional image processing methods. Besides, computational load of pulse domain techniques is limited to logical operations which appear much faster than multiplication operations utilized in conventional methods. On the other hand, pulse domain techniques appear as suitable candidate for implementation. In this paper, a sample implementation has been presented and analyzed at RTL and gate levels. Xilinx Spartan III family has been employed as programmable circuit for implementation of suppression and promotion techniques. Also, this algorithm gives good results for color images. Future work will be concentrated on realization of other image processing algorithms in pulse domain.

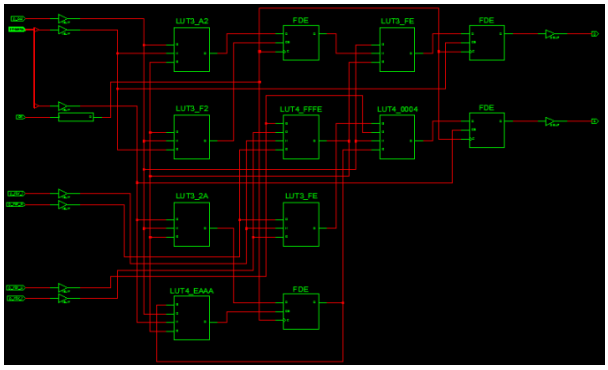


Figure 8. RTL schematic of suppression and promotion algorithms per pixel

TABLE II. Design Summary per Pixel (number of devices used)

Resource	Image Processing	Pulse to Bit Converter	Entire Pixel
No. of Flip Flops	4	38	42
No. of Slices	5	32	37
No. of LUTs	8	45	53
No. of IOBs	10	14	24
Simulation Time	1 sec	1 sec	2 secs

REFERENCES

[1] A. El Gamal and H. Eltoukhy, "CMOS Image Sensors," IEEE Circuit & Devices Magazine, May/June 2005, pp. 6-20.

[2] A. Moini, "Vision chips or seeing silicon," The Centre for High Performance Integrated Technologies and Systems, The University of Adelaide, SA 5005, Australia, 1997.

[3] C. S. Hong, "On-Chip Spatial Image Processing with CMOS Active Pixel Sensors," PhD thesis, Electrical and Computer Engineering, University of Waterloo, Ontario, Canada, 2001.

[4] J. Ohta, A. Uehara, T. Tokuda, and M. Nunoshita, "Pulse-Modulated Vision Chips with Versatile-Interconnected Pixels," International Parallel and Distributed Processing Symposium (IPDPS), Cancun, Mexico, 2000, pp. 1063-1071.

[5] R. C. Gonzalez and R. E. Woods, "Digital Image Processing," Prentice Hall, 2002.

[6] L. G. Roberts. "Machine perception of three dimensional Solids," Optical and Electro-Optical Information Processing, MIT Press Cambridge, Massachusetts, 1965, pp. 159-197.

[7] W. K. Pratt, "Digital Image Processing," New York, Wiley-Interscience, 1978.

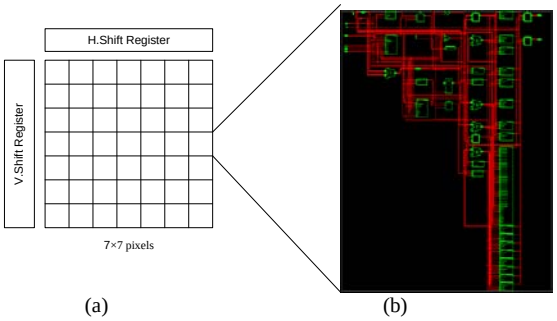
[8] M. Basu, "Gaussian-Based Edge-Detection Methods—A Survey," IEEE Transactions on Systems, Man, and Cybernetics, Vol. 32, NO. 3, August 2002, pp. 252-260.

[9] J. Canny, "A Computational Approach to Edge Detection," IEEE Transactions, Pattern Analyze, Machine Intel, Vol. PAMI-8, June 1986, pp. 679–698.

[10] D. Marr and E. C. Hildreth, "Theory of edge detection," in Proceedings of the Royal Society of London, vol. B207, 1980, pp. 187–217.

[11] E. Peli, "Feature Detection Algorithm Based on a Visual System Moddel," Proc. IEEE, vol. 90, pp. 78-93, 2002.

[12] L. A. Iverson and S. W. Zucker, "Logical/linear operators for image curves," IEEE Trans. Pattern Anal. Machine Intell., vol. 17, no. 10, pp. 982-996, Oct 1995.



(a) (b)

Figure 9. (a) Schematic diagram of an array, (b) RTL schematic per pixel

[13] J.B. Wu, Z.P. Yin, and Y.L. Xiong, "The Fast Multilevel Fuzzy Edge Detection of Blurry Images", IEEE signal processing letters, vol. 14, 2007, p.344-347.

[14] H.J. Yang and D.Q. Liang, "A New Method of Edge Detection Based on Information Measure and Neural Network", ACTA ELECTRONICA SINICA, vol. 29, 2001, pp.51-53.

[15] Z.Y. Cai, R. Chen, F.Z. Yu, and B. Li, "A V-groove Welding Seam Recognition Algorithm Based on Wavelet Transform", Journal of Image and Graphics, vol. 12,2007,pp.866-869.

[16] W.P. Gong, Y.Z. Wang, and Q. Li, "Multi scale edge detection on images using B-spline wavelets", Inflared Technology, vol. 22, 2000, pp. 15-18.

[17] D. Wang, M.D. Wu, and Y.S. Liu, " A New Mathematical Morphological Algorithm Based on Rough Sets and It's Application to Detecting Image Edge", Journal Of Engineering Graphics, vol. 2, 2007, pp. 109-113.

[18] G. H. Park, H. H. Cho, and M. R. Choi, "A Contrast Enhancement Method using Dynamic Range Separate Histogram Equalization," IEEE Transactions on Consumer Electronics, Vol. 54, No. 4, November 2008, pp.1981-1987.

[19] K. Kagawa and K. Yasuoka."Pulse-Domain Digital Image Processing for Vision Chips Employing Low-Voltage Operation in Deep-Submicrometer Technologies", IEEE Journal In Quantum Electronics, Vol.10, No.4, July/August 2004, pp. 816-828.

[20] M. Kyomasu, "A New MOS Imager Using Photodiode as Current Source," IEEE Journal of Solid State Circuits 26, pp. 1116–1122, August 1999.

Inter and Intra-Video Navigation and Retrieval in Mobile Terminals

Andrei Bursuc, Titus Zaharia
Institut Télécom; Télécom SudParis; ARTEMIS Dept.
UMR CNRS 8145 MAP 5
9, rue Charles Fourier, 91011 Evry Cedex
{Andrei.Bursuc, Titus.Zaharia}@it-sudparis.eu

Françoise Prêteux
Mines ParisTech
60, Boulevard Saint-Michel 75272 Paris Cedex, France
Francoise.Preteux@mines-paristech.fr

***Abstract*—This paper introduces a novel on-line video browsing and retrieval platform, so-called OVIDIUS (*On-line VIDeo Indexing Universal System*). In contrast with traditional and commercial main stream video retrieval platforms, where video content is treated in a more or less monolithic manner (i.e., with global descriptions associated with the whole document), the proposed approach makes it possible to browse and access video content in a finer, per-segment basis. The hierarchical metadata structure exploits the MPEG-7 approach for structural description of video content. The MPEG-7 description schemes have been here enriched with both semantic and content-based metadata. The developed approach shows all its pertinence within a multi-terminal context and in particular for video access from mobile devices. The platform has been recently (February, 2010) validated within the framework of the Médi@TIC French national project.**

Keywords - video indexing, video search engines, user interfaces, MPEG-7 standard, visual descriptors, description schemes.

I. INTRODUCTION

Over the last ten years, mobile devices have been undergoing a booming prosperity in our everyday life. A broader variety of handheld devices with audio/video playback functionalities are available in the market with a reasonable price. The huge steps forward made by third generation communication networks enable telecom operators to provide better mobile multimedia services, such as smoother streaming time and higher quality of video resolution.

A. Context and objectives

Currently we are witnessing a proliferation of powerful mobile phones capable of fast connections and high-fidelity multimedia rendering; the so-called "smartphones". According to the consultancy company Nielsen, the amount of smartphone subscribers in the United States has augmented from 14% at the end of 2008 to 29% at the beginning of 2010. The same study predicts that by the end of 2011 in the U.S. there will be more smartphones than feature phones [1].

Meanwhile, average mobile users seem to take more advantage of the new features provided by the phone, such as the faster internet connection. Thus, the use of Wi-Fi increases 10 times from 5% for feature phone owners to 50% for smartphone users [1]. A recent study made by Nielsen illustrates that in the U.S. active mobile video users grew by 57% from the fourth quarter of 2008 to the fourth quarter of 2009, from 11.2 million to 17.6 million [2]. This means that mobile subscribers are also developing a growing appetite for online video content.

However, the functionalities of such devices are constrained by their size and computational/memory capacities. Existing studies have put an emphasis around the consumer and his behavior, aiming to build user-driven interfaces and applications [3]. In this context, an important technological challenge concerns the issue of accessing/retrieving video content from mobile devices. The critical point relates to the high complexity of video content, in terms of amount of heterogeneous information included. In order to tackle the issue of complexity, appropriated presentation and search engines, as well as novel interaction modalities need to be developed. In addition, it is necessary to ensure a personalized access to *segments* of interest, defined as parts of an audio-visual document. User should have the possibility of rapidly browsing the content and identify/access the segments of interest.

Let us analyze how such aspects are treated in the state of the art.

B. Related work

Our work draws upon research in several areas concerning multimedia content management: content-based image and video retrieval, multimedia content management, user feedback management and distributed content sharing.

One of the most active areas over the last decade has been the content-based image retrieval area. Within this context, let us first mention the Multimodal Automatic Mobile Indexing (MAMI) system, which allows users to annotate and search for digital photos via speech input combined with time, date and location information [4]. Jesus *et al*. used geographical queries to retrieve personal pictures when visiting points of interest [5]. Zhu *et al.* integrate the features mentioned above in their user-centric

 ISBN: 978-1-61208-100-7

system, called iScope [6]. iScope uses multi-modality clustering of both content and context information for efficient image management and search, and online learning techniques for predicting images of interest, while supporting distributed content-based search among networked devices.

Kim *et al.* have used CBIR for visual-content recommendation [7] while Yeh *et al.* have used mobile images for content-based object search [8]. CLOVER searches sketches or photos of leaf images on a server starting from a mobile phone [9]. Photo-to-Search performs queries directly on the web using images taken with a mobile device [10].

The issue of incompatibility between video resolution and certain mobile phones is addressed by the researchers from Zhejiang University [11]. User feedback is used in order to update the metadata for video clips, including the acceptable resolution. In addition, redundant versions with lower resolution are generated for video clips by estimating their popularities, and sent to users with lower resolution directly to save the downsizing transcoding process.

The Multimedia Content Creation Platform (MMCP) [12] takes full advantage of the context information provided by the mobile device each time a new picture or video is created and added immediately as a metadata (date, time, location, etc.). The platform can be used both for generating new contact and retrieve content as well.

The system proposed in [13] allows users to retrieve video content starting from a mobile photo uploaded by the user. The search engine in the background returns videos containing key frames that are similar with the uploaded picture. An automatic key frame extractor and the Contrast Context Histogram (CCH) [14] have been used for the elaboration of the system.

Let us also mention the innovative approach proposed by Miller *et al.* Their mobile media content browser has been developed and placed on an iPhone device. MiniDiver [15] is based on four user interfaces serving mobile context sensitive video. After selecting the desired video or program via a simple interface, the user can view the content from one or two camera perspectives simultaneously. In the case of live broadcasts, users can have the moving objects (*e.g.,* hockey players) highlighted and can choose the favorite camera angles.

The existing approaches offer interesting preliminary solutions to the issue of mobile video access. However, most of them are basically mobile versions of their desktop-based counterpart and hence, do not take into account the specificity of mobile devices, environments and usages/services. In addition, they massively focus on textual queries, while an efficient search process should consider rich and multimodal queries, combining text, image, audio and video features.

Another drawback of such approaches comes from the fact that videos are considered in a monolithic manner, without taking into account the intrinsic spatio-temporal structure of the video content. However a common video document may include a huge amount of heterogeneous information that needs to be identified, described and accessed independently.

Creating dedicated tools for query formulation, metadata driven visualization / navigation and ergonomic user interaction in both fixed and mobile environments is still a challenge that needs to be addressed and solved.

C. Contributions

The analysis of the state of the art shows that currently no system fully exploits the intrinsic spatio-temporal structure of video documents. The OVIDIUS (*On-line VIDeo Indexing Universal System*) platform proposed in this paper aims at solving such limitations, ensuring all the interaction and navigation capabilities needed to access video content at a fine level of granularity, from fixed or mobile devices.

The strong points of the proposed system are the following :

- Modular and distributed architecture, achieved with the help of web services, which makes it possible to easily upgrade the system in order to keep pace with inherent future technological advances,
- Fine granularity access to video content, based on the MPEG-7 structural approach for video content description [16]
- Core interoperability achieved with open MPEG-7 standard technologies,
- Enrichment of MPEG-7 structural description schemes with semantic and content-based descriptors,
- Advanced interaction functionalities integrating browsing, search, hierarchical navigation and visualization capabilities,
- Support of both textual, content-based and hybrid queries,
- Compatibility with a vast variety of platforms.

The rest of the paper is organized as follows. Section 2 presents the description framework that we have adopted for our platform. Section 3 gives an overview of the proposed OVIDIUS system and presents the audio-video analysis techniques that have been considered in the metadata extraction engine. Finally, Section 4 concludes the paper and opens perspectives of future work.

 ISBN: 978-1-61208-100-7

II. MPEG-7 STRUCTURAL VIDEO CONTENT DESCRIPTION

The MPEG-7 structural approach for video description [16] is based on an abstract class (description scheme) of *Segment*. An MPEG-7 Segment represents an arbitrary part of a video and includes generic descriptors (*e.g.*, textual annotations, keywords, temporal localization elements for specifying the starting and the ending time stamps of a segment, etc.). Starting from this abstract structure, which cannot be instantiated, a set of media-specific segments is derived, by applying an inheritance mechanism. In our implementation, we have considered the following MPEG-7 segments: AudioVisualSegment DS, AudioSegment DS, Video DS, StillRegionDS, and MovingRegion DS.

Each segment can include dedicated descriptors, adapted to each segment type. Examples are color features for still region/video segment, audio features for audio segment, motion parameters for MovingRegion DS, etc.

A second MPEG-7 mechanism exploited is the Segment Decompostion DS, which allows the partition of a segment into sub-segments. Applied recursively, this mechanism makes it possible to represent a video as a hierarchical tree structure made of scenes, shots, transcriptions segments, and key-images.

The adopted MPEG-7 language for specifying all these descriptors and descriptions schemes is XML schema. This choice facilitates the parsing and interpretation of the descriptions, since various XML utilities are available and can be directly used (*e.g.*, Xerces parser).

Disposing of metadata is a first and essential step in the indexing process. However, appropriate video visualization and interaction capabilities need to be elaborated in order to efficiently exploit such a description. The OVIDIUS user interface integrates all the necessary interaction and navigation capabilities, as described in the following section.

III. OVIDIUS PLATFORM: SYSTEM OVERVIEW

Figure 1 illustrates the distributed architecture adopted in the OVIDIUS platform. It includes a content management module (*i.e.*, storage and editing of content and metadata), a metadata extraction engine, a MPEG-7 search engine and a web interface which can be remotely accessed from mobile and fixed environments.

Figure 1. Overview of the OVIDIUS platform.

All the components of the system can be distributed on various servers. The communication between them is performed by using web services with http requests. Such a modular approach facilitates the future extension of the platform as well as the replacement of individual components (*e.g.*, video players, search engine…).

A. OVIDIUS user interface

In order to ensure a large interoperability, the interface has been developed by using HTML, PhP and JavaScript technologies. We have successfully tested our system on iPhone 3G and 3GS devices and Android smartphones as well. Due to its construction, OVIDIUS can be accessed from other types of smartphones with a mobile internet browser JavaScript compatible (*e.g.* Android phones). The interface server automatically detects the operating system of the user terminal and adapts accordingly. The sole aspect that should be taken into consideration when switching to other types of terminals is the issue of video encoding formats.

Figure 2 illustrates the OVIDIUS user interface. The following components are included: selector of the segment type, selector of the hierarchical level of each segment, iconic representation of segments (which are dynamically derived from the media), navigation/browsing buttons, summary and keyword visualization and selection.

 ISBN: 978-1-61208-100-7

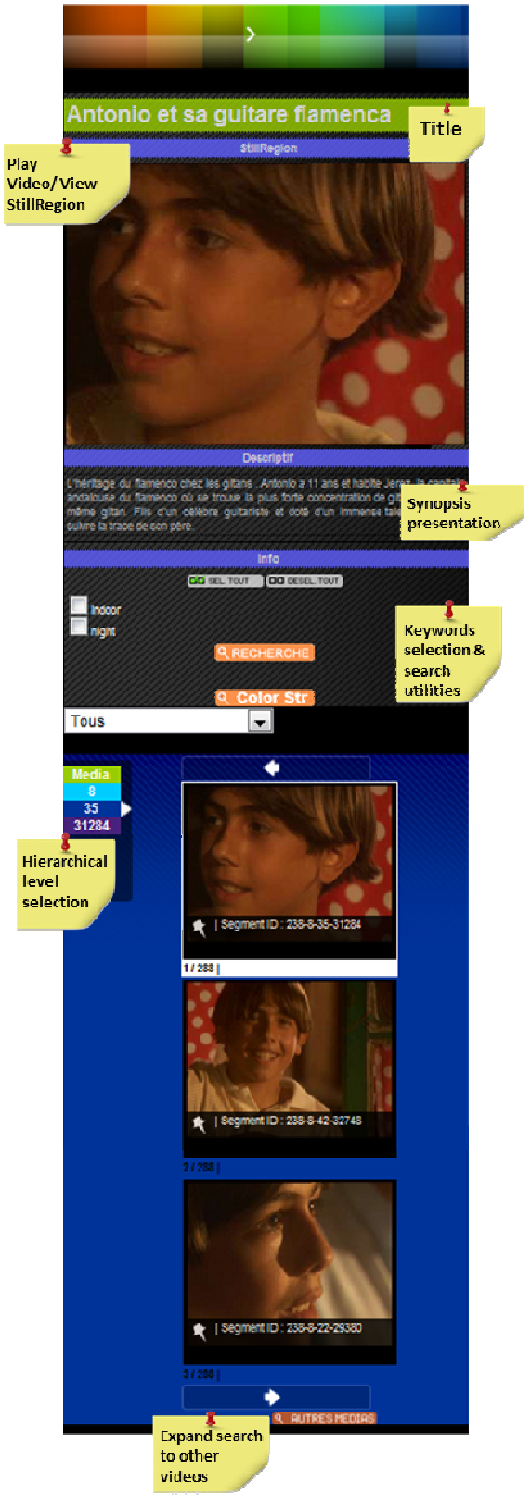


Figure 2. Components of the OVIDIUS user interface and results for a query by Color Structure descriptor.

One column of iconic preview representations is dedicated to each level of hierarchy of the video. Scenes, shots, and still regions can be browsed and accessed in a hierarchical manner, as MPEG-7 segments. Thus, the user can navigate inside a video both horizontally (through video segments on the same level) and vertically (through different levels of the scene hierarchical tree structure).

The iconic images are dynamically generated at each access with the use of a Java based frame extractor and the MPEG-7 time stamps. Each iconic image provides information about the visual content of the segment (preview image), as well as information regarding the classification and hierarchy (type of segment, segment identifier).

OVIDIUS can exploit arbitrary extraction methods and utilities, provided that the representation of the description is compliant with the MPEG-7 specification.

A specific feature has been added envisioning time and bandwidth efficiency. By taking advantage of the FFMpeg library [17], OVIDIUS can automatically cut a segment from the video at user request. Whenever a user wants to access a certain video segment, this video segment will be automatically cut from the media and sent to user. This way the user will receive the exact requested video segment.

Thanks to the web implementation of the OVIDIUS platform, a PC version is also available, offering the same navigation functionalities, features, speed and user experience (Figure 3).

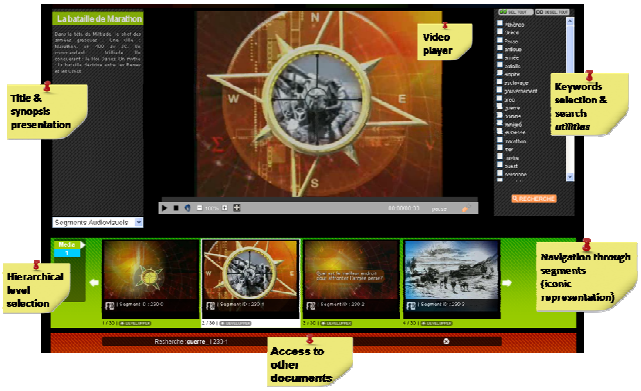


Figure 3. PC version of the OVIDIUS user interface.

B. *Metadata extraction engine*

In order to instantiate the MPEG-7 structural video description, we have implemented a video segmentation on three levels, including scenes, shots and key frames.

The first step in the segmentation process was to develop an efficient shot boundary detector. Starting from the techniques and results presented in [18], we have developed a combined two levels method able to detect both CUTs and gradual transitions. Abrupt CUTs are tackled at the first level of detection, based on similarity of color histograms.

 ISBN: 978-1-61208-100-7

In the case where no CUT is detected and the similarity score is above a certain threshold on a group of 10 frames, a second level of detection is applied. Here, we use the graph partition model proposed in [19]. This second step allows the detection of gradual transitions.

For key-frame extraction we have used an adaptation of the algorithm developed by Zhuang et al. [20] which detects multiple frames based on the visual variations in shots.

Finally, the approach described in [21] has been adopted in order to aggregate shots into scenes. The principle consists of constructing a weighted undirected graph so-called shot similarity graph, which involves a similarity measure that combines color and motion features.

The detected elements are represented using MPEG-7 segments (*cf.* §II). To each segment, appropriate descriptors are associated with. For this purpose, the entire set of MPEG-7 visual descriptors [22], related to color, texture and motion features is currently supported, based on an optimized version of the MPEG-7 reference software.

Currently we have tested the performances of the MPEG-7 Color Structure Descriptor in retrieving visually similar video sequences. The experiments have been carried out on the Médi@TIC corpus, kindly provided by LBA (Vodeo.TV) which includes about 15 documentary videos. As shown in Figure 4 the Color Structure Descriptor is very effective in retrieving instances of the same person and same environment within the video corpus.

Textual information which corresponds to the transcription of the audio track available for the Médi@TIC corpus is also included as semantic information.

For all the descriptors considered, dedicated XML schema representations have been elaborated and used to enrich the MPEG-7 schema definition.

Figure 4. Search results based on the MPEG-7 ColorStructure descriptor with queries by example.

C. *Main functionalities*

Let us summarize the main functionalities available for the OVIDIUS platform:

- browsing of a database of videos with keyword search
- selection of a video and navigation through the hierarchical structure of scenes, shots and key-images, in order to quickly discover the contents
- query formulation,
- search engine capabilities: search by visual features,
- identification of features of interest and search of segments in the whole video database.

Concerning the query formulation, OVIDIUS provides a simple keyword panel, displaying information extracted from the soundtrack of the video. Keyword of interest can be selected through checkboxes and searched within the video segments on the same level. Query results from the same video or from other videos in the data base can be browsed and accessed.

In the near future, we plan to extract also keywords describing the visual content of the video sequence (*e.g.*, indoor, outdoor, night, day, etc.).

IV. CONCLUSIONS AND PERSPECTIVES

In this paper, we have presented the OVIDIUS video indexing platform, which supports hierarchical and multigranular MPEG-7 descriptions while integrating all the necessary interaction, browsing and visualization utilities. The proposed approach makes it possible to discover a video document in a few clicks, and can be accessed from various platforms (*e.g.*, iPhone or PC).

The perspectives of future work concern the extension of the system with new, advanced descriptors and extraction engines. Notably, we will focus on the elaboration of an object detection/recognition/retrieval framework, able to identify and describe user-specific elements of interest.

ACKNOWLEDGMENT

This work has been partially supported by the French national project Medi@TIC, double labeled by the Systematic and CapDigital competitiveness clusters. The OVIDIUS platform has been developed by Institut TELECOM.

 ISBN: 978-1-61208-100-7

REFERENCES

[1] http://blog.nielsen.com/nielsenwire/consumer/smartphones-to-overtake-feature-phones-in-u-s-by-2011/, last accessed April 2010

[2] http://blog.nielsen.com/nielsenwire/online_mobile/three-screen-report-q409/, last accessed April 2010

[3] L. Wang, D. Tjondrongoro, and Y. Liu, "Clustering and visualizing audiovisual dataset on mobile devices in a topic-oriented manner," *Proceedings of the 9th international conference on Advances in visual information systems*, Shanghai, China: Springer-Verlag, 2007, pp. 310-321.

[4] X. Anguera, J. Xu, and N. Oliver, "Multimodal photo annotation and retrieval on a mobile phone," *Proceeding of the 1st ACM international conference on Multimedia information retrieval*, Vancouver, British Columbia, Canada: ACM, 2008, pp. 188-194.

[5] R. Jesus, R. Dias, R. Frias, and N. Correia, "Geographic image retrieval in mobile guides," *Proceedings of the 4th ACM workshop on Geographical information retrieval*, Lisbon, Portugal: ACM, 2007, pp. 37-38.

[6] C. Zhu, K. Li, Q. Lv, L. Shang, and R.P. Dick, "iScope: personalized multi-modality image search for mobile devices," *Proceedings of the 7th international conference on Mobile systems, applications, and services*, Kraków, Poland: ACM, 2009, pp. 277-290.

[7] C. Y. Kim, et al. , "VISCORS: A visual-content recommender for the mobile web," IEEE Intelligent Systems, vol. 19, no. 6, pp. 32-39, 2004.

[8] T. Yeh, K. Grauman, K. Tollmar, and T. Darrell, "A picture is worth a thousand keywords: image-based object search on a mobile platform," *CHI '05 extended abstracts on Human factors in computing systems*, Portland, OR, USA: ACM, 2005, pp. 2025-2028.

[9] S. Kim, Y. Tak, Y. Nam, and E. Hwang, "mCLOVER: mobile content-based leaf image retrieval system," *Proceedings of the 13th annual ACM international conference on Multimedia*, Hilton, Singapore: ACM, 2005, pp. 215-216.

[10] M. Jia, et al., "Photo-to-Search: Using camera phones to inquire of the surrounding world," *in MDM '06: Proceedings of the 7th International Conference on Mobile Data Management,* 2006, pp. 46-46.

[11] Y. Liu, Z. Yang, X. Deng, J. Bu, and C. Chen, "Media Browsing for Mobile Devices Based on Resolution Adaptive Recommendation," *Proceedings of the 2009 WRI International Conference on Communications and Mobile Computing - Volume 03*, IEEE Computer Society, 2009, pp. 285-290.

[12] S. Järvinen, J. Peltola, J. Lahti, and A. Sachinopoulou, "Multimedia service creation platform for mobile experience sharing," *Proceedings of the 8th International Conference on Mobile and Ubiquitous Multimedia*, Cambridge, United Kingdom: ACM, 2009, pp. 1-9.

[13] C. Chen, Y. Wang, H. Wang, and C. Chiu, "Digital Video Retrieval via Mobile Devices," *Proceedings of the 2008 Fourth IEEE International Conference on eScience*, IEEE Computer Society, 2008, pp. 376-377.

[14] C. Huang, C. Chen, and P. Chung, "Contrast Context Histogram - A Discriminating Local Descriptor for Image Matching," *Proceedings of the 18th International Conference on Pattern Recognition - Volume 04*, IEEE Computer Society, 2006, pp. 53-56.

[15] G. Miller, S. Fels, M. Finke, W. Motz, W. Eagleston, and C. Eagleston, "MiniDiver: A Novel Mobile Media Playback Interface for Rich Video Content on an iPhone™," *Proceedings of the 8th International Conference on Entertainment Computing*, Paris, France: Springer-Verlag, 2009, pp. 98-109.

[16] ISO/ IEC 15938-5:2003, Information technology - MultimediaContent Description. Interface - Part 5: Multimedia Description Schemes. 2003.

[17] http://ffmpeg.org/

[18] J. Yuan, H. Wang, L. Xiao, W. Zheng, J. Li, F. Lin, and B.Zhang, "A formal study of shot boundary detection," IEEE Trans. Circuits Syst. Video Technol., vol. 17, no. 2, pp. 168-186, Feb. 2007.

[19] J. Yuan, B. Zhang, and F. Lin, "Graph Partition Model for Robust Temporal Data Segmentation," *Advances in Knowledge Discovery and Data Mining*, 2005, pp. 758-763.

[20] Y. Zhuang, Y. Rui, T. Huang, and S. Mehrotra, "Adaptive key frame extraction using unsupervised clustering," *Proceedings 1998 International Conference on Image Processing. ICIP98* (Cat. No.98CB36269), Chicago, IL, USA: , pp. 866-870.

[21] Z. Rasheed, M. Shah, Detection and Representation of Scenes in Videos, IEEE Trans. on Multimedia, 7(6): 1097-1105, Dec. 2005.

[22] ISO/ IEC 15938-3:2003, Information technology - MultimediaContent Description. Interface-Part 3: Visual. 2003.

 ISBN: 978-1-61208-100-7