ICDT 2012

The Seventh International Conference on Digital Telecommunications

ISBN: 978-1-61208-193-9

April 29 - May 4, 2012

Chamonix / Mont Blanc, France

ICDT 2012 Editors

Eugen Borcoci, Politehnica University of Bucharest, Romania

Petre Dini, Concordia University, Canada // China Space Agency Center, China

ICDT 2012

Foreword

The Seventh International Conference on Digital Telecommunications [ICDT 2012], held between April 29th and May 4th, 2012 in Chamonix / Mont Blanc, France, continued a series of special events focusing on telecommunications aspects in multimedia environments. The scope of the conference was to focus on the lower layers of systems interaction and identify the technical challenges and the most recent achievements.

High quality software is not an accident; it is constructed via a systematic plan that demands familiarity with analytical techniques, architectural design methodologies, implementation polices, and testing techniques. Software architecture plays an important role in the development of today’s complex software systems. Furthermore, our ability to model and reason about the architectural properties of a system built from existing components is of great concern to modern system developers.

Performance, scalability and suitability to specific domains raise the challenging efforts for gathering special requirements, capture temporal constraints, and implement service-oriented requirements. The complexity of the systems requires an early stage adoption of advanced paradigms for adaptive and self-adaptive features.

Online monitoring applications, in which continuous queries operate in near real-time over rapid and unbounded "streams" of data such as telephone call records, sensor readings, web usage logs, network packet traces, are fundamentally different from traditional data management. The difference is induced by the fact that in applications such as network monitoring, telecommunications data management, manufacturing, sensor networks, and others, data takes the form of continuous data streams rather than finite stored data sets. As a result, clients require long-running continuous queries as opposed to one-time queries. These requirements lead to reconsider data management and processing of complex and numerous continuous queries over data streams, as current database systems and data processing methods are not suitable.

We take here the opportunity to warmly thank all the members of the ICDT 2012 Technical Program Committee, as well as the numerous reviewers. The creation of such a high quality conference program would not have been possible without their involvement. We also kindly thank all the authors who dedicated much of their time and efforts to contribute to ICDT 2012. We truly believe that, thanks to all these efforts, the final conference program consisted of top quality contributions.

Also, this event could not have been a reality without the support of many individuals, organizations, and sponsors. We are grateful to the members of the ICDT 2012 organizing committee for their help in handling the logistics and for their work to make this professional meeting a success.

We hope that ICDT 2012 was a successful international forum for the exchange of ideas and results between academia and industry and for the promotion of progress in the field of digital communications.

We are convinced that the participants found the event useful and communications very open. We also hope the attendees enjoyed their stay in the French Alps.

ICDT Advisory Committee:

Constantin Paleologu, University Politehnica of Bucharest, Romania
Tomohiko Taniguchi, Fujitsu Laboratories Limited, Japan
Jaime Lloret Mauri, Polytechnic University of Valencia, Spain
Abdulrahman Yarali, Murray State University, USA

Michael Grottke, University of Erlangen-Nuremberg, Germany
Javier Del Ser Lorente, TECNALIA RESEARCH & INNOVATION - Zamudio, Spain
Saied Abedi, Fujitsu Laboratories of Europe Ltd. (FLE), UK
Gerard Damm, Alcatel-Lucent, USA
Dan Romascanu, Avaya, Israel
Klaus Drechsler, Fraunhofer Institute for Computer Graphics Research IGD - Darmstadt, Germany

ICDT 2012 PROGRAM COMMITTEE

ICDT Advisory Committee

Constantin Paleologu, University Politehnica of Bucharest, Romania
Tomohiko Taniguchi, Fujitsu Laboratories Limited, Japan
Jaime Lloret Mauri, Polytechnic University of Valencia, Spain
Abdulrahman Yarali, Murray State University, USA
Michael Grottke, University of Erlangen-Nuremberg, Germany
Javier Del Ser Lorente, TECNALIA RESEARCH & INNOVATION - Zamudio, Spain
Saied Abedi, Fujitsu Laboratories of Europe Ltd. (FLE), UK
Gerard Damm, Alcatel-Lucent, USA
Dan Romascanu, Avaya, Israel
Klaus Drechsler, Fraunhofer Institute for Computer Graphics Research IGD - Darmstadt, Germany

ICDT 2012 Technical Program Committee

Saied Abedi, Fujitsu Laboratories of Europe Ltd. (FLE) - Middlesex, UK
Bilal Al Momani, Cisco Systems, Inc., Ireland
Antonio Marcos Alberti, INATEL - Instituto Nacional de Telecomunicações, Brazil
Maria Teresa Andrade, FEUP / INESC Porto, Portugal
Iosif Androulidakis, MPS Jozef Stefan, Slovenia
Regina B. Araujo, Federal University of São Carlos, Brazil
Khaled Assaleh, American University of Sharjah, United Arab Emirates
Andreas Aurelius, Acreo AB, Sweden
Francisco Barcelo-Arroyo, Technical University of Catalonia, Spain
Ilija Basicevic, University of Novi Sad, Serbia
Carlos Becker Westphall, Federal University of Santa Catarina, Brazil
Abdelouahab Bentrcia, King Saud University – Riyadh, Kingdom of Saudi Arabia
Andrzej (Andrew) Borys, University of Technology and Life Sciences (UTP) – Bydgoszcz, Poland
Andi Buzo, University Politehnica of Bucharest, Romania
Tijani Chahed, IT-SudParis, France
Lee-Ming Cheng, City University of Hong Kong, Hong Kong
Doru Constantin, University of Pitesti, Romania
Gerard Damm, Alcatel-Lucent, France
Karel De Vogeleer, Blekinge Institute of Technology (BTH) - Karlskrona, Sweden
Klaus Drechsler, Fraunhofer-Institut für Graphische Datenverarbeitung IGD - Darmstadt, Germany
Roger Pierre Fabris Hoeffel, Federal University of Rio Grande do Sul, Brazil
Peter Farkas, FEI STU – Bratislava, Slovakia
Gerardo Fernández Escribano, University of Castilla-La Mancha - Albacete, Spain
Mário Ferreira, University of Aveiro, Portugal
Pierfrancesco Foglia, University of Pisa, Italy
Alex Galis, University College London, UK
Andrea Giachetti, Università degli Studi di Verona, Italy
Stefanos Gritzalis, University of the Aegean, Greece
Carlos A. Gutierrez, Panamericana University – Aguascalientes, Mexico
Maryline Hélard, INSA / Institut d'Electronique et de Télécommunications de Rennes (IETR), France
Jalaa Hoblos, Kent State University, USA

K. M. Sunjiv Soyjaudah, University of Mauritius, Mauritius
Nary Subramanian, The University of Texas at Tyler, USA
Weifeng Sun, Dalian University of Technology, China
Dimitrios G. Stratogiannis, National Technical University of Athens, Greece
James K. Tamgno. ESMT - Dakar, Sénégal
Tomohiko Taniguchi, Fujitsu Laboratories Limited, Japan
Yoshiaki Taniguchi, Osaka University, Japan
Tony Thomas, Indian Institute of Information Technology and Management - Kerala, India
Božo Tomas, University of Mostar, Bosnia and Herzegovina
Georgios I. Tsiropoulos, National Technical University of Athens, Greece
Rob van der Mei, Tilburd University, The Netherlands
José Miguel Villalón Millan, Universidad de Castilla-La Mancha, Spain
Krzysztof Walczak, Poznan University of Economics, Poland
Krzysztof Walkowiak, Wroclaw University of Technology, Poland
M. L. Dennis Wong, Xi'an Jiaotong-Liverpool University - Suzhou, P. R. China
Qishi Wu, University of Memphis, USA
Jinchao Yang, Chinese Academy of Sciences, Beijing, P. R. China
Yong Yao, Blekinge Institute of Technology - Karlskrona, Sweden
Abdulrahman Yarali, Murray State University, USA
Pooneh Bagheri Zadeh, De Montfort University - Leicester, UK
Alexander Zeifman, Vologda State Pedagogical University, Russia

ICDT Additional Reviewers

Saowaphak Sasanus, TOT, Thailand

Copyright Information

Table of Contents

Wi-Fi Proximity and Context-aware Browsing

A new approach for delivering local data

Dmitry Namiot
Lomonosov Moscow State University
Faculty of Computational Math and Cybernetics
Moscow, Russia
e-mail: dnamiot@gmail.com

Manfred Sneps-Sneppe
Ventspils University College
Ventspils International Radio Astronomy Centre
Ventspils, Latvia
e-mail: manfreds.sneps@gmail.com

Abstract— **This paper describes a new model for accessing to local data for mobile subscribers. Our model uses Wi-Fi proximity ideas. In our concept, any exiting or even especially created Wi-Fi hot spot could be used as presence sensor that can open (discover) access for some user-generated content. In our approach we can discover hyper local data as info snippets that are valid (relevant) for mobile subscribers being at this moment nearby some Wi-Fi access point. And an appropriate mobile service (context-aware browser) can present that information to mobile subscribers. As the prospect use-cases we can mention for example news and deals delivery in malls, news feeds for office centers and campuses, Smart City projects, personal classifieds and real world games.**

Keywords-Wi-Fi; proximity; collaborative location; indoor positioning; context-aware computing.

I. INTRODUCTION

In the work that first introduces the term 'context-aware', Schilit and Theimer [1] refer to context as location, identities of nearby people and objects, and changes to those objects. Other authors define context awareness as complementary element to location awareness. Whereas location may serve as a determinant for resident processes, context may be applied more flexibly with mobile computing with any moving entities, especially with bearers of smart communicators. Context awareness originated as a term from ubiquitous computing, or as so-called pervasive computing, which sought to deal with linking changes in the environment with computer systems, which are otherwise static.

Modern applications adopt a context-aware perspective to manage:

a) communication among users and among systems, or between the system and the user,

b) situation-awareness, like modeling location and environment aspects (physical situation) or the current user activity (personal situation)

c) knowledge chunks: determining the set of situation-relevant information, services or behaviors [2].

In our article, we are dealing with context-aware knowledge chunks. Let us start with the base element – location.

There are many different approaches for getting location info for mobile subscribes. In general, it could be pretty standard nowadays (GPS, cell-id, assisted GPS [3]), but everything is getting more complicated as soon as we need indoor positioning. Due to the signal attenuation caused by construction materials, the Global Positioning System (GPS) loses significant accuracy indoors. Instead of satellites, an indoor positioning system (IPS) relies on nearby anchors (nodes with a known position), which either actively locate tags or provide environmental context for devices to sense. The localized nature of an IPS has resulted in design fragmentation, with systems making use of various optical, radio, or even acoustic technologies [4].

Nowadays, a great number of technologies are being used for indoor localization, such as Wi-Fi, RFID etc. However, all of them require the utilization of their own API with their own protocols. This can be a big challenge for developing heterogeneous scenarios where different localization systems have to be used for a location service.

One of the most used approaches to indoor location is Wi-Fi based positioning. A standard Wi-Fi based positioning system, such as the one offered by Ekahau [4] is completely software-based and utilizes existing Wi-Fi access points installed in a facility and radio cards already present in the user devices. Companies could deploy also Wi-Fi based radio tags that use industry standard components that adhere to the 802.11 standards. This approach allows for the use of commercial off-the-shelf hardware and drivers to produce a standards-based radio tag that can communicate bi-directionally over the 802.11 networks.

Thus, a standard Wi-Fi based positioning system can realize any type of location-aware application that involves PDAs, laptops, bar code scanners, voice-over-IP phones and other 802.11 enabled devices. For embedded solutions, there is no need for the client to include a specialized tag, transmitter, or receiver.

Because of the entire use of standards-based hardware, such as 802.11b, 802.11g, and 802.11a, a standard Wi-Fi based solution rides the installed based and economies of scale of the networks and end user devices that are proliferating today. Without the need for additional hardware, a company can install the system much faster and significantly reduce initial and long-term support costs. A common infrastructure supports both the data network and the positioning system, something companies strive for. The positioning system works wherever there is Wi-Fi coverage.

ISBN: 978-1-61208-193-9

In addition to cost savings in hardware, a standards Wi-Fi based positioning system significantly reduces the potential for RF interference. The total Wi-Fi positioning system shares the same network along with other network clients, so there is no additional installation of a separate wireless networks (as RFID requires) that may cause RF interference with the existing wireless network [5]. The cited article shows that any commodity 802.11's equipment is surprisingly vulnerable to certain patterns of weak or narrow-band interference. This enables to disrupt a link with an interfering signal whose power is 1000 times weaker than the victim's 802.11 signals, or to shut down multiple access points, multiple channel managed network at a location with a single radio interferer.

Wi-Fi location positioning is based on a grid of Wi-Fi hotspots providing, in general, 20–30 meters location accuracy. For more accuracy, there needs to be more access points. There are many articles devoted to Wi-Fi positioning. For example, we can combine a reference point-based approach with a trilateration-based one etc. Several layers of refinement are offered based on the knowledge of the topology and devices deployed. The more data are known, the better adapted to its area the positioning system can be [6].

Lets us mention also one more interesting approach: collaborative location (CL). And the most interesting approach for our future development is Collaborative Location-sensing. Cooperative Location-sensing system (CLS) is an adaptive location-sensing system that enables devices to estimate their position in a self-organizing manner without the need for an extensive infrastructure or training.

Simply saying, hosts cooperate and share positioning information. CLS uses a grid representation that allows an easy incorporation of external information to improve the accuracy of the position estimation. [7]

The motivation for CL and CLS is very transparent. In many situations, due to environmental, cost, maintenance, and other obstacles, the deployment of a dense infrastructure for location sensing is not feasible. It is exactly what we wrote about infrastructure-less system. In CLS, hosts estimate their distance from their neighboring peers. This can take place with any distance estimation method available (e.g., using signal strength). They can refine their estimations iteratively as they incorporate new positioning information.

Another are that is interesting for our approach is dynamic location based services. For example, AROUND [8] architecture is proposed as an approach for supporting location-based services in the Internet environment. AROUND provides a service location infrastructure that allows applications to select services that are specifically associated with their current location. The architecture includes a flexible scope model that defines the association between services and location, and a service location infrastructure organized by spatial criteria and optimized for location-based queries.

At this point, we are ready to make the last proposition before switching to the SpotEx model. Of course, the acronym LBS (Location Based Systems) contains the word "location". But, do we really need the location for the most of the services? As seems to us, the final goal (at least for the majority of services) is to get data related to the location, rather than location itself. Location in the classical form (latitude, longitude) here is just an intermediate result we can use as key for some requests for obtaining data (our final goal). So, why do not request data directly if we can estimate location?

II. SpotEx

What if we stop our traditional indoor positioning schema on the first stage: detection of Wi-Fi networks? This detection actually already provides some information about the location – just due to local nature of Wi-Fi network. And as the second step we add the ability to describe some rules (if-then operators, or productions) related to the Wi-Fi access points. Our rules will simply use the fact that the particularly Wi-Fi network is detected. And based on this conclusion we will open (read – make them visible) some user-defined messages to mobile terminals. Actually it is a typical example for the context aware computing. The visibility for user-defined text (content) depends on the network context.

The first time this service SpotEx (Spot Expert [9] developed by Dmitry Namiot) was described by the authors in article published in NGMAST-2011 proceedings [10]. This paper describes the next development in this approach as well as outlines the nearest plans.

Obviously, our SpotEx model is based on the ideas of Wi-Fi proximity. Wi-Fi host spots work here as presence sensors. But we are not going to connect mobile users to the detected networks and our suggestion does not touch security issues. We need only SSID for networks and any other public information.

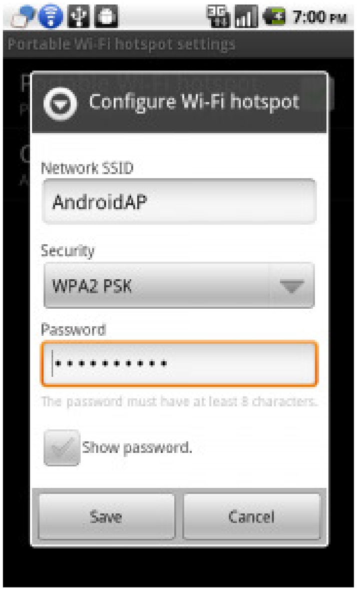
So, our service contains the following components:

- database (store) with productions (rules) associated with Wi-Fi networks
- rule editor. Web application (including mobile web) that lets users add (edit) rule-set, associated with some Wi-Fi network
- mobile applications, that can detect Wi-Fi networks, check the current conditions against the database and execute productions

How does it work? We can take any exiting Wi-Fi network (or networks especially created for this service – the most interesting case, see below) and add some rules (messages) to that network. Message here is just some text that should be delivered to the end-user's mobile terminal as soon as the above-mentioned network is getting detected via our mobile application. The word "delivered" here is a synonym for "available for reading/downloading".

The possible use cases, including commercial deployment are obvious. Some shop can deliver deals/discount/coupons right to mobile terminals as soon as the user is near some predefined point of sale. We can describe this feature as "automatic check-in" for example. Rather than directly (manually or via some API) set own presence at some place (e.g., similar to Foursquare, Facebook Places, etc.) and get deals info, with SpotEx mobile subscriber can pickup deals automatically. Campus admin can deliver news and special announces, hyper local news in Smart City projects could be tight (linked) to the public available networks and delivered via that channel etc.

Especially, we would like to point attention to the most interesting (by our opinion, of course) use case: Wi-Fi hot spot being opened right on the mobile phone. Most of the modern smart phones let you open Wi-Fi hot spots. We can associate our rules to such hot spot (hot spots) and so our messages (data snippets) become linked to the phones. Actually, we are getting dynamic LBS here: phone itself could be moved and so, the available data will be de-facto moved too.



Configure Wi-Fi hotspot

Network SSID

AndroidAP

Security

WPA2 PSK

Password

••••••••••

The password must have at least 8 characters.

☑ Show password.

Figure 1. Wi-Fi host spot on Android

This use case is probably the most transparent demonstration of SpotEx model. We can open “base” network right on the mobile phone, attach (“stick”) rules for the content to that network and it is all do we need for creating a new information channel. There is no infrastructure except the smart phone and we do not need a grid of devices as per CLS models.

Note again that this approach does not touch security and connectivity issues. You do not need to connect mobile subscribers to your hot spot. SpotEx is all about using hot spot attributes for triggers that can discover the content. The term Wi-Fi proximity is used sometimes in connection with Wi-Fi marketing and mean on practice just setting a special splash screen for hot spot that can show some advertising/branded messages for users during the connection to that hot-spot. Unlike this SpotEx threats Wi-Fi hot spots just as sensors.

How our productions data store (base of rules) looks like? Each rule looks like a production (if-then operator). The conditional part includes the following objects:

Wi-Fi network SSID,
signal strength (optionally),
time of the day (optionally),
client ID (see below).

In other words it is a set of operators like:

IF network_SSID IS ‘mycafe’ AND time is 1pm – 2pm THEN { present the coupon for lunch }

Because our rules form the standard production rule based system, we can use old and well know algorithm like Rete [11] for the processing. A Rete-based expert system builds a network of nodes, where each node (except the root) corresponds to a pattern occurring in the left-hand-side (the condition part) of a rule. The path from the root node to a leaf node defines a complete rule’s left-hand-side. Each node has a memory of facts, which satisfy that pattern. This structure presents essentially a generalized tree. As new facts are asserted or modified, they propagate along the network, causing nodes to be annotated when that fact matches that pattern. When a fact or combination of facts causes all of the patterns for a given rule to be satisfied, a leaf node is reached, and the corresponding rule is triggered [12].

The current implementation for mobile client based on Android OS. This application uses *WiFiManager* from Android SDK - the primary API for managing all aspects of Wi-Fi connectivity. This API let us pickup the following information about nearby networks:

SSID - the network name.

BSSID - the address of the access point.

capabilities - describes the authentication, key management, and encryption schemes supported by the access point.

frequency - the frequency in MHz of the channel over which the client is communicating with the access point.

level - the detected signal level in dBm.

So, actually all the above-mentioned elements could be used in our productions. And now we can prepare rules like this:

IF network_SSID IS ‘mycafe’ AND level > -60db AND time is 1pm – 2pm AND network_SSID ‘myStore’ is not visible THEN {present the deals for dinner}

Block {present the deals for dinner} is some data (information) snippet presented in the rule. Each snippet has got a title (text) and some HTML content (it could be simply a link to external site for example). Snippets are presenting coupons/discounts info for malls, news data for campuses etc.

Technically any snipped could be presented as a link to some external web site/mobile portal or as a mobile web page created automatically by the rule editor included into SpotEx. Rule editor works in both desktop and mobile web. So, once again, just having an ordinary smart phone is enough for creating (opening) information channel for delivering hyper-local news data.

In case of presenting our data as links to some existing mobile sites (portals) SpotEx works as some universal discovery tool. De facto, it lets mobile subscribers to be aware about context-relevant web resources. Owners for the web resources can describe own sites via rules rather then present for them individual QR-codes or NFC-tags for example.

In case of describing some content right in the SpotEx the whole system works in this part as a content management system. SpotEx rule editor creates mobile web page for the each provided data snippet and hosts that page on the own server. It means by the way, that for presenting our data we can use any resources that could be presented on HTML pages. In particularly, any multimedia content is also supported.

SpotEx mobile application, being executed, creates dynamic HTML page from titles (according to rules that are relevant in the given context) and presents that mobile web page to the user. It works just as a classical rule based expert system: matches exiting rules against the exiting context and makes the conclusions. Existing content here is a description for "Wi-Fi environment": list of hot spots with attributes. And conclusion here is a list of titles that can be presented as a dynamically created mobile web page. On that page each discovered title could be presented as a hyperlink that points to the appropriate data snipped. Any click on the interested title opens the snippet (shows or discovers data to mobile user).

So, for the mobile users, the whole process looks like browsing, where their browser becomes aware about hyper-local content. It is a typical example of context-aware retrieval. Context-aware retrieval can be described as an extension of classical information retrieval that incorporates the contextual information into the retrieval process, with the aim of delivering information relevant to the users within their current context [13].

The context-aware retrieval model includes the following elements:

- a collection of discrete documents;
- a set of user's retrieval needs, captured in a query;
- a retrieval task, to deliver the documents that best match the current query, rated on the basis of a relevance measure;
- the user's context, used both in the query formulation and associated with the documents that are candidates for retrieval.

It is obviously, that all the above-mentioned tasks are components of SpotEx.

As per other functionality of our context-aware browser we can highlight the following notes. At the first hand, we can note that it is the "pull model", versus the "push model" that proposed by Bluetooth marketing for example. And it could be more convenient (more safe) for the users – there are no automatically downloaded files/messages etc. But in the same time nothing prevents us from updating that dynamic web page automatically (e.g., by the timer) and simulating "pull model" in the user-safety mode.

At the second hand, we can note that because it is browsing, the whole process is anonymous. Indeed, there is no sign-in in the SpotEx. Of course, any data snippet may lead to some business web site/portal, where that site may ask about login, etc., but the SpotEx itself is anonymous. Unlike social networks like Foursquare you do not need to disclose your identity just for looking mall's deals for example.

But in the same time we still can collect some meaningful statistics in SpotEx. Because the model requires Wi-Fi to be switched on, we have automatically unique ID for the each client. It is MAC-address. It is actually global UUID. So, where we have not login info for our clients, we still can distinguish them. It let us detect for example, the same person, who did that already twice during the last week, opens that the particular data snipped.

Because mobile users in SpotEx model actually work with web pages, we can use pretty standard methods for web server log analysis for discovering user's activities.

A statistical analysis of the server log may be used to examine traffic patterns by time of day, day of week etc. So, we can detect frequent visitors, usage patterns, etc. And even more – we can use that information in our rules. E.g., some mall may offer special things for frequent visitors, etc. Data from real time analytics for our info snippets could be used in conditional parts of our rules.

The next stage of development targets the simplicity of preparing data for SpotEx model. What if instead of the separate database with rules (as it is described above) we add the ability to provide a special markup for existing HTML files?

So, rather than writing separate if-then rules we can describe our rules right in HTML code. Technically, we can add for example HTML div blocks with attributes that describe our rules (their conditions). Now, using some JavaScript code we can loop over such div blocks and simply hide non-relevant from them. For doing that we need to make sure that our JavaScript code is aware about the current context. We can achieve that via a special light implementation of local web server. This web server, being hosted right on the mobile phone (on the Android in our case) responds actually only to one type of requests. It returns the current context (Wi-Fi networks) in JSON (JSONP) format.

Why do we need a web server? It lets us stay in the web domain only. There is a simple and clear instruction for web masters:

- add SpotEx script to your page

```
<script type = ”text/javascript” src =
http://localhost:8080/spotex.js> </script>
```

- describe your info snippets as div blocks:

```
<div rel=”spotex” net=”WiFi_SSID” levelMin=”” levelMax=””>
        Your HTML code
```

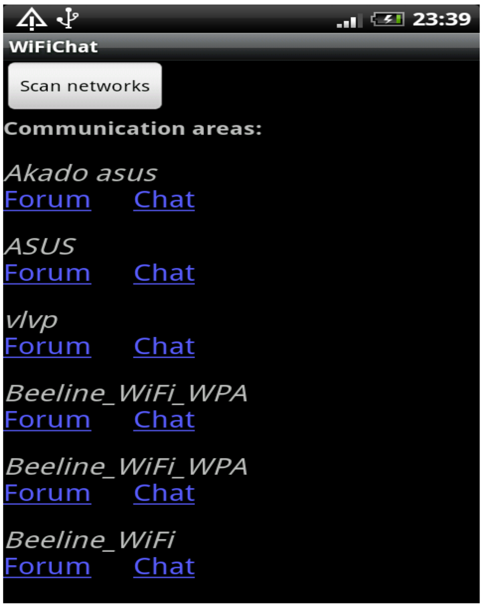
```
</div>
```

Our “old” rules could be presented via collection of attributes.

In this case, JavaScript code loaded from local server will be able to proceed all the div blocks related to SpotEx, and set visibility attributes depending on the context.

Such simple trick let us make any existing HTML page “Wi-Fi context aware”. Note that if our script is not available, the page will work as a “standard” HTML page.

There is also a “side” effect for SpotEx application – WiFiChat service [14]. This mobile application uses the principles described in this article and offers communication tools (web chat and discussions groups) for mobile users nearby the same Wi-Fi access point. Think about it as “SpotEx with predefined content”. The typical use case – we have Wi-Fi network in the train and this application automatically provides the discussions forum for the passengers. Or, keeping in mind that the “base” Wi-Fi network for this service could be opened right on the phone, this application can present personal forum (classified for example) as well as web chat for phone owner. This Android application is actually a wrapper for web mashup that combines HTML5 web chat engine and cloud based forums from Disqus:



Communication areas:

Akado asus
Forum Chat

ASUS
Forum Chat

vlvp
Forum Chat

Beeline_WiFi_WPA
Forum Chat

Beeline_WiFi_WPA
Forum Chat

Beeline_WiFi
Forum Chat

Figure 2. Wi-Fi Chat application

It is the typical tool for the ad-hoc communications on the go.

III. THE FUTURE DEVELOPMENT

Here, we see several almost obvious steps. At the first hand, it is open API. In the current implementation SpotEx front-end actually obtains data in JSON (JSONP) format from our server-side database.

As soon as API is going live, the next step is almost mandatory. It should be something that will simplify the development. The good candidates here are web intents [15] Web Intents is a framework for client-side service discovery and inter-application communication. Services register their intention to be able to handle an action on the user's behalf. Applications request to start an action of a certain verb (for example share, edit, view, pick, etc.) and the system will find the appropriate services for the user to use based on the user's preference. It is the basic.

Intents play the very important role in Android Architecture. Three of the four basic OS component types - activities, services, and broadcast receivers - are activated by an asynchronous message called as intent.

Intents bind individual components to each other at runtime (you can think of them as the messengers that request an action from other components), whether the component belongs to your application or another.

Created intent defines a message to activate either a specific component or a specific type of component - an intent can be either explicit or implicit, respectively.

For activities and services, an intent defines the action to perform (for example, to "view" or "send" something) and may specify the URI of the data to act on (among other things that the component being started might need to know). For example, our intent might convey a request for an activity to show an image or to open a web page. In some cases, you can start an activity to receive a result, in which case, the activity also returns the result in an Intent (for example, we can issue an intent to let the user pick a list of nearby images and have it returned to us - the return intent includes data in some format)

Going to our context aware browsing it means that our mobile devices will be able to present local data without low-level programming.

Web Intents puts the user in control of service integrations and makes the developers life simple.

Here is the modified example for web intents integration for the hypothetical web intents example:

1. Register some intent upon loading our HTML document

```
    document.addEventListener("DOMContentLoaded",
function() {
        var regBtn = document.getElementById("register");
        regBtn.addEventListener("click", function() {
    window.navigator.register("http://webintents.org/spotex",
undefined);
        }, false);
```

2. Start intent’s activity and pass it extra data (context info)

```
        var startButton =
document.getElementById("startActivity");
        startButton.addEventListener("click", function() {
          var intent = new Intent();
          intent.action = "http://webintents.org/spotex";
```

```
      intent.putExtra("WiFi_List", List_Of_Networks);
        window.navigator.startActivity(intent);
     }, false);
```

3. Get local info snippets (note – in JSON rather than XML) and display them in our application

```
     window.navigator.onActivity = function(data) {
       var output = document.getElementById("output");
       output.textContent = JSON.stringify(data);
     };
   }, false);
```

Obviously, that it is much shorter than the long sequence of individual calls as per any Open API.

The key point here is *onActivity* callback that returns JSON formatted data. Additionally, web intents based approach is asynchronous by its nature, so, we do not need to organize asynchronous calls by our own.

Also, we are planning to add Bluetooth measurements too. But, by our vision, we should avoid the typical Bluetooth usage cases and does not use push proxy as per classical Bluetooth marketing. We think that the end users do at least not welcome push approach and it is the source of problems with Bluetooth proximity. Vice versa, in SpotEx Bluetooth nodes will be used the same manner we are using Wi-Fi access points – as presence triggers. In other words, we will add the ability to describe rules for Bluetooth nodes too.

SpotEx approach could be extended also towards accumulating some ideas from the collaborative locations. We can add trilateration terms (conditions) to our rules, but present them in terms of fuzzy logic (close than, relatively close, etc.). It helps us incorporate grid data in case of many devices without any infrastructure preparation.

The next area we are going to pay attention to is Wi-Fi Direct specification. Wi-Fi Direct devices can connect directly to one another without access to a traditional network, so mobile phones, cameras, printers, PCs, and gaming devices can connect to each other directly to transfer content and share applications anytime and anywhere. Devices can make a one-to-one connection, or a group of several devices can connect simultaneously. They can connect for a single exchange, or they can retain the memory of the connection and link together each time they are in proximity [16].

As per Wi-Fi Direct spec a single Wi-Fi Direct device could be in charge of the Group, including controlling which devices are allowed to join and when the Group is started. All Wi-Fi Direct devices must be capable of being in charge of a Group, and must be able to negotiate which device adopts this role when forming a Group with another Wi-Fi Direct device. The device that forms the Group will provide the above described dynamically assembled web page with discovered services. It is how SpotEx could be extended to Wi-Fi Direct.

IV. Conclusion

This paper describes a new context-aware browsing model for mobile users developed on the ideas of Wi-Fi proximity. Service can use existing as well as the especially created (described) Wi-Fi networks as presence triggers for discovering user-defined content right to mobile subscribers.

The proposed approach is completely software based. It is probably its biggest advantage. For using SpotEx you need nothing except the smart phone. So, there are no prior investments in the hardware. Also this approach supports ad-hoc solutions and does not require the upfront space preparations.

This service could be used for delivering commercial information (deals, discounts, coupons) in malls, hyper-local news data, data discovery in Smart City projects, personal news, etc.

Acknowledgment

Many thanks to the anonymous ICDT-2012 reviewers for the reading and suggested improvements to the paper.

References

[1] G. Schilit and B. Theimer Disseminating Active Map Information to Mobile Hosts. IEEE Network, 8(5) (1994) pp. 22-32

[2] C. Bolchini1, G. Orsi, E. Quintarelli, F. A. Schreiber, and L. Tanca Context modeling and context awareness: steps forward in the context-addict project. IEEE Data Eng. Bull., 34(2): pp. 47–54

[3] R. Ahas, S. Silm, O. Järv, E. Saluveer, and M Tiru Using Mobile Positioning Data to Model Locations Meaningful to Users of Mobile Phones; Journal of Urban Technology Volume 17, Issue 1, 2010 pp. 3-27

[4] Comparison of Wireless Indoor Positioning Technologies http://www.productivet.com/docs-2/Wireless_Comparison.pdf, <retrieved: Jan, 2012>

[5] R. Gummadi, D. Wetherall, B. Greenstein, and S. Seshan Understanding and mitigating the impact of RF interference on 802.11 networks. SIGCOMM '07 Proceedings of the 2007 conference on Applications, technologies, architectures, and protocols for computer communications ACM New York, NY, USA ©2007 ISBN: 978-1-59593-713-1 DOI=10.1145/1282380.1282424, pp. 385-396

[6] F. Lassabe, P. Canalda, P. Chatonnay, and F. Spies "Indoor Wi-Fi positioning: techniques and systems" Annals of Telecommunications Volume 64, Numbers 9-10, pp. 651-664

[7] C. Fretzagias and M. Papadopouli Cooperative Location-Sensing for Wireless Networks, Proceedings of the Second IEEE International Conference on Pervasive Computing and Communications (PerCom'04), p.121, March 14-17, 2004.

[8] R. José, A. Moreira, H. Rodrigues, and N. Davies The AROUND Architecture for Dynamic Location-Based Services in Mobile Networks and Applications Volume 8, Number 4, pp. 377-387, doi: 10.1023/A:1024531629631.

[9] SpotEx Project: http://spotex.linkstore.ru, <retrieved: Jan, 2012>

[10] D. Namiot and M. Schneps-Schneppe About location-aware mobile messages International Conference and Exhibition on. Next Generation Mobile Applications, Services and Technologies (NGMAST), 2011 14-16 Sept. 2011 pp. 48-53 doi: 10.1109/NGMAST.2011.19

[11] E. Friedman-Hill Jess in action: rule-based systems in Java. Manning Publications Co. Greenwich, CT, USA 2003 ISBN: 9781930110892

[12] Charles L. Forgy, "RETE: A fast algorithm for the many pattern/many object pattern match problem", Artificial Intelligence 19(1): pp. 17-37, September 1982

[13] P. J. Brown and G. J. F. Jones. Context-aware retrieval: Exploring a new environment for information retrieval and information filtering. Personal Ubiquitous Computing, 5(4): pp. 253–263, 2001.

[14] Wi-Fi chat Project: http://wifichat.linkstore.ru, <retrieved: Jan, 2012>

[15] Web Intents: http://webintents.org/, <retrieved: Feb, 2012>

[16] Wi-Fi CERTIFIED Wi-Fi Direct http://www.cnetworksolution.com/uploads/wp_Wi-Fi_Direct_20101025_Industry.pdf, <retrieved: Jan, 2012>

Targeting Situational Awareness beyond the Event Horizon by Means of Sensor Element Munition

Tapio Saarelainen
Department of Military Technology
National Defence University, Finland
Helsinki, Finland
tapio.saarelainen@mil.fi

Abstract — **The main input of this paper is to examine a solution for acquiring data from beyond Event Horizon in an area of interest while operating in low-level operations in a given battlespace. As the pace of modern warfare increases, so does the necessity for maintaining accurately and timely updated Situational Awareness as well. Especially in tactical operations, the need for relevant reconnaissance data is critical in fostering effective decision making, and in this data collecting and analyzing process sensors capable of being deployable above an enemy territory play an important role. Versatile military operations in the modern battlespace strive for real-time information about enemy actions. Sensors capable of detecting seismic, acoustic and magnetic phenomena can be deployed to hostile areas with the assistance of mortars and howitzers. This paper describes basic principles concerning Sensor Element Munitions (SEMs) and discusses utilizing Sensor Elements (SEs) capable of sensing motion, magnetic, infrared and electro-optical phenomena and transmitting the accrued data to command posts in real-time to offer data and tools for rapid decision making to facilitate mission success. Rapidly deployable airborne SEMs represent versatile tools for low-level battalion and company operations.**

Keywords - Situational Awareness (SA); Common Operational Picture (COP); Sensors; Sensor Element Munitions (SEMs); Sensor Elements (SEs).

I. Introduction and Definitions

This paper introduces a method for accruing data for military troops operating in tactical level in a battlespace, namely, Sensor Element Munitions (SEMs) with encased Sensor Elements (SEs) . The objective of the introduced idea is to foster the means and technologies which increase the possibilities to facilitate collecting data for an improved Common Operational Picture (COP) utilizable in tactical operations. This rapidly deployable reconnaissance element, SEMs, is introduced as a tool for a comprehensive approach to perimeter control and intelligence, surveillance and data gathering in tactical level operations. This paper discusses how to utilize SEMs in military operations carried out in versatile battlespaces. Finally, the significance of the concept of Situational Awareness (SA) beyond the Event Horizon is discussed in relation to mission success in tactical level operations.

This paper tackles the following three research questions: 1) How to allow rapid collecting of data beyond the event horizon necessary for tactical troops with the assistance of Sensor Element Munitions? 2) What is the composition of SEMs? And lastly, 3) How to forward these gathered data to the given troops rapidly and reliably?

In particular in low-level operations, namely, those of a company and battalion, military commanders must be able to maintain an optimal COP facilitated by an equally optimal overall SA. Some of the practical data gathering means enabling fulfilling these data requirements involve issues such as Blue Force Tracking (BFT), Target Identification (TID) and Combat Identification (CID) as discussed in [1]. The term CID can be defined as a process of attaining an accurate and timely characterization of detected objects in the joint battle space to the extent that high confidence, timely application of military options and weapon resources can occur [2]. The collected data can thus be forwarded by using available existing network systems to a given entity. A definition for SA applicable is verbalized in the Army Field Manual 1-02 (September 2004): “Knowledge and understanding of the current situation which promotes timely, relevant and accurate assessment of friendly, competitive and other operations within the battle space in order to facilitate decision making. An informational perspective and skill that fosters an ability to determine quickly the context and relevance of events that is unfolding.” Now, to improve SA to ensure mission success, tools and concepts applied in Net Centric Warfare (NCW) environments can be utilized. The end-state aims at merging the data collected from a finite array of sensors and sources. SA comprises three levels: 1) perception, 2) comprehension and 3) projection [3]. Operationally, SA, or lack of it, remains a key factor in military operations and intelligence capabilities [3], [4]. SA is linked to Dismounted Battle Command System (DBCS) [5] and to Blue Force Tracking.

For the purposes of this paper, the term Event Horizon is used to denote the level which transcends the level of traditional reconnaissance capabilities of low-level military commanders. This is, commanders in battalion and company levels in militaries of small countries lack the capability of exploiting reconnaissance tools, such as Unmanned Aerial Vehicles (UAVs) and satellite services. Therefore, it is essential to introduce quickly deployable means to gather intelligence data, the means and tools that do not require procuring new types of weapons or materiel to overburden the organization in question.

Data beyond the Event Horizon refers to data collected beyond the visual horizon. In case of troops and operations concerning the battalion and lower levels, this collecting of data beyond the visual horizon is impossible because of the lacking reconnaissance tools described above. The pace of warfare is hectic in these low-level operations with multiple encounters in a tight time-frame. Thus more precise data are required for improved SA to ensure effective and timely decision making. This involves accounting for the phenomena taking place in the electromagnetic spectrum based on these observations and, moreover, the accrued data have to be in operational use in a matter of minutes. Indications of shots, explosions and acoustic and visual signatures of vehicles and their locations and actions are needed to create a Common Operational Picture (COP).

Typically, a COP comprises three types of modules: 1) information gathering sources that observe events and report information to the command and control module, 2) a command and control module that makes decisions based on both information received directly from its information gathering sources and information reported by other peers, and 3) display units at the emergency location that receive instructions from the command and control module [4].

The core capability in an optimal SA is a COP that fosters effective decision making, rapid staff actions, and appropriate mission execution [4]. The COP is employed to collect, share and display multi-dimensional information to facilitate collaborative planning and responding to security incidents.

This paper discusses recent research in regard to possibilities for increasing SA to ensure mission success in low-level military operations in battlespace. The remaining of the paper is arranged as follows: Section II introduces related work, Section III describes the composition and characteristics of SEMs and their utilization, Section IV explains the characteristics of SEMs together with the communication process, Section V concentrates on the targeting process, Section VI deals with the possibilities to analyze the collected data with Section VIII concluding the paper.

II. Related work

This paper is linked to three major areas researched by armed forces. Firstly, the key issue concerning military troops is their efficiency, which can be gained via an improved operational setting involving optimal SA, BFT and Command, Control, Computers, Communication, Information and Intelligence, (C^4I^2) [6]. Secondly, efficiency in military operations asks for optimized target identification, gained via utilizing the electro magnetic spectrum [7]. Thirdly, performance in low-level operations is currently being extensively examined, especially in conscript-based armed forces

In low-level operations, only minimal time is allocated for gaining SA data or waiting for orders. Systems applied in this level have to be simple, easy to use, and rapidly deployable. An example of this type of an operation is a dismounted company attack, where the structure of an attack process can be seen as a chain of planned events. This process requires particular services, which can be allocated to the requester of a service only if the service is available and within reach. The overall process of a company attack is explained in Figure 1.

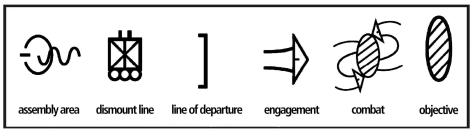


Figure 1. Company attack as a process.

Secondly, as militaries search for effective, rapid and reliable means to collect and analyze SA data in the battlespace, the realm of Wireless Sensor Networks (WSNs) [8] is relevant. As noted, WSN can be quickly-deployed, suitable for unattended monitoring and unnoticeable, representing an ideal choice for military applications [8]. For this particular purpose, remote, ground-based electronic sensors, used to collect intelligence on enemy movements and manoeuvres have been available for decades [9]. Yet, the reconnaissance and beneficial utilization of various types of sensors and sensor networks continue to be applicable. By adopting suitable sensors for appropriate platforms, the critical data can be gathered from the battlespace early enough to foster executing versatile military operations. Improved SA remains a key issue for small units operating in versatile military operations in a given battlespace and thus new and rapid means to collect data continue to remain necessary. For instance, the Finnish Defence Forces is developing its own sensors for improved skills in surveillance and intrusion detection systems to replace the anti-personnel land mines.

Obviously, all armed forces look for enhancing the performance and agility of their troops. This paper discusses a solution, the Sensor Element Munitions (SEMs) that can be produced by utilizing existing COTS-technology. The sensing elements of SEMs represent inexpensive, rapidly deployable means and draw from COTS-products to facilitate data accruing behind the event horizon.

III. The Set-Up and Utilization of Sensor Element Munitions

The data collecting and reconnaissance carried out by means of SEMs take after the standard High Explosive ammunition used in mortars and howitzers. The main difference is in the payload, in which the explosive charge has been replaced with a parachuted Sensor Element (SE). This SE is strong enough to withstand the forces of acceleration of a regular munition. The munition is delivered to a hostile area with similar procedures as standard High Explosive munition. The SE acting as a payload will be exhausted from the ammunition shell while airborne.

Structurally, the SE comprises a power source, an array of sensors, a transmitting unit and a relay-unit. The SE can act simultaneously in two roles: in accruing data and in a relaying role between two SEs. The SE does not receive data, but only transmits the data gathered, including GPS-data of its own position. The SE comprises sensors such as a visual sensor applicable to monitor targets both in daylight and low-light conditions as well as in the darkness. The central sensor is forward-looking infrared (FLIR) which is an important sensor for its advantages in night vision in securing military camps, reassuring soldier security, and detecting suspected terror activities in the battlespace [10].

The SE carries an image intensifier element and low-light sensors. It also features shortwave infrared (SWIR) and longwave infrared (LWIR). In addition, the SE includes detection elements for sensing acoustic, seismic and magnetic interference. Moreover, the sensor package features detection elements capable of detecting infrared and the movement of an individual and a wheeled or tracked vehicle.

Once the SE has been ejected out of the munition, it immediately starts to gather and transmit information from its area to own troops in an Ultra-Wideband using the frequency of 2,4 GHz for securing the transmission QoS. Another suitable method for transmitting the data is Worldwide Interoperability for Microwave Access (WiMAX), which is based on IEEE 802.16 standard utilizing frequencies of 4,4 -5,0 GHz.

The WiMAX standard 802.16d is applicable for slowly moving users whereas 802.16e is tailored for mobile users [11] and therefore we concentrate on Portable (Mobile) WiMAX, 802.16e, the channel sizes of which are 5 MHz, 8.75 MHz and 10 MHz. The usable WiMAX, 802.16e is based on orthogonal frequency division multiplexing (OFDM), orthogonal frequency division multiple access (OFDMA) [12]. In short, WiMAX combines OFDMA, an advanced multiple-input multiple-output (MIMO) as well as beamforming (BF) features [13]. These features together offer flexible bandwidth and fast link adaptation, creating a highly efficient air interface exceeding the capacity of existing 3G radio access networks [13]. These systems are suitable for military surveillance applications.

In low-level operations in the battlespace, data collecting can be facilitated by SEMs. The shell of SEMs can be manufactured of either steel, composite, or heat-treated plastics. SEMs can be deployed to a target area either by a howitzer or a mortar. In what follows, we take a closer look at SEMs and examine the processes of munition deployment, data collecting and data distribution.

Firstly, the ammunition shell of SEMs can be manufactured of various materials. One of these is composite, originally tailored for ballistic protection. The benefits of this material are its strength and suitability for munition core material in that it is lighter than steel and easily forged into the desired shape and structure. When munition is lighter, the payload can be heavier, if desired. Light-weight munition can be deployed further behind enemy lines by using the same charge as in steel munition. Moreover, composite represents a material, which can be surfaced with materials capable of absorbing radar beams, making the SEM less visible in enemy counter-artillery radars. This means that SEMs and the SEs are invisible on the screens of an adversary’s counter-artillery radars while SEMs are being deployed to enemy territory by air. Furthermore, as the SE is made of composite, when it hovers in the air above an adversary, the SE manufactured of composite is less visible compared to an SE made of traditional steel. In short, an adversary receives no early warning of the incoming munition and is unlikely to be capable of locating either the positions of an artillery weapon or the howering elements early enough. Therefore, it is highly unlikely that any counter measures be executed for there are no indications of any oncoming actions whatsoever. Figure 2 illustrates the structures of various types of SEMs with the encased SEs.

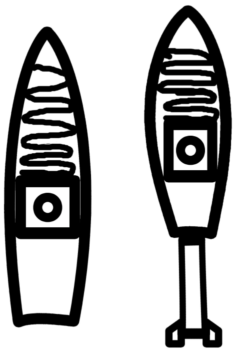


Figure 2. Structures of Sensor Element Munitions: An artillery SEM (left), a mortar SEM (right). Artillery and mortar shells can be manufactured of various materials. Composite-manufactured Sensor Elements (SEs) are packed inside the munitions together with their parachutes.

Secondly, the tactical use of the SEM-based reconnaissance system is as follows: 1) When reconnaissance data beyond the Event Horizon are needed, a commander issues the order to deploy the munition to the target area, 2) mortars or/and howitzers perform the tasks to the designated areas, 3) the SE transmits the data to own receivers, 4) resulting improved SA is utilized in commanding troops and shooters to the designated areas or targets to maximize the performance of own troops (and gain the initiative). If more data are wanted, the described phase two can be repeated and more data can be gathered. This process is explained in Figure 3.

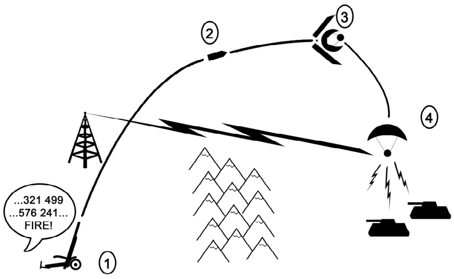


Figure 3. On deploying an SE above an enemy territory: 1) Fire Support Order is issued, 2) the SEM is airborne, 3) the SEM opens and ejects the SE, 4) the SE starts to transmit gathered data from the enemy territory and targets.

Once the critical data have been collected they have to be quickly analyzed to be used for evaluating different Courses of Action (COAs). Success depends on an accurate mission analysis and a timely evaluation process of the accrued data. Improved SA results in optimal time for mission execution and simultaneous minimizing of casualties, which increases efficiency and leads to minimum recovery times improving the overall efficiency and performance capabilities of the troops utilized.

Once commanders have access to more current reconnaissance data for mission execution, they are able to analyze different COAs and calculate the pros and cons to evaluate the best possible method to operate in any scenario prevailing. As explained in Figure 4, military commanders have by default value at least two different options for executing the mission in question. Having completed Military Decision Making Process (MDMP), the most effective operation can be executed to maximize the performance of the designated troops. In the described scenario below, the commander focuses the performance on incapacitating the Command Post (CP), the alternative number 2, instead of attacking against the armored enemy.

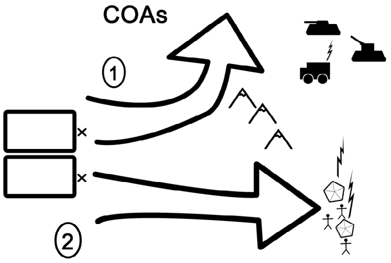


Figure 4. Possibilities of COAs.

Figure 5 explains the basic process of data gathering beyond the Event Horizon, especially in operations suitable for low-level troops. The deployed and hovering sensor element acts like loitering munition, sensing and measuring the prevailing electromagnetic spectrum, collecting and transmitting data to the receiver-station. An antenna can be installed both in a fighting vehicle and on the ground, depending on the prevailing combat-situation.

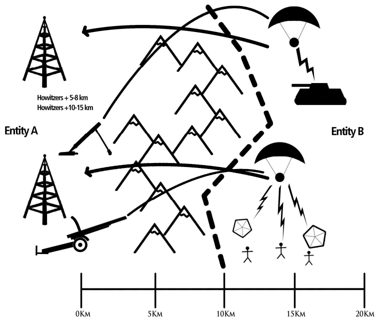


Figure 5. The data gathering process for improved SA beyond the Event Horizon and the transmitting of these data to own troops.

Time itself is a critical resource in this type of reconnaissance process. In order to avoid wasting time, the signals and data must be transmitted reliably from the SE to the receiver-station. For this purpose, the SE utilizes smart antenna technology meaning that antenna transmission and propagation pattern can be optimized for the optimal outcome. The accrued data can be transmitted reliably to the receiver, and because of the clear Line-Of-Sight (LOS), there is only a limited number of obstacles or attenuation disturbing the coded transmission process from the sensor to the receiver-station. This method is described in Figure 6.

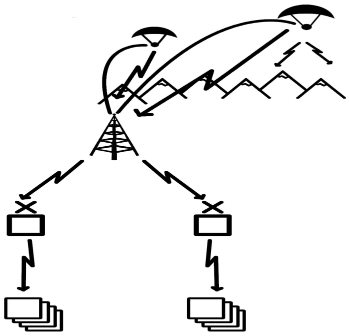


Figure 6. A method for transmitting the accrued data to the commanded troops to ensure successful operations.

IV. On Airborne sensors, SEMs and Communication

In any military operation, airborne sensors are important for missions, such as force protection, perimeter control and intelligence utilization [9]. Transmitting the accrued data to prevent fratricide and ensure success in operations presupposes optimal communications. WiMAX transmission offers applicable possibilities in forwarding collected data. The distances in the transmission process are relatively short, ranging from 1 kilometer to 20 kilometres in conditions of clear Line-Of-Sight.

The sensor package inside SEMs, namely the SE, can be made of existing COTS-products comprising sensors capable of sensing most of the phenomena occurring in the electromagnetic spectrum. In general, COTS-products are relatively inexpensive and reliable in terms of function, as explained in [14]. An SE of a SEM comprises the following sensors: acoustic-, seismic-, magnetic-, visible image-, shortwave infrared (SWIR)-, thermal-, infrared-, low-light television (LLTV)-, and sensors for laser tracking and spotting and sensors for facial recognition. In terms of automatized identification and verification, a facial Recognition System represents a computer application capable of automatically identifying or verifying a person from a digital image. One possibility is to compare selected facial features from the captured image with an existing facial database.

The analyzing centre has the capability for the fusion of all the accrued sensor information. The sensor package comprises a short lifetime battery, which can produce energy for the sensor package for the duration of 4 – 6 minutes (howering time). The battery can be equipped with capacitors or electric double-layer capacitors (EDLC) if the required energy level is inadequate with the selected sensor package.

Once an SE is airborne, it immediately starts to transmit the gathered data to own troops either directly or, if the transmission distance exceeds the capability of the transmission unit, the SE transmits the data to another airborne device, which acts as a relay station in relation to own troops. The SE communicates with the receiver station and other sensor element packages over a 2,4 GHz Ultra-Wideband Network system. The accrued data are crypted for security reasons.

V. Comprehensive Targeting Process

The cycle of a complete targeting process can be described as Detect, Identify, Decide; Engage and Assess (DIDEA) [4]. The cycle is outlined in Figure 7 below. The DIDEA provides an iterative, standardised and systematic approach supporting targeting and decision making, being generic enough to be used as a systematic process for C2 node targeting and decision making. This process is thoroughly discussed in [4].

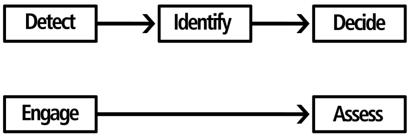


Figure 7. The simplified DIDEA process.

The decision as to whether or not to open fire is based on the visual signature of a given uniform, weapon and gear as well as magnetic, seismic or acoustic signals identified by a sensor [9]. Self-evidently, the transmission of combat-critical location and identification data play a crucial role in the battle space. After the accrued data have been transmitted and received, they flow through a dissemination process, where these data are analyzed and fused to form a COP and to increase the overall SA. Figure 8 explains the process of Signature Prediction Process (SPP). As described in Figure 8, sensors accrue data and transmit these data for analyzing centers. The data collected with SEMs are verified with the data gathered with other sensors to in order to predict and anticipate the type of target and its actions.

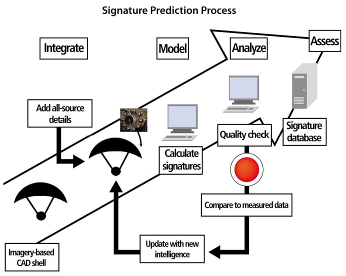
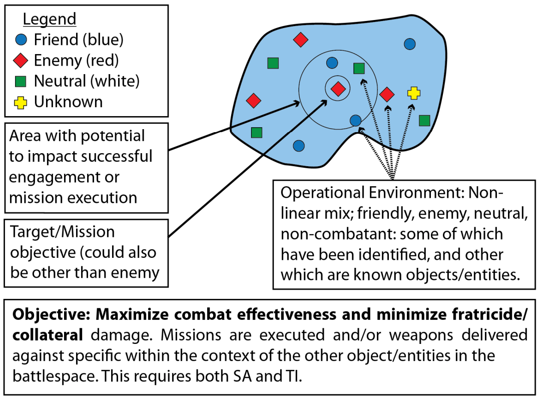


Figure 8. The Signature Prediction Process, typical of several surveillance and detection systems.

The destruction power of a given weapon system has to be optimized to account for the enemy location (forest, open area, Urban Territory), the state of movement on-the Move (OTM) or at-the-halt (ATH), and the protection-level (mounted, dismounted, dug).

The cruel reality remains that an executive commander is necessarily aware of fact the there is always the possibility of fratricide and collateral damage. Figure 9 emphasizes the importance of accurate and timely SA around the target area. The shooter has to be aware of the locations and status of both own troops and the enemy. It is critical to optimize the destruction power of a weapon system along the identification of a target. When the target represents a hierarchically critical enemy commander, he or she can be incapacitated by transmitting the coordinates and visual signature to the designated shooter, as indicated in Figure 9.


Legend
● Friend (blue)
◆ Enemy (red)
■ Neutral (white)
✚ Unknown
Area with potential to impact successful engagement or mission execution
Target/Mission objective (could also be other than enemy
Operational Environment: Non-linear mix; friendly, enemy, neutral, non-combatant: some of which have been identified, and other which are known objects/entities.

Objective: Maximize combat effectiveness and minimize fratricide/ collateral damage. Missions are executed and/or weapons delivered against specific within the context of the other object/entities in the battlespace. This requires both SA and TI.

Figure 9. The importance of accurate and timely SA around the target area.

If the commander has Close-Air Support (CAS) available, he or she can utilize the performance of the data analyzing centre (indicated as a satellite dish in Figure 10), the collected data can be forwarded directly to the shooter. This process improves the overall performance and saves time and utilizes the performance of a designated fighter (cf. Figure 10 below). As for the receiving antenna constellations, they can be both ground- and vehicle-born systems.

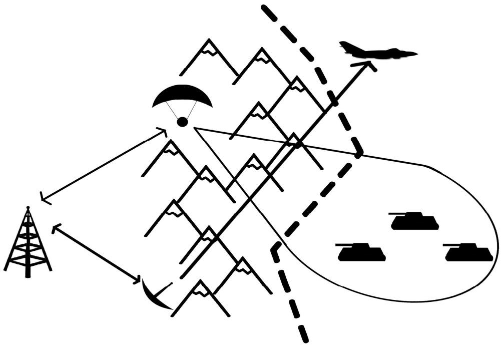


Figure 10. The process of detecting enemy forces and forwarding the accrued data to an Analyzing Center and finally to the shooter.

In a critical case when the target is a very important human being, he or she can be incapacitated by transmitting the required precise data of the target and its location to the sniper or to a team of snipers. A cellular telephone can act as a receiver. The figure caption of a sniper's cellular phone is depicted in Figure 11.

Figure 11. Adequate target identification data transmitted to an individual member of Special Forces for eliminating purposes.

VI. Means to Analyze Collected Data

When it comes to transmitting data, the following issues have been identified. As tested in [13], an 802.16e WiMAX Testbed has provided throughputs of 5.75 Mbps Upstream with a modulation and coding of 64 QAM ¾ [13]. These amounts of data seem adequate to receive all the required sensor data.

With the assistance of automated targeting programs and classifiers, it is possible to recognize faces, find hidden and concealed targets, and look for essential information by means of computation algorithms [15]. Classifiers, such as a Support Vector Machine (SVM), K-Nearest Neighbourhood classifier (KNN), and BP neural classifier, can be utilized in battlefield target identification [7]. The recognition of an end-user can be based on visual biometrics and the most conventional identification, the computer-assisted recognition of human face [16]. The ubiquitous networks and sensor data can act as assisting tools in detection, recognition and especially in target classification [15]. Moreover, the data produced by various multi-sensors can be utilized in the data refinery process to ease the recognition and identification process with the assistance of data fusion processes by resorting to computer-programs designed for data fusion processes [17]. In fact, when using the K-Nearest-Neighbour (KNN) algorithm, approximately 80 % of unknown target samples can be recognized correctly, when the known target classification accuracy remains above 95 %. This enables the use of the ATR and the Automatic Target Cuer (ATC). Face recognition schemes that combine wavelet transform, SVM and clustering can be exploited identifying human beings [18].

As for object categorization, high-definition closed-circuit television (CCTV) cameras feature many computer controlled technologies that allow them to identify, track, and categorize objects in their field of view. As defined in [19], WiMAX mobile technology is a good candidate in supporting the CCTV applications in the context of mobile users. Furthermore, Video Content Analysis (VCA) represents the capability of automatically analyzing a video to detect and determine temporal events not based on a single image. Moreover, a system utilizing a VCA can recognize changes in the environment and identify and compare objects in the database using size, speed, and color. Also, VCA analytics can be used to detect unusual patterns in a video's environment. The system can be set to detect anomalies in a crowd of people and a VCA also has the ability to track people on a map by calculating their position from the images.

VII. Conlusions

This paper has introduced a robust SEMs-based data collecting means for gathering data beyond the Event Horizon. This approach draws from the utilization of existing sensors and WiMAX technology. The composition and functionality of SEMs have been introduced. The results of this paper offer a method to improve SA, COP, CID, and TID. The concepts of SEM and SE have been introduced in Figures 2 – 3. The introduced solution is applicable in increasing the performance capabilities of modern troops.

Three research questions were raised in Section I, and the answers to questions 1 and 2 were provided in Section III with visualizations in Figures 2 and 3, and question 3 was addressed in Section III with visualizations in Figures 5 and 6.

As evident, in all operations and in low-level tactical operations in particular, critical Situational Awareness data have to be collected rapidly, since mission success is time-dependent. Figures 4 – 6 concentrated on describing the data accruing process and utilization of data in military operations. Once data have been collected, a battle can be won only by careful mission planning, comparing different COAs and rapidly executing successful operations. Figures 7 – 9 illustrated the process of targeting. Thereby the adoption of existing COTS-technologies, when appropriately applied, offers a key to ensuring the desired success.

Operational time spent in the battlespace can be minimized by careful mission analysis and thorough evaluation of Courses of Actions (COAs). Critical data have to be forwarded to shooters, as illustrated in Figures 10 – 11.

So far, all the decision-making processes in battlespace settings have culminated in a human being making the final decision to apply performance in missions. In the future, this decision maker's position may be manned by Artificial Intelligence (AI). AI can be benefitted as an assisting power of a commander to ensure mission success. Figure 12 describes this process in brief.

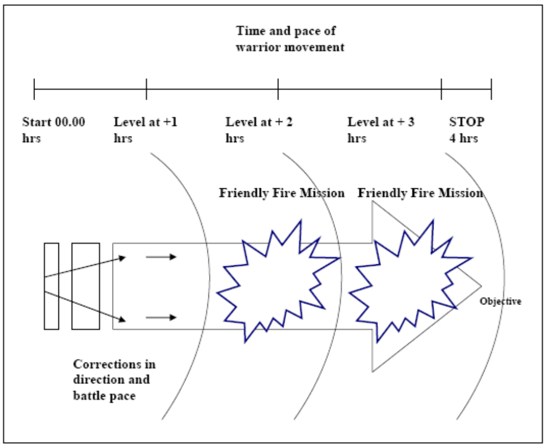


Figure 12. Once COP and SA have been fused, a computer can assist in fostering the movement of own troops along the pace and direction.

This mid-term solution remains applicable until the current armed forces are being replaced by robotic militaries of the future. Before this, however, humans have to continue coping with their own intelligence assisted with optimal data accruing and analyzing tools in order to be able to make judgements that keep incorporating both the probability of success and affordable costs.

REFERENCES

[1] T. Saarelainen, Enhancing Situational Awareness by Means of Combat-ID to Minimize Fratricide and Collateral Damage in the Theater, The Sixth International Conference on Digital Telecommunications, (ICDT2011), Budapest, 17 – 22.4.2011, Hungary.

[2] T. Saarelainen, White Force Tracking, in Proceedings of the *8th European Conference on Information Warfare and Security, ECIW2009*, 6 – 7 July 2009, Lisbon, Portugal, pp. 216 – 223.

[3] M. Endsley, and Connors, E., Situation awareness: State of the art, in Proceedings of *Power and Energy Society General Meeting - Conversion and Delivery of Electrical Energy in the 21st Century*, 20-24 July 2008, Pittsburg, PA, pp. 1 – 4, doi 10.1109/PES.2008.4596937

[4] T. Saarelainen, and J. Timonen, Tactical Management in near real-time Systems, in Proceedings of *IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (Cogsima 2011)*, CogSIMA2011, Miami Beach, Fl, U.S.A., pp. 240 - 247, doi 10.1109/COGSIMA.2011.5753452.

[5] T. Saarelainen, and J. Jormakka, Computer-Aided Warriors for Future Battlefields, The 9th European Conference on Information Warfare and Security ECIW2009, Lisbon, 6 – 7 July, 2009, Portugal, pp. 224 – 233.

[6] Saarelainen, T. and J. Jormakka, C4I2-Tools for the Future Battlefield Warriors, in Proceedings of *IEEE International Conference on Digital Communications (ICDT)*, 13 –19 June 2009, Athens, Greece, pp. 38 – 43, doi 10.1109/ICDT.2010.15.

[7] J. Li, C. Zhang, and Z. Li, Battlefield Target Identification Based on Improved Grid-Search SVM Classifier, International Conference on Computational Intelligence and Software Engineering, 2009 (CiSE 2009), 11-13 Dec. 2009, pp. 1 – 4, Doi 10.1109/CISE.2009.5365100.

[8] T. Jing, Y. Shengwei, Y. Bing, and M. Shilong, Study On Wireless Sensor Networks, International Conference on Intelligent System Design and Engineering Application (ISDEA), 13-14 Oct. 2010, Vol 2, ISBN 978-1-4244-8333-4, pp. 510 – 521, doi 10.1109/ISDEA.2010.392.

[9] P. Buxbaum, Denying Access, *Special Operations Technology*, July 2010, Vol 8, Issue 5, pp.26 – 27.

[10] I. Zafar, U. Zakir, I. Romanenko, R. Jiang, and E. Edirisinghe, Human Silhouette Extraction on FPGAs for Infrared Night Vision Military Surveillance, IEEE, International Conference onSecond Pacific- Asia Conference on Circuits, Communications and System (PACCS), 2010, pp. 63 – 66, doi 10.1109/PACCS.2010.5627025.

[11] J. Martin, B. Li, W. Pressly, and J. Westall, WiMAX Performance at 4.9 GHz, pp. 1 – 7, doi 10.1109/AERO.2010.5446943.

[12] M. Al-Adwany, A performance study of wireless broadband access (WiMAX), 1st International Conference on Energy, Power and Control (EPC-IQ), Nov. 30 2010-Dec. 2 2010, pp. 320 – 324, INSPEC Accession Number: 11989694.

[13] K. Etemad, Overview of Mobile WiMAX Technology and Evolution, IEEE Communications Magazine, October 2008, pp. 31 – 40, doi 10.1109/MCOM.2008.4644117.

[14] R. Kozma, L. Wang, K. Iftekharuddin, E. McCracken, M. Khan, K. Islam, R. Demir, Multi-modal sensor system integrating COTS technology for surveillance and tracking, Radar Conference, 10-14 May 2010, pp. 1030 – 1035, doi 10.1109/RADAR.2010.5494467.

[15] Frankot, R.T., Performance of cued target acquisition systems: Impact of automatic target recognition, in Proceedings of *42nd Conference on Asilomar Signals, Systems and Computers (ACSSC2008)*, pp. 1619 – 1623, doi 10.1109/ACSSC.2008.5074697.

[16] R. Luo, Y. Chou, T. Chung, C. Liao, C. Lail, and C. Tsai, NCCU Security Warrior: An Intelligent Security Robot System, in Proceedings of *33rd IEEE Conference on Industrial Electronics Society (IECON 2007)*, 5-8 Nov. 2007, Taipei, pp. 2960 – 2965, doi 10.1109/IECON.2007.4460380.

[17] Libiao, T. ; Lu, W., Qi, D., Kai, Z., Target Recognition Based on Seismic Sensors and Neural Network, in Proceedings of *8th International Conference on Electronic Measurement and Instruments (ICEMI '07)*, Aug. 16 2007-July 18 2007, Xi'an, pp. 1-18 – 1-21, doi 10.1109/ICEMI.2007.4350418.

[18] Luo, B; Zhang, Y., and Pan, Y-H., Face recognition based on wavelet transform and SVM, Conference on Information Acquisition, in Proceedings of *IEEE International Conference on Information Acquisition*, 27 June-3 July 2005, doi pp. 373 – 377, doi 10.1109/ICIA.2005.1635115.

[19] M. Aguado, E. Jacob, J. Matias, C. Conde, and M Berbineau, Deploying CCTV As an Ethernet Service Over the Wimax Mobile Network in the Public Transport Scenario, International Conference on Communications Workshops, 2009. (ICC Workshops 2009), 14-18 June 2009, pp. 1 – 5, doi 10.1109/ICCW.2009.5208011.

Investigations of Resource Allocation Schemes Between Multi-hop Backhaul Network and Access Network

Wang, Xinglin; Tan, Minqiang; Yang, Xiaokun; Li Zheng
Research and Technology Department
Nokia Siemens Networks Technology (Beijing) Co. Ltd.
Beijing, China
xinglin.wang@nsn.com; minqiang.tan@nsn.com;
xiaokun.yang@nsn.com; zheng.li@nsn.com

Cheng, Xiaohui
Department of mechanical and electrical engineering
Beijing Vocational College of Labour and Social Security
Beijing, China
huigirl@126.com

***Abstract*—In this paper, the trade-off on resource allocation between multi-hop backhaul and access is investigated. Multi-hop, treated as a special case of mesh, is very useful in non-contention based network. We assume in-band relay backhaul and access share the same resource. Based on calculation and simulation, the relationship among resource allocation, cell coverage and channel status is revealed under two relay schemes.**

Keywords-Multi-hop; resource allocation; Shannon capactiy; in-band;out-band; relay.

I. Introduction

Wireless multi-hop networks have attracted lots of attentions in recent years as the next evolutionary step for wireless data networks. It is more feasible and effective than pure mesh structure, especially in non-contention based wireless network. Non-contention wireless network contains, such as Worldwide Interoperability for Microwave Access (WiMAX) [1], Long Term Evolution (LTE) [2], High Speed Uplink Packet Access (HSPA) [3] and so on. These networks carefully schedule the radio resource to avoid interference and efficiently utilize the radio resource. Pure mesh structure is not so easy to be implemented in such a network, due to synchronization, interference and so on.

Currently, WiMAX and LTE both setup relay work group to study how to build multi-hop backhaul in access cell.

WiMAX technology is becoming increasingly popular as a number of service providers are deploying WiMAX to provide wireless broadband connectivity to customers. IEEE 802.16j work group is focusing on multi-hop relay networks that will enable multi-hop communication in mobile WiMAX (IEEE 802.16e) networks. In such a network, mobile stations or subscriber stations may communicate with a Relay Station (RS) instead of communicating directly with the Base Station (BS) [4][5].

Similarly, people also proposed relay system in LTE-Advanced [6][7]. Generally, relay is essentially backhaul function plus access function in one node. LTE based wireless backhaul can be classified into in-band and out-band backhaul solutions. In-band backhaul, such as LTE in-band relay and IEEE802.16j [4][5], will share the radio resource with access. Out-band backhaul will use another independent radio resource from access. In this paper, we focus on in-band multi-hop relay.

In [9], the coverage and capacity of in-band relay in urban area were simulated. The realistic performance of relay in suburban area was illustrated in [10].

In this paper, we will build an Orthogonal Frequency Division Multiplexing (OFDM)-based two-hop relay network to cover most cases, e.g., LTE, WiMAX. Then we will study the resource allocation balance between backhaul and access under different cell radius, various channel status, and different relay schemes. The performance will be compared in in-band and out-band relay, so as to indicate the respective use cases.

This paper is organized as follows. In Section II, system model is described and two relay schemes are defined for multi-hop backhaul part. Here, we will investigate a two-hop in-band backhaul system with access. The assumptions are also given. In Section III, the calculation steps are given and a static system level simulation is built to help get the final results [8]. The obtained results are analyzed in detail as well. In Section IV, we get the final conclusions.

II. System Model

Since access and backhaul parts share the same radio resource, an efficient resource management will be very important to avoid congestion regardless in access or backhaul. Generally, there are two kinds of methods—dynamic allocation and static allocation.

Comparatively, dynamic allocation is more efficient. The system obtains the statistics of the access/backhaul requirement and the channel qualities, and calculates the resource trade-off between access part and backhaul part instantly.

Here, we assume an ideal dynamic allocation to avoid any congestion or unbalance between backhaul and access, which means that BS knows all instant channel information of all links and allocation granularity is very small.

We also assume two configurations at two-hop relay part. The first scheme is the traditional one, which means different relay backhaul links will use orthogonal resources. The second scheme allows relay backhaul links to utilize Spatial Division Multiplex Access (SDMA) to save radio resource. That means same resources can be spatially reused among different backhaul links with directional antennas.

Details are as follows.

A. Relay Scheme 1

We assume a cell deployment structure shown in Figure 1. All base stations (BSs) or relays are located at the center of hexagon cell. The access radio resource is separated into three parts as shown in Figure 2 and reused among the cells shown in Figure 1. Similarly, the backhaul resource is also reused as shown in Figure 1 and Figure 2. Different backhaul links will use orthogonal resources.

Additionally, we call the central cell as egress cell. In egress cell as shown in Figure 1, all access traffics from these 19 cells are collected and backhauled to this egress BS, which is called Donor eNB in LTE-A [6][7]. We treat this node as wireless backhaul egress, since generally there will be a fiber connection on this node to continue to backhaul all traffics to core network.

Around egress cell, there are 6 cells called 1st tier cells as shown in Figure 1. Around 1st tier, 2nd tier consists of 12 cells. Backhaul links connecting egress cell and 1st tier cells are called 1st hop, while those connecting 1st tier and 2nd tier are called 2nd hop.

Note that the first hop backhaul generally occupies more resources than the second hop due to the much more backhaul traffic. In Figure 1 thicker backhaul line means more radio resource occupation.

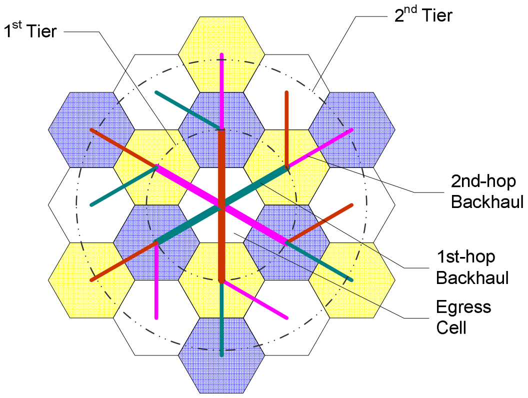


Figure 1. Cell structure of relay scheme 1.

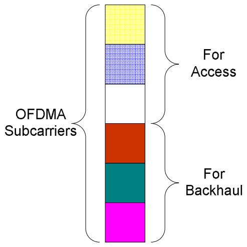


Figure 2. OFDM subcarrier assignment for relay scheme 1.

B. Relay Scheme 2

We assume another cell deployment shown in Figure 3. Similarly, the access radio resource is separated into three parts as shown in Figure 4 and orthogonally reused in Figure 3. Here, the backhaul resource is spatially reused as shown in Figures 3 and 4. Different backhaul links will use the same resource by for example Spatial Division Multiplex Access (SDMA) through directional antennas. Note that the first hop still occupies more resource than the second hop due to the much more backhaul traffic. In Figure 3, thicker backhaul line means more radio resource occupation.

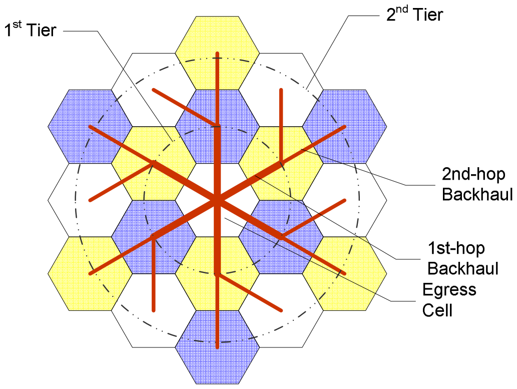


Figure 3. Cell structure of relay scheme 2.

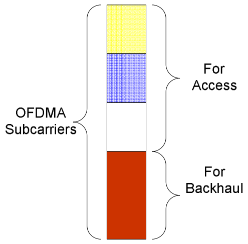


Figure 4. OFDM subcarrier assignment for relay scheme 2.

C. Assumptions in System Model

We assume that each cell has the same user density and the same traffic requirement of each user (calculated by Shannon Capacity). All cells have many users and are full-loaded.

Obviously, if a cell is close to the backhaul egress, the backhaul requirement will be much higher, because this cell has to backhaul not only its own traffic but also those of its child nodes. The cell close to the egress surely will consume more radio resource for backhaul.

Here, we assume all cells have the same radius *r*.

The other assumptions are as follows. All the cell access parts have the same path loss factor γ_A, and all the backhaul channels also have the same path loss factor γ_{BH}. Assuming carrier frequency is f_c; the transmit powers for access and backhaul are $\boldsymbol{P_A}$ and $\boldsymbol{P_{BH}}$ respectively; the noise power is $\boldsymbol{N}$. B_A and B_{BH} are the bandwidths for access and backhaul parts *in the second tier and second hop* respectively. The total bandwidth is ***B***. ***E(.)*** is the expectation operation.

The path loss for access in each cell is

$$PL_A(d,\gamma_A)[dB] = PL(d_0,\gamma_A)[dB] + 10\gamma_A \log\left(\frac{d}{d_0}\right) + X_\sigma \quad (1)$$

Similarly, the path loss for backhaul is

$$PL_{BH}(d,\gamma_{BH})[dB] = PL(d_0,\gamma_{BH})[dB] + 10\gamma_{BH} \log\left(\frac{d}{d_0}\right) + X_\sigma ,(2)$$

where $PL(d,\gamma) = \left(\frac{4\pi f_c}{c}\right)^2 d^{\gamma}$.

Here, d is the distance from transmitter to receiver (e.g., BS to User Equipment (UE)), d_0 is the reference distant, and X_σ is the shadow fading.

In following simulation and analysis, we follow the parameters defined in Table I.

TABLE I. SYSTEM PARAMETERS

Parameters		***Value***
Access channel path loss factor γ_A		3.5
Two-hop backhaul channel path loss factor γ_{BH}		2.5
Carrier frequency f_c		2.5GHz
Cell radius		1 km
Downlink transmit power of BS and relay		33dBm@BS; 18dBm@relay
Total bandwidth ***B***		10MHz
UE antenna gain		0dBi
BS or relay node antenna gain		11dBi
Noise Figure		5dB@BS or Relay, 9dB@UE
Traffic Density	district town	D=9.196bps/m^2
	semi rural area	D=1.522bps/m^2
	rural area	D=0.298bps/m^2
Antenna Configuration		SISO
Shadow fading standard deviation, X_σ		8dB

Since access downlink (DL) always has higher power and higher throughput, we only consider DL here.

III. CALCULATION AND SIMULATION RESULTS

Here, we build a static system level simulation, and use Monte-Carlo method to get the results according to [8].

We use Shannon Capacity to calculate the throughput for users and backhaul part.

A. *Relay Scheme 1*

At the second tier, in order to get the balance between access and backhaul parts, we have

$$B_A \cdot E\left[\log_2\left(1+\frac{P_A * G_A}{N * PL_A}\right)\right] = B_{BH} \cdot E\left[\log_2\left(1+\frac{P_{BH} * G_{BH}}{N * PL_{BH}}\right)\right] (3)$$

Here, B_A and B_{BH} are the bandwidths for access in each cell and backhaul parts in the second tier respectively. $\boldsymbol{G_A}$ and $\boldsymbol{G_{BH}}$ are the antenna gains in access part and in backhaul part respectively. As shown in Figure 1, the backhaul link in the first tier will transmit more traffic, including its own and its child cells'. Therefore, the backhaul bandwidth in first tier should be **3*$\boldsymbol{B_{BH}}$**, since it will backhaul three cells' traffic.

Finally, assuming the total bandwidth is B, according to Figure 1, we have

$$3 * B_A + 5 * B_{BH} \le B \quad (4)$$

It is hard to obtain a close-form result of the left side in (3). One way is to use static system level simulator to do Monte Carlo simulation [8].

An example is as follows. With the parameters in Table I and ***B***=10MHz, according to Monte Carlo simulation, we have

$$E\left[\log_2\left(1+\frac{P_A * G_A}{N * PL_A}\right)\right] = 1.91 \quad \text{and}$$

$$E\left[\log_2\left(1+\frac{P_{BH} * G_{BH}}{N * PL_{BH}}\right)\right] = 6.014 \quad (5)$$

According to (3), (4) and (5), we can obtain the bandwidth requirement for access and backhaul in the second tier as

$\boldsymbol{B_A}$**:** $\boldsymbol{B_{BH}}$=6.014:1.91

We use equal mark in (4) and obtain

$\boldsymbol{B_A}$=2.18MHz and $\boldsymbol{B_{BH}}$=0.6922MHz

For the egress cell and the cells in the first tier and the second tier, the resource for access in each cell is same, i.e., $\boldsymbol{B_A}$. For the 1st hop backhaul from the first tier to the egress cell, we require 3*$\boldsymbol{B_{BH}}$. For the 2nd hop backhaul from the second tier to the first tier, we require $\boldsymbol{B_{BH}}$.

According to the mentioned example, we can calculate the bandwidth for access and backhaul in case of different parameters.

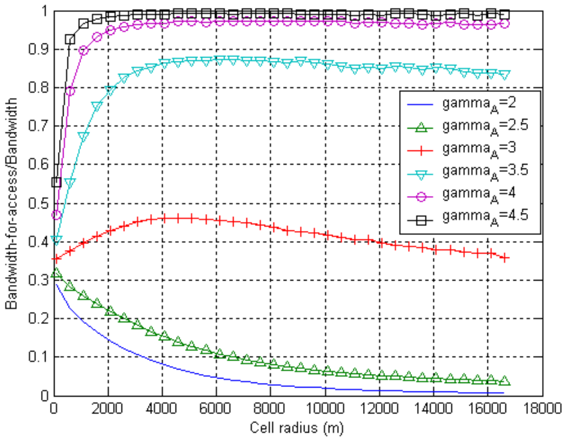


Figure 5. Proportion of Total Access Bandwidth over Total Bandwidth for Realy Scheme 1.

With the parameters in the mentioned example and different path loss factors γ_A in access (gamma$_A$ in Figure 5), we can get Figure 5. In Figure 5, the proportion of total access bandwidth in case of different path loss factors in access is shown, i.e., $(3*B_A)/B$. It is shown that we require more access resource with higher path loss in access.

B. Relay Scheme 2

At the second tier, in order to get the resource balance between access and backhaul parts, we have

$$B_A \cdot E\left[\log_2\left(1+\frac{P_A * G_A}{N * PL_A}\right)\right] = B_{BH} \cdot E\left[\log_2\left(1+\frac{P_{BH} * G_{BH}}{N * PL_{BH}}\right)\right] \quad (6)$$

Here, B_A and B_{BH} are the bandwidths for access in each cell and backhaul parts in the second tier. Similarly, the backhaul bandwidth in first tier should be **3*$\boldsymbol{B_{BH}}$**. Assuming the total bandwidth is ***B***, according to Figure 3, we have

$$3 * B_A + 3 * B_{BH} \le B \quad (7)$$

Similarly, we use Monte Carlo simulation to obtain the result of left side in (6).

An example is as follows. With the parameters in Table I and ***B***=10MHz, according to Monte Carlo simulation [8], we can get

$$E\left[\log_2\left(1+\frac{P_A * G_A}{N * PL_A}\right)\right]=1.91 \text{ and}$$

$$E\left[\log_2\left(1+\frac{P_{BH} * G_{BH}}{N * PL_{BH}}\right)\right]=6.014 \quad (8)$$

According to (6), (7) and (8), we can obtain the bandwidth requirement for access in each cell and backhaul in the second tier as

$\boldsymbol{B_A}$**:** $\boldsymbol{B_{BH}}$=6.014:1.91

We use equal mark in (7) and obtain

$\boldsymbol{B_A}$=2.53MHz and $\boldsymbol{B_{BH}}$=0.80347MHz

For the egress cell and the cells in the first tier and the second tier, the resource for access in each cell is same, i.e., $\boldsymbol{B_A}$. For the 1st hop backhaul link, we require 3*$\boldsymbol{B_{BH}}$ for each. For the 2nd hop backhaul link, we require $\boldsymbol{B_{BH}}$ for each.

With the parameters in the mentioned example and different path loss factors γ_A in access (gamma$_A$ in Figure 6), we can get Figure 6. In Figure 6, the proportion of total access bandwidth in case of different path loss factors in access is shown, i.e., $(3*B_A)/B$. It is shown that we require more access resource with higher path loss in access.

C. Simulation Results Analysis

According to Figure 5 and Figure 6, it is shown that when access has bad channel status, i.e., high path loss factor, access will occupy more bandwidth. In this case, backhaul will consume little bandwidth due to good channel status. If we meet large cell radius, access will occupy almost all bandwidth as shown in figures.

If access has relatively good channel status (γ_A=3), i.e., low path loss factor, the proportion of access bandwidth has a peak value. As shown in Figure 5 and Figure 6, the peak value of scheme 1 is about 0.46 at cell radius equal to about 4000 meters, while the peak value of scheme 2 is about 0.6 at cell radius equal to about 4500 meters.

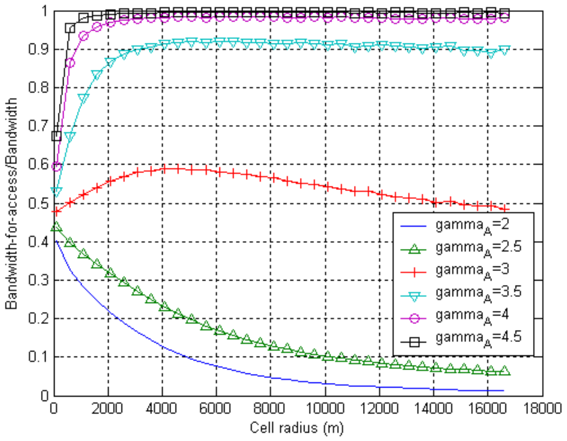


Figure 6. Proportion of Total Access Bandwidth over Total Bandwidth for Relay Scheme 2.

In this case, as cell radius enlarging, the proportion of access bandwidth will increase, but after achieving peak value, the proportion will decrease. The reason is that if the cell radius is too large, path loss in multi-hop backhaul part will relatively increase faster and require more bandwidth.

In a word, the radio resource trade-off between multi-hop backhaul and access yields different characteristics under different channel status.

D. Impact on Access Coverage

The throughput in each cell is $B_A \cdot E\left[\log_2\left(1+\frac{P_A * G_A}{N * PL_A}\right)\right]$, where PL_A is a function of distance. Here, we assume that the access traffic density is ***D***bps/m^2 shown in Table I. The access traffic requirement in a cell is $D*\frac{3\sqrt{3}}{2}r^2$ bps. With the parameters in Table I, we can get Figure 7 for scheme 1 and scheme 2. The capacity of traditional cell without *in-band* multi-hop backhaul (BH) is also shown (*"cell capacity w/o in-band backhaul" in Figure 7*).

Note that a cell without in-band backhaul can be a cell with out-band wireless backhaul or wired backhaul, such as fiber, ATM, and so on. We also call this cell as *traditional cell*.

In Figure 7, the real throughput requirements in each cell based on different traffic densities are shown (different ***D*** in Figure 7). It also shows the throughput provided by access cells in case of schemes 1 and 2.

Obviously, only if the provided cell throughput is larger than the real traffic requirement, the user communication can be satisfied, i.e., the three red lines in Figure 7 should be on the upper of the other lines. According to Figure 7, we can get the coverage radius limit under different cases in Table II.

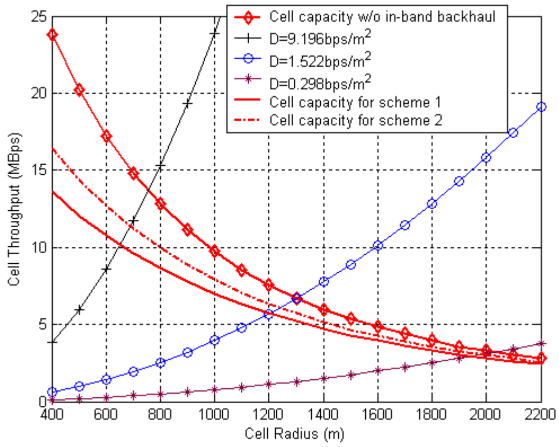


Figure 7. Cell Throughput and Throughput Requirement..

TABLE II. COVERAGE RADIUS LIMITATION

Senarios	district town D=9.196bps/m^2	semi rural area *D*=1.522bps/m^2	rural area *D*=0.298bps/m^2
Scheme 1	651m	1200m	1950m
Scheme 2	690m	1240m	1985m
w/o in-band BH (traditional cell)	755m	1301m	2036m

Thus in order to cover a specific area of 100km^2, the required number of BSs is listed in Table III. *Here, BS means BS or relay, i.e., any access node.*

TABLE III. NUMBER OF BSS TO COVER A SPECIFIC AREA OF 100KM^2

Parameters	district town D=9.196bps/m^2	semi rural area *D*=1.522bps/m^2	rural area *D*=0.298bps/m^2
Scheme 1	91	27	11
Scheme 2	81	25	10
w/o in-band BH (traditional cell)	68	23	10

From this table, we can see that schemes 1 and 2, i.e., in-band backhaul, are more suitable for rural area or the area of low traffic density, since in-band relay results in similar number of access nodes as traditional cells.

In urban area, scheme 1 and 2 result in much more BSs than traditional cell, which may cause cost increasing and more handoff overhead. However, scheme 2 causes fewer BSs than scheme 1, which means that SDMA among multi-hop backhaul is an efficient method to save radio resource.

In rural area or semi rural area, comparison between scheme 1 and scheme 2 shows that SDMA yield little gain. Scheme 2 even results in same number of BSs as traditional cell.

IV. CONCLUSIONS

In this paper, we studied a non-contention based OFDM in-band multi-hop system. Under two relay schemes, the allocation results for multi-hop backhaul and access are analyzed. It is shown that if access part has much worse channel status than backhaul part, the access will occupy more and more resources with increased cell radius. If backhaul part has similar channel status with access part, the access part will occupy more resources at the beginning, but the occupied resources will be decreased with continuing increased cell radius.

If we use SDMA at multi-hop backhaul part, resource will be saved, and relay can cover larger area. However, it is much more effective in urban area than in rural area. Further study revealed that in-band relay is more suitable for low traffic density area.

ACKNOWLEDGMENT

This work is sponsored by Nokia Siemens Networks Technology（Beijing）Co. Ltd.

REFERENCES

[1] IEEE Std 802.16-2004, “IEEE Standard for Local and metropolitan area networks Part 16: air inteface for fixed broadband wireless access systems,” Oct. 1, 2004.

[2] 3GPP TS 36.401, “E-URTAN Architecture description,” v8.6.0, Dec. 2009.

[3] 3GPP TS 25.308, “High Speed Downlink Packet Access (HSDPA); Overall description; Stage 2,” v8.11.1, Dec. 2011.

[4] S. W. Peters and R. W. Heath Jr., “The future of WiMAX: multi-hop relaying with IEEE 802.16j,” IEEE Communications Magazine, vol. 47, pp. 104-111, January 2009.

[5] IEEE 802.16J-2009, “Multihop Relay Specification,” June 2009.

[6] S. Parkvall, E. Dahlman, A. Furuskar, et al., “LTE-advanced—evolving LTE towards IMT-advanced,” in IEEE 68th Vehicular Technology Conference (VTC08 Fall), pp. 1–5, September 2008.

[7] 3GPP TR 36.814, “Further Advancements for E-UTRA, Physical LayerAspects,” v1.5.1, December 2009.

[8] 3GPP TR36.942, “Radio Frequency System Scenarios,” v8.1.0, Dec. 2008.

[9] Irmer, R. and Diehm, F., “On coverage and capacity of relaying in LTE-advanced in example deployments,” in IEEE 19th International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), pp. 1-5, Sept. 2008.

[10] Coletti, C., Mogensen, P., and Irmer, R., “Performance Analysis of Relays in LTE for a Realistic Suburban Deployment Scenario,” in IEEE 73rd Vehicular Technology Conference (VTC11 Spring), pp. 1-5, May 2011.

Assessment of LTE Uplink Power Control with Different Frequency Reuses Schemes

Mohamed M. El-Ghawaby
Electronics and Communications
AAST
Cairo, Egypt
ghawaby25@gmail.com

Hesham El-Badawy
Network Planning Department
National Telecom Institute
Cairo, Egypt
heshamelbadawy@IEEE.com

Hazem H. Ali
Electronics and Communications
AAST
Cairo, Egypt
hazemhali@gmail.com

***Abstract*— Single Carrier Frequency Division Multiple Access (SC-FDMA) is the access scheme chosen by 3GPP for uplink UTRAN Long Term Evolution project (LTE). As SC-FDMA provides intra-cell orthogonality, one of the main reasons for performance degradation is the Inter-Cell Interference (ICI). This degradation is accentuated by the frequency reuse of 1 deployed in the system, Since the Frequency Reuse (FR) and Power Control (PC) functionalities is a strong tool for co-channel interference mitigation, using them critical issues in cellular Orthogonal Frequency Division Multiple Access (OFDMA)/LTE networks. In this paper, we compare between the Open Loop Power Control (OLPC) and Closed Loop Power Control (CLPC) performance using different frequency reuse schemes. Simulation results show that large differences exist between the performance of different (FR) schemes and the optimal case in the overall cell throughput, as well as the cell-edge user performance. Also the closed loop power control has shown more cell and edge throughput gain over OLPC.**

Keywords*— *Open Loop Power Control; Closed Loop Power Control; Hard Frequency Reuse; Fractional Frequency Reuse; Soft Frequency Reuse.

I. INTRODUCTION

LTE introduces a number of innovations that, in aggregate, continue to push ever closer to the theoretical maximum data rates defined by Shannon's Law [4]. Advances in multi-antenna techniques, OFDMA methods, wider bandwidth, interference mitigation, and protocol efficiencies are fundamental to deliver the promise of 4G Mass Market Wireless Broadband. The amazingly high data rates and sector throughputs (capacity) per cell are fundamental to supplying the ever increasing demand for wireless broadband.

Effective reuse of resources in a cellular system can highly enhance the system capacity. With a smaller Frequency Reuse Factor (FRF), more available bandwidth can be obtained by each cell. So, in this sense the classical FRF of 1 is desirable see Fig. 2a. However, with the usage of FRF-1, the most User Equipments (UEs) are seriously afflicted with heavy ICI, especially near the cell edge. And that causes severe connect outages and consequently low system capacity. The conventional method to figure out this problem is through increasing the cluster-order, which can mitigate the ICI efficiently, nonetheless at the cost of a decrease on available bandwidth for each cell. This leads to restricted data transmissions and lower system spectrum efficiency. To take aim at improving cell-edge performance while retaining system spectrum efficiency of reuse-1.

There are many techniques which can be used to mitigate interference in E-UTRA uplink. The basic approaches are classified into different type such as Power Control, Inter-cell-interference randomization, Coordination/avoidance, and Frequency domain spreading.

Recent researches are focused at OLPC and CLPC performance evaluation. This is due to its capability of interference mitigation as well as increasing the system throughput. Many investigations for the performance and configurations of the OLPC and CLPC [4][7]. Results show that the different configuration is directly effect on both cell edge users and cell center users.

Also, many recent researches are focused at FR techniques such as Hard Frequency Reuse HFR, Fractional Frequency Reuse FFR, Soft Frequency Reuse SFR and performance evaluation and developing [5][6], Results show great performance, especially for the cell edge throughput due to interference mitigation.

ISBN: 978-1-61208-193-9

The current paper investigates about the ICI as a result of uplink PC and FR. In addition, it will combine between each PC techniques and the three FR schemes to achieve better performance.

The paper is organized as follows; Section II describes the general interference mitigation concepts for E-UTRA followed by detail description of OLPC, CLPC and the most famous frequency reuse schemes which will be used with both OLPC & CLPC. Section III is discussing proposed system model. Section IV illustrates results and its analysis. Finally, the conclusion is presented in Section V.

II. INTERFERENCE MITIGATION

PC and FR schemes are representing the main building blocks of the proposed system model.

A. Open Loop Power Control:

PC refers to set output power levels of transmitters, Base Stations (BSs) in the downlink and UEs in the uplink. A PC formula has been already agreed in a 3GPP meeting for the Physical Uplink Shared Channel (PUSCH) [2]. Fig. 1 is based on an OLPC algorithm and CLPC adjustments can also be applied.

The 3GPP specifications [3] defines the setting of the UE transmit power P for PUSCH by the following equation

$$P = \min\{P_{\max}, P_0 + 10 * \log_{10} M + \alpha * PL + \delta_{msc} + f(\Delta_i)\} \quad (1)$$

where $P_{\max}$ is the maximum UE transmit power, P_0 is a parameter that has a cell specific and nominal part. It is measured in dBm/Hz, expressing the power to be contained in one Physical Resource Block (PRB), *M is* the number of assigned PRBs to a certain user, α is the cell-specific path-loss compensation factor that can be set to 0.0 and from 0.4 to 1.0 in steps of 0.1, *PL* is the downlink path-loss measured in the UE, δ_{msc} is a UE-specific parameter (optionally cell-specific), and $f(\Delta_i)$ is a UE-specific close-loop correction value with a relative or absolute increase.

The scope of PC is to define the transmitting power in one PRB according to (1), letting the UE scale it to the assigned transmission bandwidth (BW). This implies that ultimately it will transmit with a constant power in each assigned PRB, For this reason, the term $10 \cdot \log_{10} M$ can be extracted from (1). Finally, removing the closed loop term, the Power Spectral Density (PSD) formula results in (2), which is referred to as the Fractional Power Control (FPC) formula.

$$PSD = P_0 + \alpha * PL \qquad \text{“dBm/Hz”} \quad (2)$$

It is preferred to work with the path gain information which is the linear inverse of the path loss. Then, (2) is re-written as (3) in dBm

$$PSD = P_0 - \alpha * PG \qquad \text{“dBm/Hz”} \quad (3)$$

where PG is the path gain of the user to the serving BS.

If α =0, a case referred to as **no compensation**. All UEs will transmit at full power which results in high interference level and poor cell edge performance. With α =1, a case referred to as **full compensation**. The equation reduces to traditional slow power control scheme where all UEs are received at the same power resulting in poor spectral efficiency. By letting 0<α <1, one can achieve both good edge performance and high spectral efficiency by letting UEs with good channel condition transmit at relatively low power level to reduce the interference. At the same time, UEs with bad channel condition are transmitting at relatively high power level to achieve high spectral efficiency.

Regarding to one of the references [4], we will use α =.8 and $P_0 = -81$ dBm/PRB which achieve both good edge and cell throughput.

- Impact on the CINR Distribution

The Carrier to Interference plus Noise Ratio (CINR) is one of the factors that determine the user throughput. Therefore, a discussion of the impact of the OLPC parameters on each UE experienced CINR would be very helpful for the operator. Let's define the experienced CINR per user

$$isd_j = \sum_{k=s(j)} E[psd_k] * pg_{k,j} \qquad \text{“mW/Hz”} \quad (4)$$

where, isd_j is the average interference spectral density perceived by a given BS, *s(j)* denotes the users not served by BS j and allocated to transmit on the observed PRB, psd_k is the power spectral density for user k which is not serving by the given BS, $pg_{k,j}$ is the path gain between user k and the given BS.

$$\text{CINR}_i = E\left[\frac{psd_i * pg_{i,s(i)}}{isd_{s(i)} + n}\right] \quad (5)$$

where n is the thermal noise.

ISBN: 978-1-61208-193-9

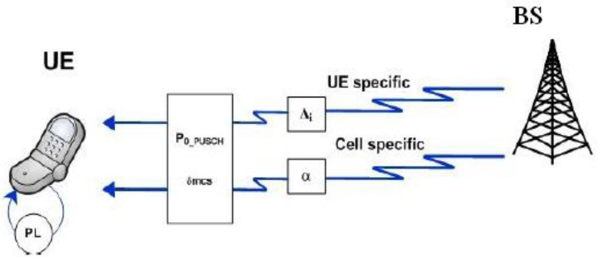


Figure 1. PUSCH power control parameters broadcasted by BS towards the UEs

- Impact on the Cell and Edge throughput

  In EUTRAN LTE UL, the Modulation and Coding Scheme MSC is chosen according to the state of CINR, higher orders are used when this is higher. Equation (6) shows how the user throughput is calculated for a given user from its experienced CINR and allocated bandwidth. [4]

$$C_i = BW_{eff} * v * M * BW_{PRB} * \log_2\left(1 + \frac{CINR_i}{S_{eff}}\right) \text{“bps”} \qquad (6)$$

where BW_{eff} is the bandwidth efficiency Set to 0.72, v is a correction factor set to 0.68, M is the number of allocated PRBs, BW_{PRB} is the bandwidth of one PRB Equal to 180 KHz, S_{eff} is the CINR efficiency at system level Set to 0.2 dB. By taking one PRB to be compatible with the fractional frequency reuse which will be discussed later, so there will be difference between our edge throughput and the reference edge throughput [4].

Equation (7) is to calculate the cell throughput.

*T=E[C]**total number of PRBs at the system “bps” (7)

where *T* is the cell throughput, *E[C]* is the average UEs throughput.
Edge throughput is the lowest 5 % of Cumulative Distribution Function (CDF) of the total cell throughput.

B. *Closed Loop Power Control:*

There are different techniques are used in CLPC because it does not have standardization. But the main idea of the closed loop is to start with OLPC then the UEs also sends feedback to the BS, which is then used to correct the user Transmitted T_X power.

There are two main techniques used for CLPC, Generalized Interference Based Power Control GI-PC, which take in the consideration the path loss to the serving BS, and the generated interference from the UEs to the neighbour BS.

The second technique is Cell Interference Based Power Control C-IPC, which proposed for each UE to have not less the minimum reference CINR.

In our work, we will use the GI-PC as the second PC reference.

The power spectral density can be obtained from (8)

$$PSD_i = I_0 - PG_s * \beta - PG_i * \gamma \text{ “dBm/Hz”} \qquad (8)$$

where I_0 is interference power spectral density limit, it work as p_0 in OLPC but the main difference is that I_0 is the power spectral density per hertz but p_0 is the total power contained in one PRB, PG_s is the path gain to the serving BS, PG_I is the path gain to the nearest interfered BS from the UE_i, β is a parameter that affects the impact of PG_s on the T_X PSD, γ is a parameter that affects the impact of PG_i on the T_X PSD.

- Impact on the CINR Distribution

  The CINR can be easily obtained same as OLPC but the main difference will be only in the PSD term.

$$S_i = \frac{I_0 * PG_s^{1-\beta}}{PG_I^{\gamma} * [I + N]} \qquad (9)$$

  where I is the average interference spectral density perceived by a given BS and N is the thermal noise.

  For the cell and edge UEs throughput will be the same as OLPC, other assumption will be at the Table 3.

C. *Frequency Reuses Schemes:*

There are three major techniques used

- Hard Frequency Reuse (HFR), hard frequency reuse splits the system bandwidth into a number of distinct sub-bands according to a chosen reuse factor and lets neighboring cells transmit on different sub bands see Fig. 2b.
- Fractional Frequency Reuse (FFR), Fractional frequency reuse [5] splits the given bandwidth into an inner and an outer part. The inner part is completely reused by all BSs, the outer part is divided among the BSs with a frequency reuse factor greater, as one seen in Fig. 2c.

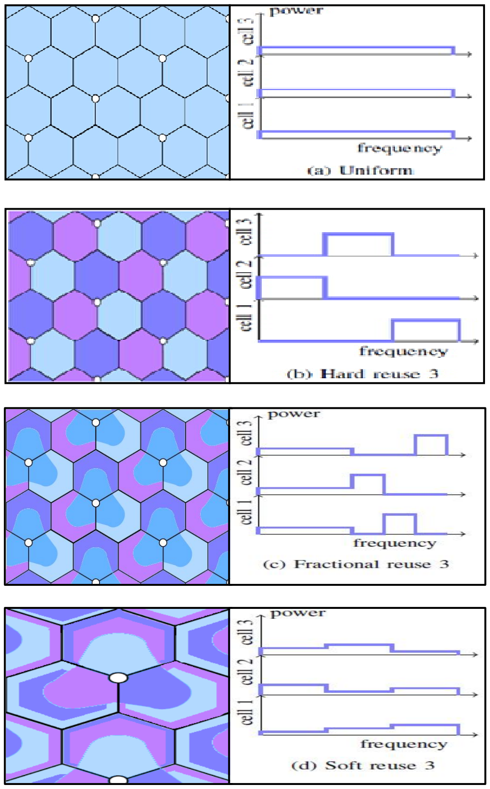


Figure 2. Different frequency reuses techniques

- Soft Frequency Reuse (SFR), soft frequency reuse [6][8][9], the overall bandwidth is shared by all base stations (reuse factor of one is applied), but for the transmission on each sub-carrier the BSs are restricted to a certain power bound see Fig. 2d.

III. PROPOSED SYSTEM MODEL

In this section, the system model is discussed; details are shown in tables 1, 2.

Following the 3GPP guidelines [1], the cell simulation layout consist of a wrap around Macro-cell scenario reference case 1; see Fig. 3. Composed by a grid of 19 sites with 3 sectors each (19 BS with 3 sectors, total cells are 19*3=57cells), each cell has 100 user, the inter site distance is 500 meters and each sector is modeled by a hexagon

The operating bandwidth is divided in 50 PRBs (48 PRB for users and 2 for signaling) with a bandwidth of 180 KHz each. There is a Maximal Ratio Combining (MRC) in the specifications, used to constructively combine the multiple received signals in the antennas. It is modeled here as a constant gain of 3 dB in the received signal.

The total path loss between an UE and a BS is modeled as in (10).

TABLE 1 SYSTEM MODEL DETAIL

Simulation case	**ISD meters**	**BW MHz**	**PLoss dB**	**Speed Km/h**
1	500	10	20	3

TABLE 2 SYSTEM MODEL DETAIL

Parameter	**Assumptions**
Cellular Layout	Hexagonal grid, 19 cell sites, 3 sectors per site (wrap around)
Distance-dependent path loss	L=128.1 + 37.6*log10(R) R in kilometers
Penetration Loss	20 dB
Antenna pattern(horizontal) (For 3-sector cell sites with fixed antenna patterns)	$A(\theta) = -\min\left(12 * \left(\frac{\theta}{70}\right)^2, 20\right)$
Shadowing modeled as a log-normal distribution (SF)	Mean =0, standard deviation= 8dB
Total path loss	*L+A(θ)+SF*
Max UE Tx power	24 dBm
Number of users in system	100*3*19=5700 user

$$PL=L+A(\theta)+SF \quad \text{“dB”} \qquad (10)$$

where *L* is the path loss between BS and UE, *A(θ)* is the modeled antenna gain and *SF* is the shadowing

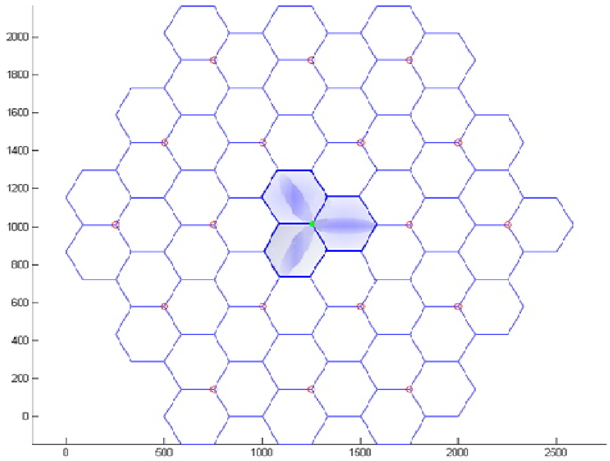


Figure 3. Wrap around Macro cell model

TABLE 3 PC PARAMETERS

Parameter	Value	Unit
Bandwidth efficiency	.72	bps/Hz
PRB bandwidth	180	KHz
Max UE power	250	mW
Number of PRBs per user	1	-
Outage	5	%
Thermal noise level	-174	dBm/Hz
Total number of PRBs	48+2 for signalling	-
Number of users per cell	100	user
MRC gain	3	dB
α	.8	-
P_0	-81	dBm/Hz
I_0	-157	dBm/Hz
β	.7	-
γ	.3	-

For the PC parameter we will take the same assumption as [4], except the PRB for each user will be one PRB, all parameters are shown in Table 3.

For the frequency reuse

- For HFR, we will divide the total used PRBs for the 3 sectors which will give 16 PRBs for each sector.
- For FFR, we will divide the total PRBs to two groups each group is 24 PRBs, 24PRBs for the centre UEs (Ues, which have path loss less than 120dB), and 24 PRBs is distributed to the three sectors (8PRBs for each sector for the UEs which have path loss more than 120dB).
- For SFR, we divide the total PRBs to three [9] groups, the first group include the UEs which have path loss less than 110dB, the second group include the UEs, which have a path loss between 110dB and 120dB and the last group include the UEs, which have a path loss more than 120dB.

IV. RESULTS AND ANALYSIS

The implementation and simulations are carried out using a multi-cell radio network dynamic simulator implemented in MATLAB to evaluate the PC with different FR schemes.

The results show that all techniques start from the lowest cell throughput and edge throughput and both of them increase to a certain point, peak edge throughput observed when the first user reaches the maximum UE power limitation. Sudden decreasing appears in edge throughput due to interference increasing regarding to the many UEs reach the maximum power limitation which leads to average PSD increasing, which is responsible of edge throughput decreasing.

We will divide the results to three main parts, validation results, OLPC with different FR schemes and CLPC with different FR schemes

A. *Validation results*

Fig. 4 shows a comparison between the obtained results and that had been presented in [4] in the same operational conditions. It is shown that the obtained results get more gains and have the same behaviour of [4] taking in the consideration that in [4] there are 6 PRB for each user ,but in our case there are only 1 PRB for each user to be compatible with each FR scheme.

B. *OLPC with different FR schemes*

Fig. 5 illustrates different schemes of FR. It is shown that by decreasing the interference level by using different FR there will be an increasing in the CINR. The obtained results may be categorized into two main sections. The first one is the edge throughput and the other one is cell throughput.

1. Impact on edge throughput

All FR schemes obtained edge throughput gain over the ordinary OLPC.

OL-HFR has become the highest obtained edge throughput, on the other hand OL-SFR is acting as the lowest edge throughput.

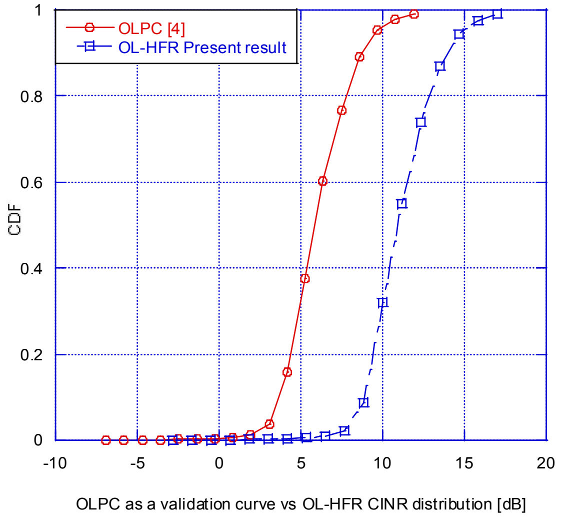


Figure 4. Shows there are CINR shift towards increasing with HF reuse scheme

Copyright (c) IARIA, 2012.   ISBN: 978-1-61208-193-9

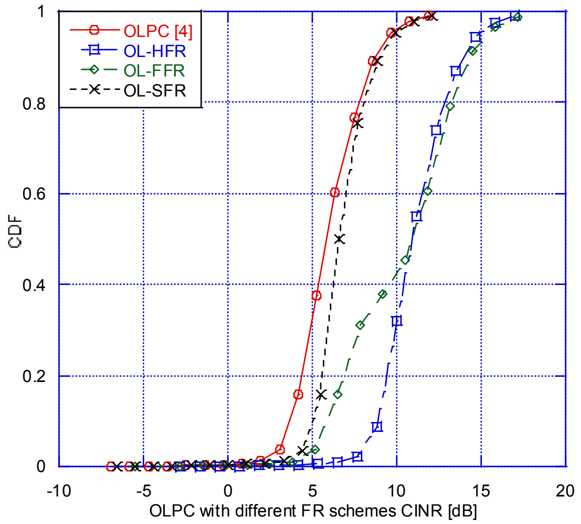


Figure 5. CINR distribution of OLPC with $\alpha = .8$, $P_0 = -81$ dBm /Hz with different FR schemes, there are an increasing in CINR for all FR

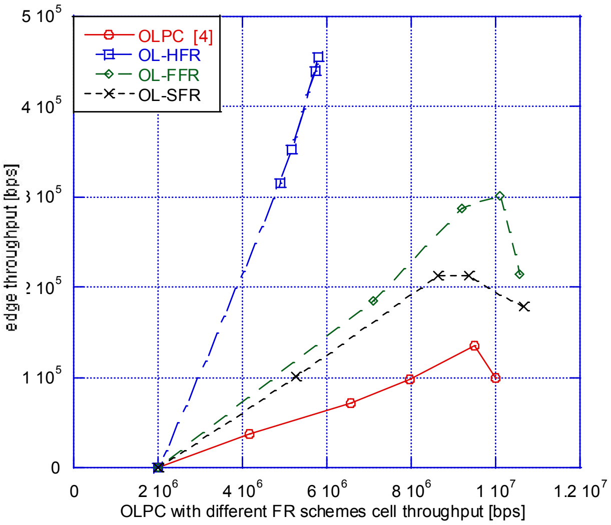


Figure 6. Edge throughput vs. Cell throughput of OLPC with $\alpha = .8$, $P_0 = -81$ dBm/Hz and with all FR schemes, there are edge throughput increasing for all FR over OLPC

The results may be explained as follows;

- OL-HFR: As a result of taking sixteen PRBs only for each cell, The interference level is decreased by 1/3 compared with ordinary OLPC; see Fig. 6.
- OL-FFR: Has a moderate edge throughput due to degradation of interference level by 1/3; see Fig. 6.
- OL-SFR: Has the lowest edge throughput due to the increasing of the interference level when it is compared to the other FR schemes; see Fig. 6.

2. Impact on cell throughput

Both OL-FFR and OL-SFR obtained cell throughput gain over ordinary OLPC on the other hand CL-HFR has lower cell throughput than ordinary OLPC.
The result may be explained as follows;

- OL-HFR: Has the lowest cell throughput as the total number of PRBs is decreased to 16 PRBs only; see Fig. 6.
- OL-FFR: Has a good cell throughput regarding to decreasing the amount of interference which is generated from the edge UEs; see Fig. 6.
- OL-SFR has the highest cell throughput because of decreasing the total amount of interference for the cell; see Fig. 6.

C. *CLPC with different FR schemes*

Fig. 7 illustrates different schemes of FR. It is shown that by decreasing the interference level by using different FR there will be an increasing in the CINR. The obtained results may be categorized into two main sections. The first one is the edge throughput and the other one is cell throughput.

1. Impact on edge throughput

All FR schemes obtained edge throughput gain over the ordinary CLPC.

CL-HFR has become the highest obtained edge throughput, on the other hand CL-SFR is acting as the lowest edge throughput.
The results may be explained as follows;

- CL-HFR: The interference level is decreased by 1/3 compared with ordinary CLPC; see Fig. 8.
- CL-FFR: Has a moderate edge throughput due to degradation of interference level by 1/3; see Fig. 8.
- CL-SFR: Has the lowest edge throughput due to interference level is higher than the other two FR schemes; see Fig. 8.

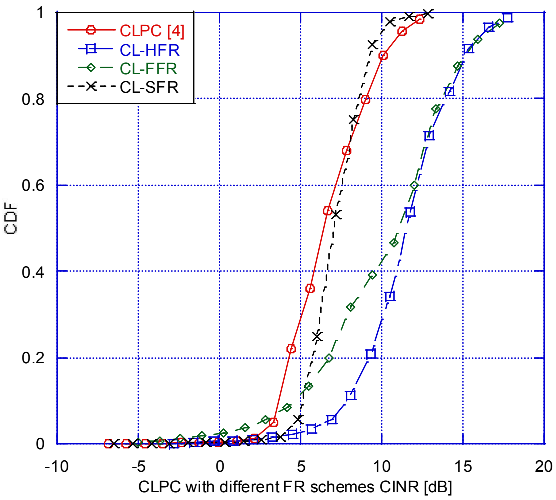


Figure 7. CINR distribution of CLPC with $\beta = .7$, $\gamma = .3$, $I_0 = -157$ dBm/Hz with different FR schemes, there are an increasing in CINR for all FR

2. Impact on cell throughput

Both CL-FFR and CL-SFR obtained cell throughput gain over ordinary CLPC on the other hand CL-HFR has lower cell throughput than ordinary CLPC.

The result may be explained as follows;

- CL-HFR: Has the lowest cell throughput as the total number of PRBs is decreased to 16 PRBs only; see Fig. 8.

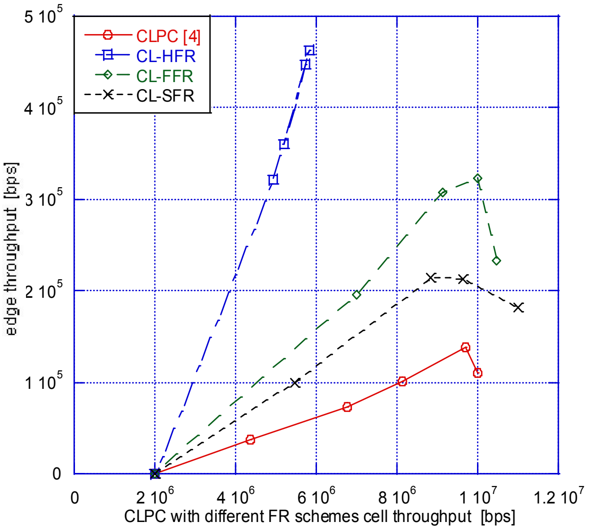


Figure 8. Edge throughput vs. Cell throughput of CLPC with $\beta = .7$, $\gamma = .3$, $I_0 = -157$ dBm/Hz and with all FR schemes, there are an increasing in all edge throughput

- CL-FFR: Has a good cell throughput regarding to decreasing the amount of interference which is generated from the edge UEs; see Fig. 8.
- CL-SFR has the highest cell throughput because of decreasing the total amount of interference for the cell; see Fig. 8.

V. CONCLUSION AND FUTURE WORK

As the FR and PC functionalities is a strong tool for co-channel interference mitigation, using them critical issues in cellular (OFDMA)/LTE networks**.**

Both of OLPC & CLPC techniques had been investigated.

The novelty of the current work is presented via considering both of FR schemes as well as the PC techniques.The obtained results shows gain in CINR for all FR schemes.

The closed loop power control has shown more cell and edge throughput and system gain.

During this work PC techniques with different FR schemes were analyzed by the means of a fixed bandwidth, balanced load and specific boundries of PL for FR schemes.

Future work could investigate the impact of variable bandwidth and unbalanced load. An important contribution would be to find a mechanism to automatically set the optimum boundries of PL for FR schemes and the ability to switch between different FR schemes to obtain the best performance

REFERENCES

[1] 3GPP TR 25.814, Physical layer aspects for evolved universal terrestrial radio access (UTRA) (release 7), Tech. report, v7.1.0, 2006.

[2] R1-073224, Way forward on power control of PUSCH, 3GPP TSG-RAN WG1 49-bis, 2007.

[3] 3GPP TS 36.213 V8.2.0, E-UTRA Physical layer procedures, 2008.

[4] Nestor J. Quintero, Advanced Power Control for UTRAN LTE Uplink, October 2008.

[5] Yong Soo Cho, Jaekwon Kim, Won Young Yang, and Chung G. Kang, MIMO-OFDM wireless communication with matlab, 2010.

[6] R1-050507, 'Soft Frequency Reuse Scheme for UTRAN LTE', Huawei.3GPP TSG RAN WG1 Meeting #41, Athens, Greece, May 2005.

[7] Carlos Ubeda Castellanos, Dimas Lopez Villa, C Rosa, Klaus I Pedersen, F D Calabrese, Per-Henrik Michaelsen, Jurgen Michel, Performance of Uplink Fractional Power Control in UTRAN LTE, VTC Spring 2008 IEEE Vehicular Technology Conference (2008) Publisher: Ieee, Pages: 2517-2521.

[8] Ashley Mills, David Lister, and Marina De Vos, Understanding Static Inter-Cell Interference Coordination Mechanisms in LTE, Journal of communications, vol. 6, no. 4, July 2011.

[9] Mathias Bohge, James Grossy, and Adam Wolisz, Optimal Power Masking in Soft Frequency Reuse based OFDMA Networks, in Proc. of the European wireless conference 2009 (EW '09), Aalborg, Denmark, May 2009.

MediaSense – an Internet of Things Platform for Scalable and Decentralized Context Sharing and Control

Theo Kanter
Department of Computer and System Sciences
Stockholm University
SE-164 40 Kista, Sweden
kanter@dsv.su.se

Stefan Forsström, Victor Kardeby, Jamie Walters,
Ulf Jennehag, and Patrik Österberg
Department of Information Technology and Media
Mid Sweden University
SE-851 70 Sundsvall, Sweden
{stefan.forsstrom, victor.kardeby, jamie.walters,
ulf.jennehag, patrik.osterberg}@miun.se

***Abstract*—Research in Internet-of-Things infrastructures has so far mainly been focused on connecting sensors and actuators to the Internet, while associating these devices to applications via web services. This has contributed to making the technology accessible in areas such as smart-grid, transport, health, etc. These early successes have hidden the lack of support for sensor-based applications to share information and limitations in support for applications to access sensors and actuators globally. We address these limitations in a novel open-source platform, MediaSense. MediaSense offers scalable, seamless, real-time access to global sensors and actuators via heterogeneous network infrastructure. This paper presents a set of requirements for Internet-of-Things applications support, an overview of our architecture, and application prototypes created in order to verify the approach in a test bed with users connected from heterogeneous networks.**

Keywords: Internet of Things, Context awareness, Sensors, Actuators, Open source

I. Introduction

Applications that utilize information from sensors to provide more personalized, automatized, or even intelligent behavior to the user are commonly referred to as Internet-of-Things (IoT) applications[1] or Machine-to-Machine (M2M) applications[2]. The reasoning is that these kinds of applications will become widespread when connected to form the IoT where everyday objects can display intelligent behavior. IoT applications can display context-aware behavior, since they may associate a user or an object with information about the surroundings and the current situation[3].

IoT applications exist in a variety of areas, such as environmental monitoring (pollution, earth quake, flooding, forest fire), energy conservation (optimization), security (traffic, fire, surveillance), safety (health care, elderly care), and enhancement of social experience and comfort. IoT applications are projected to have a big impact on how we interact with the world, people and things in the future. In order to enable IoT applications to make intelligent decisions, it is paramount to support timely access to a wide range of information sources on a global scale.

This paper therefore specifies a set of requirements that need to be considered when designing a platform for IoT applications. Related work is presented and described in relation to the requirements, along with limitations in existing solutions, foremost in the lack of support for applications to share information and provide timely access to information from global sensors and actuators. We present an overview of the MediaSense platform and how it addresses the requirements in order to offer scalable, seamless, and real-time access to global sensors and actuators via heterogeneous network infrastructures.

Section II outlines the requirements for IoT applications. Section III puts these requirements in relation to related work. Section IV outlines our solution called the MediaSense platform, whereas section V discusses our current results. Finally, section VI presents the conclusions and the research that still remains to be undertaken.

II. Requirements

Internet-of-Things applications put certain requirements on the supporting architecture and infrastructure. In our analysis, we derived a list of requirements that must be satisfied in order to provide adequate Quality of Service (QoS) and Quality of Experience (QoE) for various types of Internet-of-Things applications. In detail, these requirements are: *a*) **Scalable** – logarithmic or better scaling of communication load in end-points. *b*) **No central point of failure** – fully distributed and several ways to connect to the platform. *c*) **Bidirectional** – capable of communicating with both sensors and actuators. *d*) **Fast** – capable of signaling in real-time between end points. *e*) **Current** – all data retrieved should be the most current values. *f*) **Lightweight** – able to run on mobile devices with limited resources. *g*) **Seamless** – capable of handling multi-NAT traversal, heterogeneous infrastructures, and different end user devices. *h*) **Stable** – reliably handle transient nodes joining and leaving with high churn rates, while making sure that all queries into the platform should return an answer. *i*) **Extensible** – capable of adding new features and modules without complete redistribution, such as persistence, authentication, and reasoning.

III. Related Work

Related work has mainly been focused on brokering of sensor information on the Internet, via different types of web services. Examples of these typical IoT architectures which utilize centralized servers or cloud-based web services are SenseWeb[4] and Pachube[5]. In detail, these approaches broker the information through a centralized web-service based architecture and thus they do not support requirement a), b), c), d), and e), i.e., **Scalable**, **No central point of failure**, **Bidirectional**, **Fast**, and **Current**.

Cloud-based infrastructures such as CeNSE[6] claim to address the scalability issues. However, cloud-based infrastructures centralize components for authentication and brokering, therefore not satisfying requirements b) **No central point of failure** and e) **Current**.

Project SENSEI[7] proposes a logical architecture for the IoT, but has as yet not provided answers about how sensors are integrated and how such information will be made available in a real-time and scalable fashion, therefore not fulfilling requirements a) **Scalable** and d) **Fast**. Additionally, its architecture contains components which centralize brokering and therefore does not satisfy requirement b) **No central point of failure**.

The SOFIA architecture[8] offers a scalable middleware approach to context aware applications. SOFIA is based on an ontological data model, which can provide filtering of information in relation to related context to create context awareness. The reasoning over an ontological model is inherently slow and the solution there does not satisfy requirement d) **Fast** or e) **Current**.

The COSMOS system[9] is also a middleware for context-centric access control for wireless architectures. COSMOS applies context information to create a novel security model for context-centric access control where mobile agents acts as proxies for mobile devices inside the middleware. This introduces an additional step in data communication which does not satisfy requirement d) **Fast**.

IV. The MediaSense Platform

In response to the shortcomings of earlier solutions in regards to fulfilling the requirements presented in section II, this paper presents a novel architecture for developing applications on the Internet of Things, which satisfies the requirements. The architecture is encapsulated as the MediaSense platform and it is a distributed architecture that enables IoT applications based on sensor and actuator information. An overview of the platform and its components is presented in Figure 1, which show how the platform is distributed over a number of entities connected to the Internet. The figure show how an application that is running a client of the MediaSense platform (a MediaSense instance) communicates with other entities running the platform. A client can acquire sensor and actuator information of the other participants. Furthermore, the platform can act as both a producer and consumer of sensor and actuator information at the same time, enabling bidirectional exchange of context information.

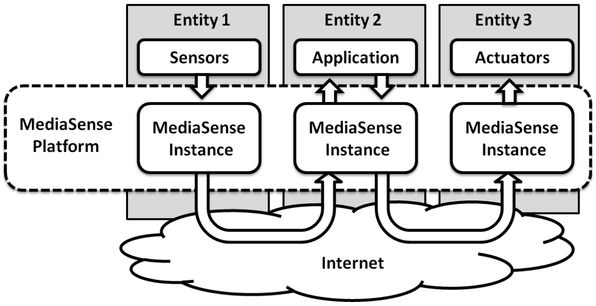


Figure 1. Overview on the function of the MediaSense platform.

A more detailed overview of the whole architecture, including all the layers and components, is shown in Figure 2. This paper will focus on this figure and the remainder of this section will explain the purpose and operation of each layer and their components.

A. Interface Layer

The interface layer is the public interface through which applications interact with the MediaSense platform. The interface layer includes a single component, the MediaSense application interface, which is a generic and standardized Application Programming Interface (API) for developers to build their own IoT applications.

1) MediaSense Application Interface: The purpose of the MediaSense application interface is to provide a single entry point for developers to create applications on top of the MediaSense platform. The interface is thus a standardized API and it provides access to all of the available functionality that the platform provides. Hence, the MediaSense application interface provides many different means of interacting with the MediaSense platform, such as accessing the dissemination core directly or trough any running add-ins, all depending on the applications demands, requirements, and sought after QoE.

B. Sensor and Actuator Layer

The purpose of the sensor and actuator layer is to enable a generalized method to produce information and provide it to the MediaSense platform. The problem is that there exist a large number of different sensors and actuators, which use many different technologies. This needs to be addressed in order to provide the platform with the information and functionality that applications require. The sensor and actuator layer is therefore separated into four components: the actual sensors and actuators, different sensor and actuator

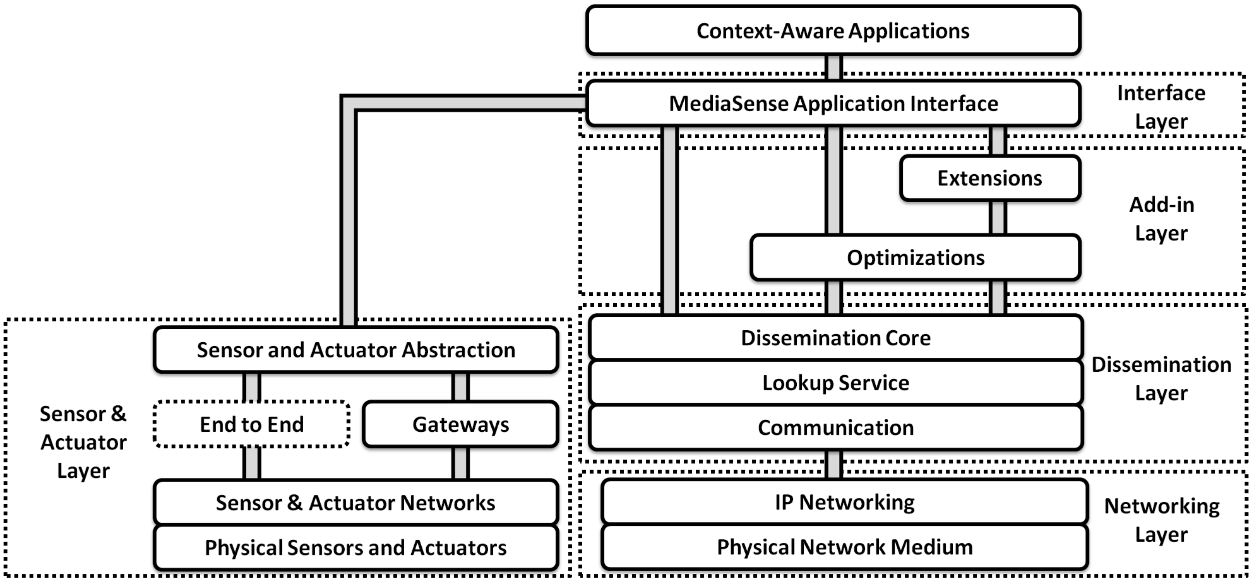


Figure 2. Overview of the MediaSense platform's architecture.

networks, a sensor and actuator gateway, and an abstraction component.

1) Physical Sensors and Actuators: The sensors and actuators provides the actual connection to the physical world. Sensors sense their surroundings and are thus the sources of context information to the whole system. By sensors we mean anything that can produce contextual information, for example, GPS location, temperature, pressure, humidity, and health status. But also context information that are difficult to physically sense, such as name, favorite food, mood, preferences, etc. By actuators we mean any type of object which can access the physical world and perform some form of actuation in it. Typical actuators include controlling devices such as light switches and heating temperature settings. But actuators can also include any type of shared resource made available to interact with via the MediaSense platform, such as shared data storage.

2) Sensor and Actuator Networks: Sensors and actuators are usually connected through some form of sensor or actuator network. These are for example, wireless sensor networks (WSN) or wireless sensor/actuator networks (WSAN). The purpose of these networks is to gather data from many connected physical sensors through a network of sensors and actuators. WSN/WSAN is used in order to achieve larger area coverage, higher quality of service, lower energy consumption, and cheaper hardware.

A sensor and actuator network commonly designate a node to coordinate access to and from the outside world, it is not uncommon for such a node to be more advanced in terms of processing power or battery reserves. Based on the properties of the network, this will be the node that communicates upward to either a gateway or directly to the sensor and actuator abstraction.

3) Gateways: As a consequence of the wide range of protocols and technologies used in sensors and actuator networks, gateways are sometimes required. In detail, a gateway translates the sensor specific network technology into a common communication protocol. Thus a gateway mediates communication with each specific sensor and actuator type. Therefore a separate gateway has to be built for each new sensor or actuator network that wants to connect into the MediaSense platform. The gateway then provides access to the sensors and actuators regardless of the underlying technology used by the sensor and actuator networks.

However, if the sensor and actuator network has the ability to directly talk to the abstraction component, the gateway can be ignored for that particular network. This is denoted as the "End to End" part inside the sensor and actuator layer. These are for example IPv6 capable sensor and actuator networks [10] or other types of networks with extend capacities and computational power.

4) Sensor and Actuator Abstraction: The abstraction component provides a standardized method of interaction with all sensors and actuators. It abstracts all sensors and actuators into a generalized and standardized format, which is connected into the MediaSense platform through the application interface as any other application. It thus provides the platform with access to all types of sensors and actuators in one of two ways. Either through a gateway that translates any communication protocol and access method used by the sensor and actuators network, or by directly communicating with the sensor and actuator network if the sensor and actuator network is powerful enough.

C. Add-in Layer

The purpose of the add-in layer is to enable developers to add optional functionality and/or optimization algorithms to the MediaSense platform. An add-in can be used in order to make the MediaSense platform meet specific application requirements, sought after quality of experience, or available capacity in regards to computational power and bandwidth. Thus the add-in layer manages different extensible and pluggable add-ins, which can be loaded and unloaded in runtime when needed. The add-in layer can include any number or type of add-ins, but they are divided into two categories, optimization components and extension components. Whereas the optimization components offer ways of optimizing the behavior and functionality of the system, and the extension components enables extended functionality which applications might demand.

1) Extensions: The extension components provide add-ins for enabling extended functionality, such as context-awareness and reasoning, in the MediaSense platform. These extension add-ins can for example include, logical context objects, semantics, reasoning, ranking of context information, search engines, query languages, and context agents.

2) Optimizations: The optimization components provides add-ins for optimizing the MediaSense platform in many different forms. These optimizations add-ins can for example include clustering of information, caching, persistence, intelligent routing, and decision making to determine when to optimally send data.

D. Dissemination Layer

The dissemination layer enables dissemination of information between all entities that participate in the system and are connected to the IoT. In detail, the Distributed Context eXchange Protocol (DCXP) [11] is used. DCXP offers reliable communication among entities that have joined a peer-to-peer network, which is used to enable exchange of context information in real-time. The operation of the DCXP includes resolving of so called Universal Context Identifiers (UCI) and subsequently transferring context information directly with a resolved entity. Therefore, the dissemination layer includes three components, a dissemination core, a lookup service, and a communication system. The dissemination core exposes the primitive functions provided by DCXP, the lookup service find and resolve other entities who has joined the system, and the communication component abstracts a transport layer communication.

1) Dissemination Core: DCXP offers primitive functions for publishing, retrieving, and transferring information in a peer-to-peer manner, as well as joining and leaving the peer-to-peer network. Hence, it is the dissemination core that exposes these primitive functions to the above layers, thus making these services available to the MediaSense platform. Furthermore the dissemination core hides the underlying lookup service and communication technology from the above layers. Thus allowing different choices for lookup service and communication technology without any changes to the other layers.

2) Lookup Service: The lookup service provides the means of resolving UCI's to find the location of a sought after piece of information or entity. The lookup service can be implemented in a number of different ways, for example as a distributed hash table, distributed graph, or cloud server. Examples of already tested and evaluated systems are Chord and PGRID, which was done in [12] and [13], respectively.

3) Communication: The communication component offers the possibility to exchange the communication protocol of the dissemination layer and thus of the whole MediaSense platform. It also makes it possible to provide multiple concurrent communication protocols, such that the components can request different quality of service based on their chosen communication protocol. Examples of possible communication protocols are: TCP, UDP, Reliable-UDP, and Stream Control Transmission Protocol (SCTP).

E. Networking Layer

The MediaSense platform is designed to operate over heterogeneous infrastructure, including wireless and mobile. The purpose of the networking layer is to connect different entities over current IP based networking infrastructure. In general the networking layer has two components, an IP network and a physical network medium.

1) IP Networking: The IP network component is the IP endpoint for a particular entity, which is running an instance of the MediaSense platform. The IP networking components thus provides the ability to communicate with other entities on the Internet, regardless of type the type of connection. In detail, this can include both IPv4 and IPv6 networks.

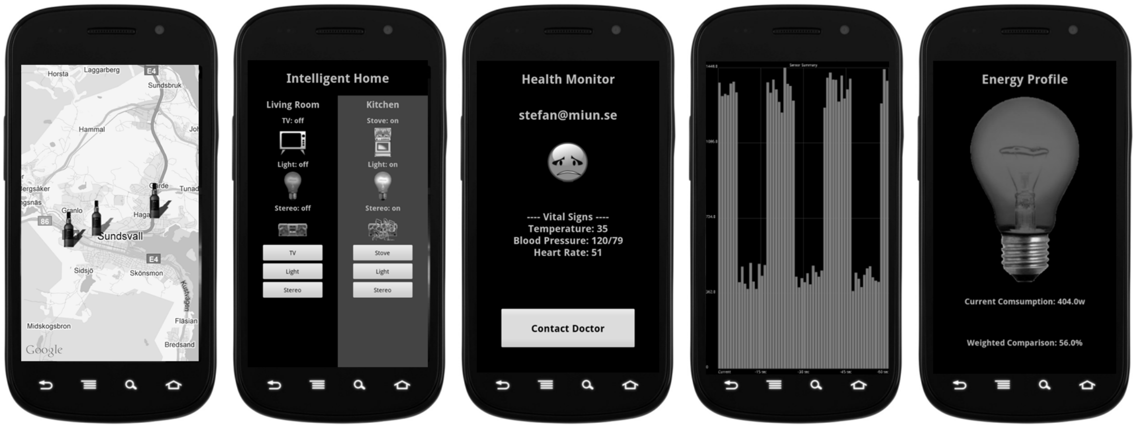
2) Physical Network Medium: The physical network medium component denotes that the MediaSense platform is agnostic of the underlying infrastructure. Hence, the MediaSense platform can run over heterogeneous networks and via different types of physical infrastructures. This includes different technologies such as Ethernet, 802.11 b/g/n, and other variants of mobile broadband and fiber optic networks.

V. Results and Discussion

The current results include development and launching of an open source development website (www.mediasense.se) for the MediaSense platform. This website will act as a portal for all developers who want to utilize the MediaSense platform in their applications. The MediaSense platform is provided free and under an open source license, in order to make it available for anyone to use.

A. Verification

Initial testing and evaluation of the open source platform has been conducted with users in a testbed with fixed and mobile access to the Internet. Proof-of-concept applications



stefan@miun.se

---- Vital Signs ----
Temperature: 35
Blood Pressure: 120/79
Heart Rate: 51

Current Comsumption: 404.0w

Weighted Comparison: 56.0%

(a) Object tracking (b) Intelligent home (c) Health monitoring (d) Energy consumption (e) Energy profiling

Figure 3. Examples of applications using the MediaSense platform.

have been built both on set-top boxes and smartphones, in order to show that the MediaSense platform can be applied in a wide range of scenarios, e.g., health care, intelligent home, object tracking, and social applications. Figure 3, shows some of these proof-of-concept applications that use the MediaSense platform. From left to right they represent object tracking (for tracking sensor enabled objects), intelligent home automation (for interacting with the intelligent home), health monitoring (for medical status and alerts), energy consumption (for smart energy monitoring), and energy profiling (for energy awareness). The initial testing and evaluation indicates that the MediaSense platform is on par with UDP traffic over mobile Internet access, which is demanded by real-time applications.

The application shown in Figure 3a addresses object tracking and verifies the importance of requirements a) scalable, b) **No central point of failure**, f) **Lightweight**, and h) **Stable**, due to the large number of transient entities and objects.

The application in Figure 3b targets the intelligent home and verifies the importance of requirements c) **Bidirectional** and g) **Seamless**, due to that it both handles sensors and actuators, as well as different networks.

Figure 3c shows a health monitoring application that verifies the importance of requirements d) **Fast**, e) **Current**, h) **Stable**, and i) **Extensible**. This is due to the highly sensitive data which requires secure authentication and real-time delivery to minimize delay of critical health care.

The application in Figure 3d targets energy consumption and verifies the importance of requirements d) **Fast**, and e) **Current**, requiring a steady stream of current data to monitor the changes in energy consumption.

Lastly, the energy profiling application in Figure 3e verifies the importance of requirements c) **Bidirectional**, d) **Fast**, e) **Current**, and i) **Extensible**, because profiling also has to include reasoning in order to create energy awareness in the application.

B. Addressing the Requirements

The presented architecture and platform can support all of the posed requirements in section II. In detail, the requirements are addressed as follow.

a) **Scalable** is addressed by using a scalable lookup service in the platform. For example, the Chord and P-grid solutions which scale logarithmic with the amount of entities in the whole system.

b) **No central point of failure** is addressed by using a fully distributed system in the dissemination layer, without any centralized component. A fully distributed system is needed due other requirements, such as fast, current, and scalability.

c) **Bidirectional** is addressed by allowing two-way communication in the dissemination layer and accepting both sensors and actuators in the system.

d) **Fast** is addressed by using a distributed lookup service with logarithmic, or better, lookup delays that utilize a communication protocol which can provide real-time communication. This, in order to support real-time applications, and that no obsolete information is being considered in application logic.

e) **Current** is addressed by using a peer-to-peer system as the dissemination layer. This because a peer-to-peer system is communicating with the source, and thus is always proving the most current value.

f) **Lightweight** is addressed with the possibility of choosing lightweight components and only loading the required add-ins. For example only loading reasoning, semantics, etc., when needed. Thus avoiding computational heavy components if they are not required. Our platform is therefore able to run on mobile devices such as smartphones.

ISBN: 978-1-61208-193-9

g) **Seamless** is addressed by having the platform agnostic of the underlying infrastructures. The networking layer supports heterogeneous IP-based infrastructures if the lookup service and commutation protocol can penetrate Network Address Translation (NAT).

h) **Stable** is addressed by utilizing a stable lookup service. The lookup service must thus support transient nodes leaving and joining, as well as intermittent disruptions and unexpected disconnections. Furthermore, the reliability is solved by using a reliable lookup service and commutation protocol. This is also required in order to support real-time applications, with application logic depending on fresh information. Thus the lookup service and the communication protocol must provide reliable services.

i) **Extensible** is addressed by allowing multiple add-ins to be dynamically loaded on demand. Which is solved in the add-in layer, and is a prerequisite for other requirements, such as persistence and authentication. Persistence can be addressed by adding such an add-in to the platform. Relevant data can be stored based on importance, in order to provide better lookup, and reliability. Furthermore, controlling the reach of information can be addressed by creating an add-in that can authenticate entities and encrypt the data.

VI. Conclusion and Future Work

The contributions of this paper begins in Section II with the specification of a set of requirements that need to be considered when designing a platform for IoT applications. Then existing related work is evaluated in Section III, where several of these approaches address some of the requirements, but they all have their limitations. We therefore propose the MediaSense platform and its components in Section IV. It provides a distributed IoT infrastructure that offers scalable, seamless, real-time access to global sensors and actuators. The sensors and actuators are connected via devices that act as end-points in the peer-to-peer based infrastructure. Applications running locally on such end-point devices can thereby access information from sensors, and control actuators, connected to any other end-point. The proposed MediaSense platform fulfills all the stated requirements, as shown by the demo applications and descriptions in Section V. Further, the platform is provided for free, under an open source license.

Current efforts are directed toward interfacing between the connected-things infrastructure and the world of our experience, through extending the platform with a semantic layer and datamining capabilities, decision making, reasoning, and optimizations. Moreover, we are working on a generic integration with IoT and cloud infrastructures in new projects.

Acknowledgment

This research is partially funded by VINNOVA, the Swedish Governmental Agency for Innovation Systems. The authors also want to thank partners from industry and academia, in particular Acreo AB, Ericsson, and Stockholm University.

References

[1] L. Atzori, A. Iera, and G. Morabito, "The internet of things: A survey," *Computer Networks*, vol. 54, no. 15, pp. 2787–2805, 2010.

[2] G. Wu, S. Talwar, K. Johnsson, N. Himayat, and K. Johnson, "M2m: From mobile to embedded internet," *Communications Magazine, IEEE*, vol. 49, no. 4, pp. 36–43, 2011.

[3] J. Hong, E. Suh, and S. Kim, "Context-aware systems: A literature review and classification," *Expert Systems with Applications*, vol. 36, no. 4, pp. 8509–8522, 2009.

[4] W. Grosky, A. Kansal, S. Nath, J. Liu, and F. Zhao, "Senseweb: An infrastructure for shared sensing," *Multimedia*, vol. vol. 14, pp. pp. 8–13, 2007.

[5] O. Haque. Pachube. [Online]. Available: www.pachube.com

[6] Central Nervous System for the Earth (CeNSE). Hewlett Packard. [Online]. Available: http://www.hpl.hp.com/research/intelligent_infrastructure/

[7] A. Gluhak, M. Bauer, F. Montagut, V. Stirbu, M. Johansson, J. Vercher, and M. Presser, "Towards an architecture for a real world internet," *Towards the Future Internet: a european research perspective*, pp. 313–324, 2009.

[8] A. Toninelli, S. Pantsar-Syväniemi, P. Bellavista, and E. Ovaska, "Supporting Context Awareness in Smart Environments: A Scalable Approach to Information Interoperability," in *Proceedings of the International Workshop on Middleware for Pervasive Mobile and Embedded Computing*, 2009.

[9] P. Bellavista, R. Montanari, and D. Tibaldi, "Cosmos: A Context-Centric Access Control Middleware for Mobile Environments," in *Mobile Agents for Telecommunication Applications*, 2003, pp. 77–88.

[10] A. Dunkels, B. Gronvall, and T. Voigt, "Contiki-a lightweight and flexible operating system for tiny networked sensors," in *Local Computer Networks, 2004. 29th Annual IEEE International Conference on*. IEEE, 2004, pp. 455–462.

[11] T. Kanter, S. Pettersson, S. Forsstrom, V. Kardeby, R. Norling, J. Walters, and P. Osterberg, "Distributed context support for ubiquitous mobile awareness services," in *Communications and Networking in China, 2009. ChinaCOM 2009. Fourth International Conference on*, Aug. 2009, pp. 1–5.

[12] T. Kanter, P. Österberg, J. Walters, V. Kardeby, S. Forsström, and S. Pettersson, "The MediaSense Framework," in *Proceedings of Fourth IARIA International Conference on Digital Telecommunications (ICDT)*, Colmar, France, July 2009.

[13] J. Walters, T. Kanter, and E. Savioli, "A Distributed Framework for Organizing an Internet of Things," in *the 3rd International ICST Conference on Mobile Lightweight Wireless Systems*, 2011.

Mobile Robot Localization by RFID Method

Ya-Chuan Chen
Department of Engineering Science
National Cheng Kung University
Tainan, Taiwan
yachuan.tw@gmail.com

Jung-Hua Chou
Department of Engineering Science
National Cheng Kung University
Tainan, Taiwan
jungchou@mail.ncku.edu.tw

***Abstract*—This paper explores the merits of the indoor positioning system using Radio-frequency identification (RFID) technology for mobile robots. The indoor positioning function by the Received Signal Strength Indicator (RSSI) and the Link Quality Indicator (LQI) method is developed first to determine the location of the robot. The RSSI and the LQI based indoor positioning system employs wireless signal intensities to build the intensity distribution map and compares the signal intensity of the transmitters obtained by the receivers to determine the location of the robot. In the end, the implementation is to use a predetermined indoor environment, which is set up beforehand to validate the RFID based indoor positioning system developed by a mobile robot. Validation results show that the precision of the positioning system in one-dimensional is 10 cm and the average accuracy rate is 97.23%. The precision of the positioning system in two-dimensional is 10cm x 10cm and the average accuracy rate is 79.73%. Finally, the average error of the robot tracking in two-dimensional is 14.16cm.**

Keywords-RFID technology; RSSI; LQI; indoor positioning; mobile robot

I. INTRODUCTION

Currently, various kinds of robots are used to make work efficiently and avoid the dangerous situations for human beings. In order to make the robot perform the designated activities like human beings in an unknown environment, the robot must be able to perceive its own locations. By being able to perceive its position in the environment, the robot can then move autonomously to achieve its tasks.

For the purpose of knowing the location of the robot, different methods have been developed, including multiple sensors for obstacle detection in the environment such as infrared sensor [1]-[5], ultrasonic sensor [6]-[7], image methods [8]-[14], lasers [15]-[17], color belt method [13], landmark [18]-[22] and RFID [20]-[31], etc. The RFID tag includes a small RF transmitter and receiver. An RFID reader transmits an encoded radio signal to interrogate the tag. The tag receives the message and responds with its identification information. This study adopts the RFID method for robot localization for its simplicity and wireless features.

In this paper, Section I introduces the structure of the paper. Section II introduces the information about the RFID system. Section III and Section IV introduce the basic principles of positioning and the positioning techniques. Section V introduces the system design and the implementation. In the end, Section VI shows the results and the discussion and Section VII is to sum up the paper.

II. RFID SYSTEM

RFID is a technology that transfers data between a reader and an electronic tag attached to an object for the purpose of identification and tracking via wireless signals and has attracted great attentions recently for its tracking capability. It was originally intended in industry as an alternative to the bar code and logistic applications. Its advantage includes no requirement of direct contact or line-of-sight scanning.

The purpose of an RFID system is to enable data to be transmitted to and from the tag. The data may provide identification or location information, or specifics about the product being tagged, such as price, color, date of purchasing, etc. A basic RFID system consists of three components, namely an antenna, a transceiver (with a decoder often combined into the reader) and a transporter (an RF tag, electrically programmed with unique information). The antenna transmits radio signals to activate the transponder.

Three types of RFID tags are commonly used, including active, passive and semi-passive ones. Active RFID tags broadcast their signal to the reader, and are typically more reliable and accurate than passive ones. They send stronger signals and thus are more adaptable to environments that are hard to transmit other types of tags, such as under water situations or those requiring a longer distance of communications. However, their required power source makes them larger and more expensive. Passive RFID tags, on the other hand, do not need internal power supplies and rely on the RFID reader to transmit data so that they are smaller and cheaper. A small electrical current is received through radio waves by the RFID antenna which has enough power to transmit a response. They are more suited for warehousing environments where the interference is not severe and the distances are relatively short. Semi-passive RFID tags are similar to the active ones with an internal power supply, but it does not broadcast any signal until the reader initiates it.

III. THE BASIC PRINCIPLES OF POSITIONING

A. RSSI (Received Signal Strength Indicator)

RSSI [32]-[33] uses a channel path loss propagation model to describe the signal decay with distance. RSSI represents the signal strength. When the transmitter is near to the receiver, the RSSI value is larger and vice versa. For localization, three receivers are required at least. In the free space propagation model, the transmitted and received signal power ratio is as follows:

$$P_r = \frac{P_t G_t G_r \lambda^2}{(4\pi d)^2}$$

ISBN: 978-1-61208-193-9

where P_r (dB), P_t (dB), G_r, G_t, λ (m) and d (m) denote the received signal power, the transmitted signal power, the receiver antenna gain, the transmitter antenna gain, the signal wavelength and the distance between the transmitter and the receiver, respectively

In an indoor environment, the received and transmitted power ratio is susceptible to both multi-path interference and shadowing effects. Therefore, the signal strength will be different from that in free space. In addition, the resulted distance will have errors due to refractions at edges and from the media, different propagation speeds, polarization and scatterings, etc. That is, the deduced location will not be at one point but will fall into an area. Therefore, the total received power from the tag has to take into account the effects of multi-reflections and other error factors which may cause the power loss. Thus, the total received power previously given is modified as follows:

$$P_r \approx P_{ref} + P_{err} + \frac{P_t G_t G_r \lambda^2}{(4\pi d)^2}$$

where P_{ref} (dB) and P_{err} (dB) represent the power due to reflections and other error factors, respectively.

B. *LQI (Link Quality Indicator)*

In practical applications, the total received power can not be used directly due to affecting factors mentioned above. Hence, in this study we use the link quality indicator (LQI) method for our distance reduction to avoid the problems.

LQI [34] is determined by two characteristic of the receiving packets, namely, signal strength and quality. It is specified by IEEE802.15.4 to assess the quality of the communication link between a receiver and transmitter. LQI is based on signal to-noise ratio or energy density of the signal in the frequency band used and provides average correlation values for each incoming packet. As with RSSI, LQI allows users to assess the communication link considering the environmental effects on a single transmitter/receiver pair. However, LQI provides a more thorough estimate of the quality of the link than RSSI since it assesses all possible frequencies in the physical layer of the transmission.

The basic idea of using LQI for distance determination is that the transmission quality is better when the distance is smaller and vice versa. In general, the values of LQI are expressed in binary digits of a byte from 0 to 255 with 255 indicating the highest link quality. That is, the received signal is divided into 256 parts equally from the weakest to strongest by the manufacturer.

IV. POSITIONING TECHNIQUES

The positioning technique adopted in this study includes pattern matching and related algorithms as follows.

A. *Pattern matching location technology*

The pattern matching between the received signal and the database of the training samples for positioning is known as the fingerprint matching method. The matching process is divided into two phases as shown in Figure 1; namely off line (off stage) and on line (real time) as follows.

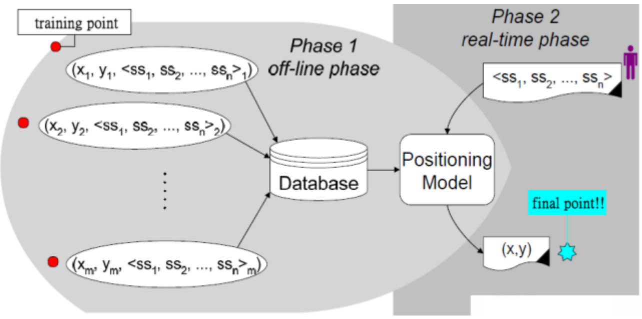

training point
Phase 1
off-line phase
Phase 2
real-time phase
$(x_1, y_1, <ss_1, ss_2, ..., ss_n>_1)$
$(x_2, y_2, <ss_1, ss_2, ..., ss_n>_2)$
$(x_m, y_m, <ss_1, ss_2, ..., ss_n>_m)$
Database
Positioning Model
$<ss_1, ss_2, ..., ss_n>$
final point!!
(x,y)

Figure 1: Pattern matching for location

· Off line (Off stage): establishing the feature database.

This is the training phase in which the signal strengths at a number of locations are measured and recorded to the database first. The format of the received signal strength of each record is denoted by $(x,\ y,\ \langle ss_1,\ ss_2,\ \dots,\ ss_n\rangle\ )$, where $(x\ ,\ y)$ is the coordinate of the training location and $ss_i(i = 1, 2, \dots, n)$ is the received signal strength at the training point from the i^{th} transmitter. Once the database is established, positioning can be accomplished by pattern matching between the received signal strength and that in the database.

· On line (Real time)

After establishing the database, real time position localization is performed. In this situation, a number of training locations is selected at which the received signal strength from every sender is recorded and the received signal pattern is then compared with the training pattern in the database for positioning. In general, positioning model will give several possible estimated locations combined with their probabilities. In the present method, only the highest likelihood location is selected.

B. *Probability-based method*

Probability method is the use of conditional probability statistics. Under current conditions, use the current location of the user to find the most likely location of the user. Firstly, establish the characteristic vector for every training location to form the database of the characteristic vectors. The characteristic vector is denoted by $C_i = \langle C_1^i,\ C_2^i,\ \dots,\ C_n^i\rangle$, where the C_j^i is the i^{th} training location receiving the distributed signal strength from the j^{th} transmitter. Using statistical methods to collect all possible distributed signal patterns from every transmitter of the training location, and then to set the threshold value of the number of the correct matching data. When the number of the correct matching data is greater than the threshold value, the training location is used to estimate the position.

V. SYSTEM DESIGN AND IMPLEMENTATION

The procedure of location planning is set in two steps; first, one-dimensional line planning; then two-dimensional positioning planning. By varying the distance between the robot and the signal transmitter, the law of the received signal by LQI is established to determine the positioning relationship.

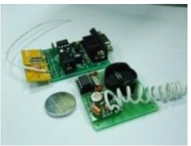


Figure 2: RFID module

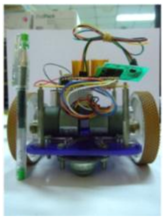


Figure 3: the mobile robot

In Figure 2 is RFID module, which is operated at the frequency of 315M Hz. The effective radius of RFID module is 15 meters with an antenna. In Figure 3, the mobile robot uses a micro-processor AT89S52 to control the motors for movements and uses the Bluetooth for remote control via a personal computer. The robot size is145mm long by 115mm wide by 120 mm high and weighs 1kgw.

Table I. The specifications of RF8315RT

Specifications	Description
Receiver Supply Voltage :	9VDC via wall adaptor if necessary
Receiver Supply Current :	4mA Typical
Receiver Output :	RS232, 9600 Baud, 8 bit words,1 stop bit,1 start bit, no parity
Transmitter Power Supply :	CR2032 (7,000 hours)
Operating Frequency :	315 Mhz
Data Format :	4 characters
Effective Radius :	15 meters by using antenna

Table I is the specifications of the RFID module. As shown in Figure 4, the positioning system consists of the mobile robot, RFID modules and the personal computer. It has four active RFID tags with an RFID receiver for its position module. The RFID receiver is responsible for transmitting the received signal strength of the RFID tags via RS232 to the personal computer for data processing. The personal computer processes the received signal strength of each RFID tag and displays the positioning results.

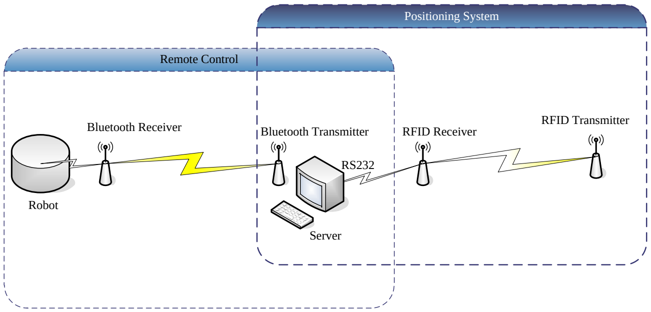
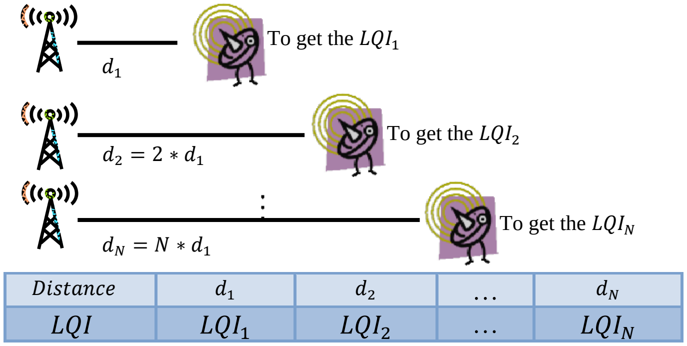


Figure 4: System configuration

VI. RESULT AND DISCUSSION

A. LQI statistics and the deduced distance database table

For real time positioning, we need to establish the reference database of LQI versus known distance first. This is accomplished by fixing the transmitter location, varying the receiver distance step by step from the transmitter and measuring the corresponding LQI value at each location. In In Figure 5, it shows that how is the LQI reference table being established. It shows that at distance d_1, the corresponding LQI_1 is measured. Then, the receiver is moved to location 2 where the distance is $d_2 = 2d_1$ to measureLQI_2, and so on. The LQI values are measured a total of 30 data at each location so that data variance can be reduced to obtain an interval of the average value and thus to strengthen the reliability of the received signal.



$Distance$	d_1	d_2	…	d_N
LQI	LQI_1	LQI_2	…	LQI_N

Figure 5: Establishing the LQI reference table with known distance.

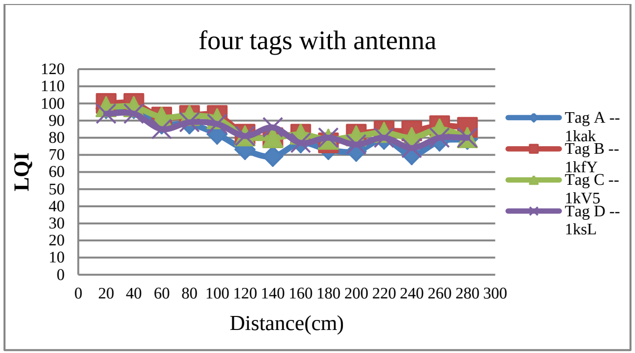


Figure 6: LQI distributions versus distance for four RFIDs, each with an antenna.

Typical signal strength variations versus distance are shown in Figure 6 for four RFID tags. It can be observed that the signal strength decreases with some oscillation as the distance increases from 0 cm to 140 cm. For distances greater than 140 cm, the LQI value levels off to about a fixed range. The signal can be received to a distance as far as 1500 cm. It is clear that the gain of each RFID is not the same because the curves shown in Figure 6 do not coincide to each other. To unveil the reason of this inconsistency among the RFIDs, the antenna of the RFID is removed and the measurements are taken again with the same conditions as that of Figure 6. The results are shown in Figure 7. It can be seen that data variation among the RFIDs is reduced considerably. That is, the variation in antennas of the RFIDs is the main reason causing the data spreading shown in Figure 6.

The results depicted in Figure 7 indicate that the signal strength decreases as the distance increases from 0 cm to 80 cm. The oscillatory variation shown in Figure 6 is greatly reduced. Furthermore, the decreasing rate of the LQI value is larger than that with antennas. For the distance greater than 80 cm, the LQI value is reasonably constant. The signal can be received as far as 8 m, shorter than that with the antenna of 15 m. That is, the antenna gain factor is about 1.75. Because a constant LQI value is not suitable for distance determination, the LQI in the range from 0 to 80 cm illustrated in Figure 7 is used for distance determination in this study due to the larger negative slope.

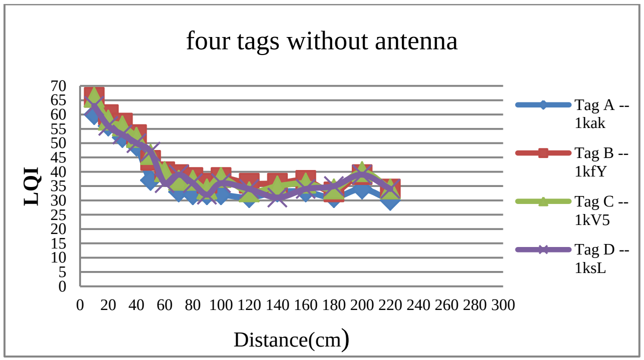


Figure 7: LQI distributions versus distance for the RFIDs without any antenna.

B. *Experimental one-dimensional positioning for mobile robot*

Table II. The table of one-dimensional position and the actual distance prediction (measuring a total of 30 data)

Location	10 cm	20 cm	30 cm	40 cm	50 cm	60 cm
the numbers of true positioning	30	30	30	29	26	30
the numbers of false positioning	0	0	0	1	4	0
accuracy	100 %	100 %	100 %	96.67 %	86.67 %	100 %

➢ The average accuracy:97.23%

For the one dimensional positioning experiments, at each distance, the measurements are conducted a total of 30 data for each location and the distance is changed with an interval of 10 cm. The results are shown in Table II. The number of false positing given in the table means the number of incorrect measurement among the 30 measurements at each location. The accuracy is deduced by dividing the number of accurate positing to the total measurement number 30. It can be observed that within the distance of 30 cm and at the distance of 60 cm, no error occurs. However, at the distances of 40 cm and 50 cm, the errors are 10 cm and ±10 cm, respectively. That is, at these two locations, an error of one measurement interval (10 cm) is observed due to the oscillatory behavior of LQI shown in Figure 7 at these distances.

C. *Pattern matching location technology to establish the signal strength database*

In the off-line stage, we establish the characteristics of the positioning database. The environment is divided into 36 cells by an area of 10 cm x 10 cm each first. Then we measure the signal strength LQI of each position to establish the location reference database. In real-time applications, we use the positioning module and positioning rules described above to estimate the receiver position. Figure 8 shows the experimental environment for real time testing. It can be seen that a tag was installed at each corner of the four corners of the environment while a robot moves inside the area enclosed by the tags.

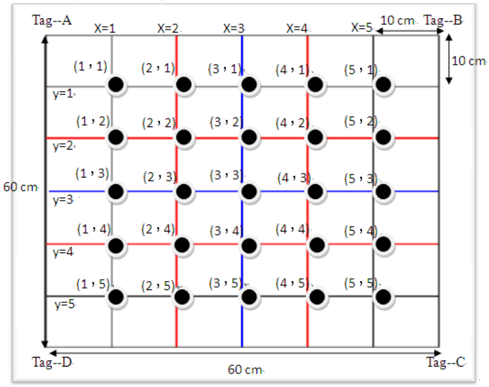


Figure 8: The testing environment.

Figures 9(a)-(d) illustrate the variation of the signal strength for each RFID tag. The result shows that the signal strength of the tag varies from the distance in the indoor environment. That is, the received signal strength is smaller when the distance is farther. This is true for all of the four tags investigated. However, it is also evident that at the same distance, the received signal strength from each tag. That is, when using the received signal strength of the tags to determine a specific location in the environment, the deduced coordinate is not a single value but during an interval.

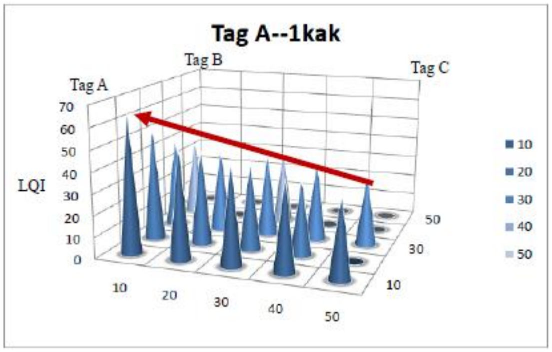


Figure 9(a) Signal strengths from Tag A

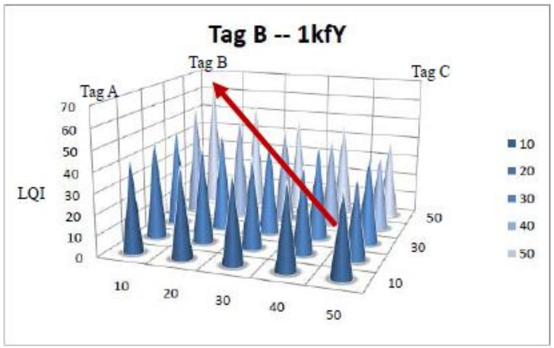


Figure 9(b) Signal strengths from Tag B

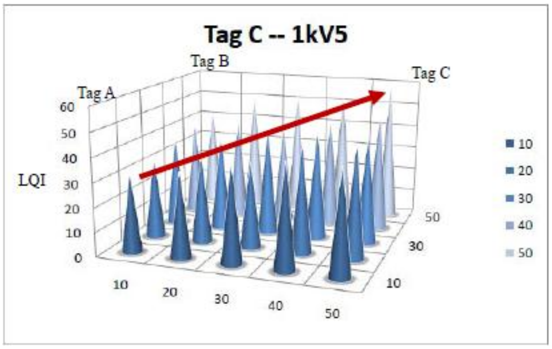


Figure 9(c) Signal strengths from Tag C

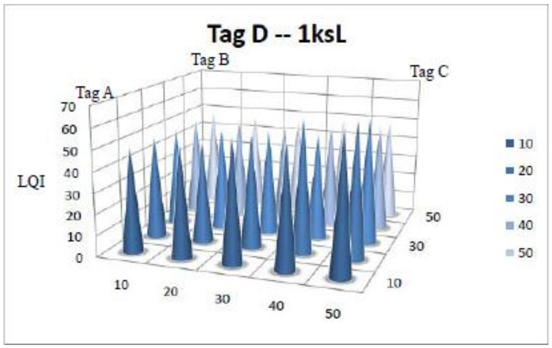


Figure 9(d) Signal strengths from Tag D

D. E Two-dimensional localization for tracking the mobile robot

This experiment is adding the mobile robot for tracking observation. Place the receiver in the mobile robot and observe the localization results. In Figure 10, this is the result of two-dimensional localization for mobile tracking and the tracking path is (5 , 5)→(5 , 4)→(5 , 3)→(4 , 3)→(3 , 3)→(3 , 2)→(3 , 1).

The wireless location signal strength is affected by fluctuations, so real-time tracking is a great challenge. According to experimental the results, it could be observed that the accuracy rate is low at the coordinate of (1 , 2), (1 , 3), and(1 , 4). It is clear that when the robot makes a turn, the receiver will follow the turning action and causes a positioning error. Therefore, the robot must wait for a while for the signal to be steady so that the positioning results can be more accurate.

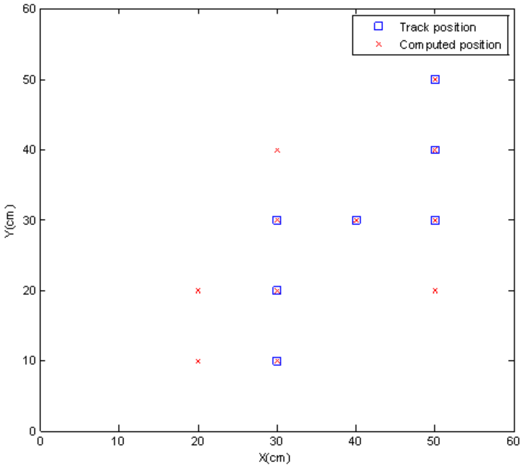


Figure 10: Two-dimensional localization for tracking mobile

In Figure 11, it shows the estimated localization average error of walking path for mobile robot. The average localization error is 14.16 cm.

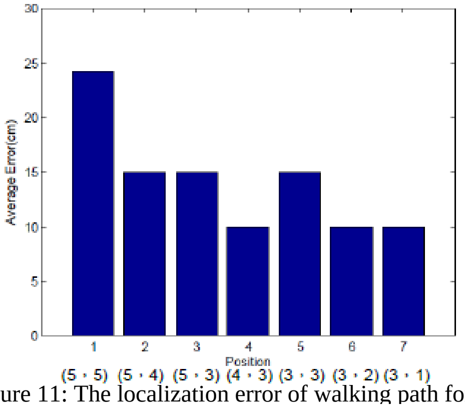


Figure 11: The localization error of walking path for robot

VII. CONCLUSION

This study uses the property of the signal strength for positioning. The accuracy of one-dimensional localization without antennas is 10 cm and the average accuracy rate is about 97%. Therefore, in one-dimensional positioning by using the robot for adaptive positioning control is good.

For two-dimensional positioning in the area of 60 cm x 60 cm, the location average accuracy is about 79.7%, not as good as the one dimensional case due to errors generated by the wireless signal interference from the surrounding objects.

To expand the scope of the present localization method, in addition to increasing the number of RFID tags, the other approach is to unify the antenna specifications with high-gains.

REFERENCES

[1] L. Geunho and C. Nak Young (2011). "Low-Cost DualRotating Infrared Sensor for Mobile Robot Swarm Applications." Industrial Informatics, IEEE Transactions on, vol. **7**(2), pp. 277-286.

[2] J. Pugh, X. Raemy, C. Favre, R. Falconi, and A. Martinoli (2009). "A Fast Onboard Relative Positioning Module for Multirobot Systems." Mechatronics, IEEE/ASME

Transactions on, vol. 14(2), pp. 151-162.. [3] G. Lee, S. Yoon, N. Y. Chong, and H. I. Christensen (2009). "Self-configuring robot swarms with dual rotating infrared sensors." Intelligent Robots and Systems, 2009. IROS 2009. IEEE/RSJ International Conference on, pp. 4357-4362.

[4] J.F. Roberts, T.S. Stirling, J. C. Zufferey, and D. Floreano (2009). "2.5D infrared range and bearing system for collective robotics." Intelligent Robots and Systems. IROS 2009. IEEE/RSJ International Conference on, pp. 3659-3664.

[5] G. Benet, F. Blanes, J. E. Simó, and P. Pérez (2002). "Using infrared sensors for distance measurement in mobile robots." Robotics and Autonomous Systems, , vol. 40(4), pp. 255-266.

[6] T. H. S. Li, Y. C. Yeh, J. D. Wu, M. Y. Hsiao, and C. Y. Chen (2010). "Multifunctional Intelligent Autonomous Parking Controllers for Carlike Mobile Robots." Industrial Electronics, IEEE Transactions on, vol. 57(5), pp. 1687-1700.

[7] L. Kyoungmin and C. Wan Kyun (2009). "Effective Maximum Likelihood Grid Map With Conflict Evaluation Filter Using Sonar Sensors." Robotics, IEEE Transactions on, vol. 25(4), pp. 887-901.

[8] T. Sasaki , D. Brscic, and H. Hashimoto (2010) "Human-Observation-Based Extraction of Path Patterns for Mobile Robot Navigation." Industrial Electronics, IEEE Transactions on, vol. 57(4), pp. 1401-1410.

[9] J. Chen, D. Sun, J. Yang, and H. Chen (2010) "Leader-Follower Formation Control of Multiple Non-holonomic Mobile Robots Incorporating a Receding-horizon Scheme." The International Journal of Robotics Research, vol. 29(6), pp. 727-747.

[10] D. Pizarro , M. Mazo , E. Santiso , M. Marron , D. Jimenez , and S. Cobreces, et al. (2010). "Localization of Mobile Robots Using Odometry and an External Vision Sensor." Sensors, vol. 10(4), pp. 3655-3680.

[11] Ni. Moshtag, N. D. Michae, A. Jadbabaie, and K. Daniilidi (2009). "Vision-Based, Distributed Control Laws for Motion Coordination of Nonholonomic Robots." Robotics, IEEE Transactions on, vol. 25(4), pp. 851-860.

[12] A. Turgut, H. Ç elikkanat, F. Gökçe, and E. ¸Sahin (2008). "Self-organized flocking in mobile robot swarms." Swarm Intelligence, vol. 2(2), pp. 97-120.

[13] L. E. Parker, B. Kannan, T. Fang, and M. Bailey (2004). "Tightly-coupled navigation assistance in heterogeneous multi-robot teams." Intelligent Robots and Systems. (IROS 2004). Proceedings. 2004 IEEE/RSJ International Conference on, pp. 1016-1022.

[14] J. Fredslund and M. J. Mataric (2002). "A general algorithm for robot formations using local sensing and minimal communication." Robotics and Automation, IEEE Transactions on, vol. 18(5), pp. 837-846.

[15] H. Surmann, A. Nüchter, and J. Hertzberg (2003). "An autonomous mobile robot with a 3D laser range finder for 3D exploration and digitalization of indoor environments." Robotics and Autonomous Systems, vol. 45(3-4), pp. 181-198.

[16] C. Ye (2007). "Navigating a mobile robot by a traversability field histogram." Ieee Transactions on Systems Man and Cybernetics Part B-Cybernetics, vol. 37(2), pp. 361-372.

[17] R. Kurazume, Y. Noda, Y. Tobata, K. Lingemann, Y. Iwashita, and T. Hasegawa (2009). "Laser-based geometric modeling using cooperative multiple mobile robot." Robotics and Automation, 2009. ICRA '09. IEEE International Conference on, pp. 3200-3205.

[18] M. Betke and L. Gurvits (1994). "Mobile robot localization using landmarks." Intelligent Robots and Systems. 'Advanced Robotic Systems and the Real World', IROS '94. Proceedings of the IEEE/RSJ/GI International Conference on, pp. 135-142.

[19] A. Lazanas and J. C. Latombe (1995). "Landmark-Based Robot Navigation." Algorithmica, vol. 13(5), pp. 472-501.

[20] L. M. Ni, Y. L. Yiu, C. Lau, and A. P. Patil (2004). "LANDMARC: Indoor Location Sensing Using Active RFID." Wireless Networks, vol. 10(6), pp. 701-710.

[21] T. Tsukiyama (2002). "Global navigation system with RFID tags", SPIE.

[22] K. Yamano, K. Tanaka, M. Hirayama, E. Kondo, Y. Kimuro, and M. Matsumoto (2004). "Self-localization of mobile robots with RFID system by using support vector machine." Intelligent Robots and Systems. (IROS 2004). Proceedings. 2004 IEEE/RSJ International Conference on, pp. 3756-3761.

[23] H. D. Chon, S. Jun, H. Jung, and S. W. An (2005). "Using RFID for Accurate Positioning." Journal of Global Positioning Systems, vol. 3, pp. 32-39.

[24] S. Han, H. Lim, and J. Lee (2007). "An Efficient Localization Scheme for a Differential-Driving Mobile Robot Based on RFID System." Industrial Electronics, IEEE Transactions on, vol. 54(6), pp. 3362-3369.

[25] P. Sunhong, R. Saegusa, and S. Hashimoto (2007) "Autonomous navigation of a mobile robot based on passive RFID." Robot and Human interactive Communication. RO-MAN 2007. The 16th IEEE International Symposium on, pp. 218-223.

[26] W. Gueaieb and M. S. Miah (2008). "An Intelligent Mobile Robot Navigation Technique Using RFID Technology." Instrumentation and Measurement, IEEE Transactions on, vol. 57(9), pp. 1908-1917.

[27] P. Sunhong and S. Hashimoto (2009). "Autonomous Mobile Robot Navigation Using Passive RFID in Indoor Environment." Industrial Electronics, IEEE Transactions on, vol. 56(7), pp. 2366-2373.

[28] G. Enriquez, P. Sunhong, and S. Hashimoto (2010). "Wireless sensor network and RFID sensor fusion for mobile robots navigation." Robotics and Biomimetics (ROBIO), 2010 IEEE International Conference on, pp. 1752-1756.

[29] R. Tesoriero, R. Tebar, J. A. Gallud, M. d. Lozano, and V. M. R. Penichet (2010). "Improving location awareness in indoor spaces using RFID technology." Expert Systems with Applications, vol. 37(1), pp. 894-898.

[30] Y. Park, J. W. Lee, D. Kim, J. J. Jeong , and S. W. Kim (2010). "Mathematical formulation of RFID tag floor based localization and performance analysis for tag placement." Control Automation Robotics & Vision (ICARCV), 2010 11th International Conference on, pp. 1547-1552.

[31] S. S. Saad and Z. S. Nakad (2011). "A Standalone RFID Indoor Positioning System Using Passive Tags." Industrial Electronics, IEEE Transactions on, vol. 58(5), pp. 1961-1970.

[32] J. Blumenthal, R. Grossmann, F. Golatowski, and D. Timmermann (2007) "Weighted Centroid Localization in Zigbee-based Sensor Networks." Intelligent Signal Processing. WISP 2007. IEEE International Symposium on, pp. 1-6.

[33] I. Yamada, T. Ohtsuki, T. Hisanaga, and L. Zheng (2007) "An indoor position estimation method by maximum likelihood algorithm using RSS." SICE, 2007 Annual Conference, pp. 2927-2930.

Joint Design of RFID Reader and Tag Anti-Collision Algorithms: A Cross-Layer Approach

Ramiro Sámano-Robles and Atílio Gameiro
Instituto de Telecomunicações, Campus Universitário, Aveiro, 3810-193, Portugal.
emails:ramiro@av.it.pt; amg@ua.pt

***Abstract*—This paper investigates the potential interactions between reader and tag anti-collision algorithms of passive RFID (radio frequency identification) systems. Conventionally, reader and tag anti-collision algorithms are designed by assuming that they are independent from each other. In practice, however, readers and tags usually operate in the same frequency band. Therefore, contention between their transmissions can also potentially arise. Furthermore, reader anti-collision policies directly influence the way in which tags are activated, and thus also the way in which they collide when responding to reader's requests. In view of this and considering the growing numbers of readers and tags, independence of both schemes can not longer be considered as a realistic assumption. This paper partially fills this gap by proposing a new cross-layer framework for the joint evaluation and optimization of reader and tag anti-collision algorithms. Furthermore, the paper proposes a new approach, based on a Markov model, which allows capacity and stability analysis of asymmetrical RFID systems (i.e., when readers and tags experience different channel and queuing states). The model captures the dynamics of tag activation and tag detection processes of RFID. It also represents a first step towards a joint design of physical (PHY) and medium access control layers (MAC) of RFID. The results indicate that the proposed approach provides benefits in terms of stability and capacity over conventional solutions even when readers and tags operate in different channels. The results also provide useful guidelines towards the cross-layer design of future RFID platforms.**

***Index Terms*—RFID anti-collision algorithms, cross-layer design, random access theory.**

I. Introduction

A. RFID technology and previous works

RFID (Radio Frequency Identification) is a technology that uses radio frequency signals for purposes of identification and tracking of objects, humans or animals [1]. In passive systems, where tags reuse the energy radiated by the reader, coordination capabilities are considerably limited, thereby leading to signal collisions. Therefore, an efficient medium access control layer (MAC) is crucial to the correct operation of RFID [2].

Two types of RFID MAC collision can be distinguished: tag and reader collision. A tag collision arises when several tags simultaneously respond to the same reader request, thus causing the loss of information. To address this issue, tag anti-collision schemes such as ALOHA and binary tree algorithms are commonly employed [2]. Improvements on these solutions have been further proposed by using tag estimation methodologies [3], and modified frame structures [2]. Two types of reader collision can be also identified: multiple-reader-to-tag and reader-to-reader collision [4]. To address these issues, reader anti-collision algorithms based on scheduling or coverage control have been proposed. Typical scheduling schemes are frequency division multiple access (FDMA) [5] or listen-before-talk (LBT) [6]. Advanced schemes such as Colorwave in [7] and Pulse in [8] implement inter-reader control mechanisms to assist in collision avoidance, whereas HiQ in [9] uses analysis of collision patterns to improve scheduling. In coverage-based algorithms, we find schemes that reduce the overlapping coverage area between readers (e.g., [10]), and those that monitor interference to adapt power levels accordingly (e.g., [12]).

B. Paper contributions

Despite these advances in RFID MAC design, several issues remain open. This paper addresses some of these issues and proposes several advances over previous solutions. The paper addresses for the first time in the literature (to the best of our knowledge) the joint design of reader and tag anti-collision algorithms. To achieve this goal, a novel framework for cross-layer design of MAC and PHY (physical) layers of RFID is also proposed. Based on this framework, a Markov model is further presented for the study of capacity and stability of asymmetrical RFID systems, which is also new in the literature. More details on these objectives and the rationale behind them are next explained.

1) Joint design of reader and tag anti-collision: In conventional RFID system design, reader and tag anti-collision algorithms are considered as independent from each other. This means that reader anti-collision schemes ignore tag collisions, and viceversa, tag anti-collision schemes ignore reader collisions. The reason for this is that the number of readers is low in typical RFID applications, which means that reader collisions rarely occur. However, recent years have seen an increasing numbers of readers and tags. Also, readers and tags of passive systems usually operate in the same frequency band, which increases the probability of collision between their transmissions. Furthermore, reader anti-collision schemes directly induce tag collision patterns. Therefore, the assumption of independence of these two schemes does not longer hold. The objective of this paper is to fill this gap by studying the interactions between all the elements of a multi-tag and multi-reader RFID network.

2) Cross-layer design in RFID: In order to achieve an accurate evaluation of multi-tag and multi-reader systems, this paper proposes a novel theoretical framework which includes relevant PHY and MAC layer parameters. Previous works on RFID MAC design have used simplistic formulations of the PHY layer which are inaccurate when modeling real-life systems. In general, cross-layer design has been scarcely used in the literature of RFID. At the MAC/PHY level, some anti-collision algorithms based on power control and reader scheduling can be considered as cross-layer solutions (e.g.,[12]), but they have not been explicitly designed with a cross-layer methodology. At upper layers, only a few cross-layer solutions using context aware analysis have been shown to significantly improve reading reliability levels (e.g., [13]) and security/privacy features (e.g.,[14]). By contrast, in conventional wireless networks cross-layer design has shown considerable benefits, particularly at the MAC/PHY level [11]. Therefore, there is a big potential in using cross-layer design to improve RFID. The proposed framework in this paper represents an initial step towards a full joint design of RFID PHY and MAC layers [15]. Stochastic reception models for correct tag activation and correct tag detection probabilities considering channel and queuing states are here proposed. This stochastic framework can also be used, for example, to describe advanced multiuser detection schemes for RFID applications.

3) Asymmetrical scenarios, results and future work: RFID MAC algorithms have been conventionally modeled in symmetrical scenarios, i.e. when all elements are statistically identical. However, this assumption is unrealistic and can lead to inaccurate design. This paper also proposes a Markov model that allows capacity and dynamic stability analysis of asymmetrical RFID systems (i.e., readers and tags have different channel and queuing states). This approach is, to the best of our knowledge, new in the literature of RFID, as it captures the dynamics of tag activation and detection processes. The results in this paper indicate that joint cross-layer optimization of reader and tag anti-collision algorithms provides considerable benefits in terms of capacity and stability even when readers and tags operate in different channels. The proposed approach was found particularly helpful in the asymmetrical case. Future work will include the use of advanced algorithms such as beam-forming, retransmission diversity, and multi-packet reception.

C. Paper organization

Section II describes the proposed framework for cross-layer optimization with the signal models for down-link and up-link reception. Section III describes the proposed metrics, the tag reception and activation probabilities and the Markov model for dynamic analysis. Section IV presents the optimization of the throughput and the results obtained in different scenarios. Finally, Section V presents the conclusions of the paper.

II. System model and cross-layer framework

Consider the slotted RFID network depicted in Fig. 1 with a set $\mathcal{R}$ of K readers, i.e., $\mathcal{R} = \{1, \dots K\}$, and a set $\mathcal{T}$ of J tags, i.e., $\mathcal{T} = \{1, \dots, J\}$. Two main processes can be distinguished in the RFID network in Fig. 1: Tag activation by the transmission of readers, also called the down-link transmission; and the backscattering response towards readers by previously activated tags, also called up-link transmission (see Fig. 1). In the down-link, the transmit power of reader k will be denoted by $P_{r,k}$ while its probability of transmission will be denoted by $p_{r,k}$. The subset of active readers at any given time will be denoted by $\mathcal{R}_t$. Tags are activated whenever the energy received from a reader is above an activation threshold. The set of activated tags will be denoted here by $\mathcal{T}_P$ ($\mathcal{T}_P \subseteq \mathcal{T}$). These active tags proceed to transmit a backscatter signal to the readers using a randomized transmission scheme. The subset of tags that transmit a backscatter signal once they have been activated will be given by $\mathcal{T}_t(\mathcal{T}_t \subseteq \mathcal{T}_P \subseteq \mathcal{T})$, where each tag $j \in \mathcal{T}_t$ will transmit with a power level denoted by $P_{t,j}$.

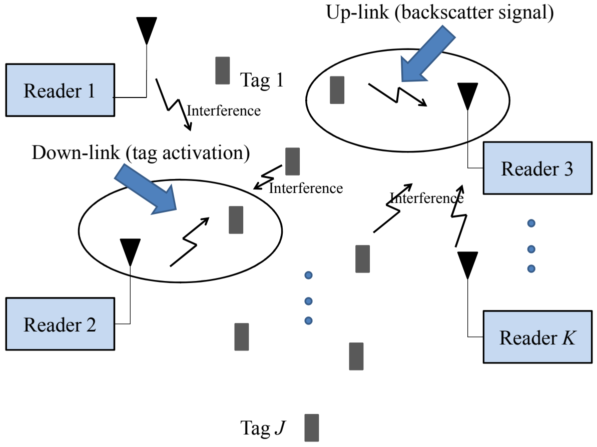


Fig. 1. Multi-tag and Multi-reader deployment scenario.

A. Tag activation: Down-link model

For convenience in the analysis, consider that the channel between reader k and tag j is given by $h_{k,j}$. Similarly, the channel between reader k and reader m is given by $g_{k,m}$, and the channel between tag i and tag j is given by $u_{i,j}$. Therefore, the signal-to-interference-plus-noise ratio (SINR) experienced by tag j due to a transmission of reader k will be denoted by $\gamma_{k,j}$, and it can be expressed as follows:

$$\gamma_{k,j} = \frac{P_{r,k}|h_{k,j}|^2}{I_{r_{k,j}} + I_{t_j} + \sigma^2_{v,j}}, \qquad k \in \mathcal{R}_t \qquad (1)$$

where $I_{r_{k,j}} = \sum_{m\in\mathcal{R}_t, m\neq k} P_{r,m}|h_{m,j}|^2$ is the interference created by other active readers, $I_{t_j} = \sum_{i\in\mathcal{T}_t, i\neq j} P_{t,i}(|u_{j,i}|^2)$ is the interference created by other contending tags, and $\sigma^2_{v,j}$ is the noise component. If the SINR experienced by tag

j is above the tag sensitivity threshold $\widetilde{\gamma}_j$, then the tag is powered-up and is then considered as active. The probability of tag j being activated can be written as $\Pr\{j \in \mathcal{T}_P\} = \Pr\{\max_k \gamma_{k,j} > \widetilde{\gamma}_j\}$.

B. Backscattering reply and tag detection: up-link model

Once a given tag j has been activated, it starts a random transmission process to prevent collisions with other active tags. This random transmission control will be characterized by a Bernoulli process with parameter $p_{t,j}$, which is also the transmission probability. We consider the backscattering factor β_j as the fraction of the received power reused by the tag to reply to the reader. Therefore, the transmit power of tag j can be calculated as $P_{t,j} = \beta_j P_{r,k}|h_{k_{opt},j}|^2$, where $k_{opt} = \arg\max_k \gamma_{k,j}$ denotes the reader that has previously activated the tag. The SINR of the signal of tag j received by reader k can then be written as:

$$\widehat{\gamma}_{j,k} = \frac{P_{t,j}|h_{j,k}|^2}{\widehat{I}_{r,k} + \widehat{I}_{t_{j,k}} + P_{r,k}\eta_k + \widehat{\sigma}^2_{v,k}}, \qquad j \in \mathcal{T}_t \qquad (2)$$

where $\widehat{I}_{r,k} = \sum_{m \neq k} P_{r,m}|g_{m,k}|^2$ is the interference created by other active readers, $\widehat{I}_{t_{j,k}} = \sum_{i \neq j} P_{t,i}|h_{i,k}|^2$ is the interference created by other active tags, η_k is the power ratio leaked from the down-link transmission chain, and $\widehat{\sigma}^2_{v,k}$ is the noise. Tag j can be detected by reader k if the received SINR is above a threshold denoted by $\check{\gamma}_k$. The set of detected tags by reader k will be denoted by $\mathcal{T}_{D,k}$, while the probability of tag j being in $\mathcal{T}_{D,k}$ will be given by $\Pr\{j \in \mathcal{T}_{D,k}\} = \Pr\{\widehat{\gamma}_{j,k} > \check{\gamma}_k\}$.

III. Performance metrics and Markov model

The main performance metric to be used in this paper is the average tag throughput, which can be defined as the long term ratio of correct tag readings to the total number of time-slots used in the measurement. Before providing an expression for this metric, it is first necessary to define the network state information, as well as the tag activation and tag reception probability models, and the definition of the Markov model for the dynamic analysis of an RFID network.

A. Network state information and tag activation model

The network state information can be defined as all the parameters that completely describe the network at any given time. In our case, the network state information $\mathcal{N}(n)$ at time slot n is defined as the collection of the sets of active readers $\mathcal{R}_s(n)$ and contending tags $\mathcal{T}_t(n)$:

$$\mathcal{N}(n) = \{\mathcal{R}_s(n), \mathcal{T}_t(n)\}.$$

Once the network state information has been defined, we can define the probability of tag j being activated in slot n conditional on a given realization of the network state information $\mathcal{N}(n)$ as follows:

$$Q_{j|\mathcal{N}(n)} = \Pr\{j \in \mathcal{T}_P(n+1)|\mathcal{N}(n)\} =$$

$$\Pr\{\max_k \gamma_{k,j}(n) > \widetilde{\gamma}_j\}.$$

For convenience in the analysis, let us rewrite this tag activation probability in terms of the set of active tags $\mathcal{T}_P(n)$ by averaging over all values of $\mathcal{N}(n)$ where $\mathcal{T}_t(n) \in \mathcal{T}_P(n)$:

$$Q_{j|\mathcal{T}_P(n)} = \sum_{\mathcal{N}(n); \mathcal{T}_t(n) \in \mathcal{T}_P(n)} \Pr\{\mathcal{N}(n)\} Q_{j|\mathcal{N}(n)}$$

where $\Pr\{\mathcal{N}(n)\}$ is the probability of occurrence of the network state information $\mathcal{N}(n)$. This term can be calculated by considering all the combinations of active tags and readers as follows:

$$\Pr\{\mathcal{N}(n)\} = \prod_{k \in \mathcal{R}_t} p_{r,k} \prod_{m \notin \mathcal{R}_t} \overline{p}_{r,m} \prod_{j \in \mathcal{T}_t} p_{t,j} \prod_{i \notin \mathcal{T}_t} \overline{p}_{t,i}$$

where $\overline{(\cdot)} = 1 - (\cdot)$. This concludes the definition of the tag activation probability and the network state information.

B. Markov model

In order to define the Markov model for dynamic analysis of the system, let us now calculate the probability of having a set of active tags $\mathcal{T}_P(n+1)$ in time slot $n+1$ conditional on having the set of active tags $\mathcal{T}_P(n)$ during the previous time-slot. This transition probability must consider all the combinations of tags that either enter (i.e., they are activated in time slot n) or leave the set of active tags (i.e., they transmit in time slot n). This can be expressed as follows:

$$\Pr\{\mathcal{T}_P(n+1)|\mathcal{T}_P(n)\} = \prod_{j \in \mathcal{T}_P(n), j \notin \mathcal{T}_P(n+1)} p_{t,j}$$

$$\times \prod_{i \notin \mathcal{T}_P(n), i \in \mathcal{T}_P(n+1)} Q_{i|\mathcal{T}_P(n)} \prod_{l \notin \mathcal{T}_P(n), l \notin \mathcal{T}(n+1)} \overline{Q}_{l|\mathcal{T}_P(n)}$$

$$\times \prod_{w \in \mathcal{T}_P(n), w \in \mathcal{T}_P(n+1)} \overline{p}_{t,w}.$$

Let us now arrange the probability of occurrence of all the possible sets of activated tags $\Pr\{\mathcal{T}_P\}$ into a one-dimensional vector given by $\mathbf{s} = [s_0, \dots s_{J^J}]^T$, where $(\cdot)^T$ is the transpose operator (see Fig. 2). This means that we are mapping the asymmetrical states into a linear state vector where each element represents the probability of occurrence of one different state $\Pr\{\mathcal{T}_P\}$. In the example given in Fig. 2 we have only two possible tags, where the first system state is given by both tags as active, the second state with only tag 1 as active, the third state with only tag 2 as active, and the fourth state with none tag active. Once these states are mapped into the state vector $\mathbf{s}$, the transition probabilities between such states ($\Pr\{\mathcal{T}_P(n+1)|\mathcal{T}_P(n)\}$) can also be mapped into a matrix $\mathbf{M}$, which defines the Markov model for state transition probabilities (see Fig. 2). The i, j entry of the matrix $\mathbf{M}$ denotes the transition probability between state i and state j. The vector of state probabilities can thus be obtained by solving the following characteristic equation:

$$\mathbf{s} = \mathbf{M}\mathbf{s},$$

using standard eigenvalue analysis or iterative schemes. Each one of the calculated terms of the vector $\mathbf{s}$ can be mapped back to the original probability space $\Pr\{\mathcal{T}_P\}$, which can then be used to calculate relevant performance metrics.

C. Tag detection model

Before calculating the tag throughput, first we must define the correct reception probability of tag j at the reader side conditional on the network state information $\mathcal{N}(n)$ as follows:

$$q_{j|\mathcal{N}(n)} = \Pr\{j \in \mathcal{T}_P(n+1)\} = \sum_{k\in\mathcal{R}} \Pr\{\hat{\gamma}_{j,k}(n) > \check{\gamma}_k\}$$

It is also convenient to re-write this reception probability in terms of the set of active tags $\mathcal{T}_P(n)$ by averaging over all values of $\mathcal{N}(n)$ where $\mathcal{T}_t(n) \in \mathcal{T}_P(n)$:

$$q_{j|\mathcal{T}_P(n)} = \sum_{\mathcal{N}(n);\mathcal{T}_t(n)\in\mathcal{T}_P(n)} \Pr\{\mathcal{N}(n)\} q_{j|\mathcal{N}(n)}$$

D. Tag throughput and stability

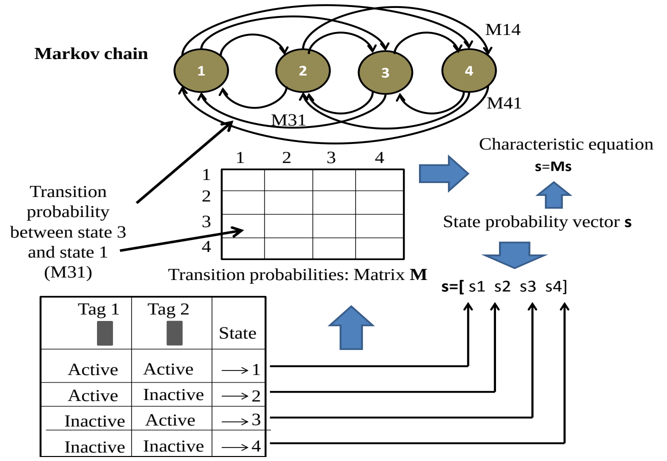
The tag throughput can be finally calculated by adding all the contributions over the calculated probability space $\Pr\{\mathcal{T}_P\}$ using the Markov model presented in previous subsections. This can be mathematically expressed as:

$$T_j = \sum_{\mathcal{T}_P} \Pr\{\mathcal{T}_P\} q_{j|\mathcal{T}_P}. \qquad (3)$$

As a measure of stability we will use the average number of activated tags, which can be calculated as follows:

$$E[|\mathcal{T}_P|] = \sum_{\mathcal{T}_P} \Pr\{\mathcal{T}_P\} |\mathcal{T}_P|. \qquad (4)$$

A high number of active tags means that stability is compromised, while a relatively low number indicates that the algorithm is more stable.


Markov chain
1
2
3
4
M14
M41
M31
Transition probability between state 3 and state 1 (M31)
1 2 3 4
1
2
3
4
Transition probabilities: Matrix **M**
Characteristic equation
s=Ms
State probability vector **s**
s=[s1 s2 s3 s4]

Tag 1	Tag 2	State
Active	Active	⟶1
Active	Inactive	⟶2
Inactive	Active	⟶3
Inactive	Inactive	⟶4

Fig. 2. Example of the Markov model for a two-tag system.

IV. Optimization and Results

The parameters to be optimized are the vector of reader transmission probabilities $\mathbf{p}_r = [p_{r,1}, \ldots p_{r,K}]^T$, the vector of reader transmit powers $\mathbf{P}_r = [P_{r,1}, \ldots P_{r,K}]$ and the vector of transmission probabilities of the active tags $\mathbf{p}_t = [p_{t,1}, \ldots p_{t,J}]$. The objective of the optimization is the total tag throughput, so the optimization problem can thus be written as follows:

$$\{\mathbf{P}_r, \mathbf{p}_t, \mathbf{p}_r\}_{opt} = \arg \max_{\{\mathbf{P}_r, \mathbf{p}_t, \mathbf{p}_r\}} \sum T_j$$
$$\text{subject to} \quad \mathbf{P}_r < \mathbf{P}_{r,0} \qquad (5)$$

where $\mathbf{P}_{r,0}$ is the reader transmit power constraint vector. Since the explicit optimization of the expressions is difficult to achieve, particularly when considering the Markov model proposed in the previous section, in this section we will simplify the optimization problem by applying the previous concepts to an ALOHA protocol implemented both at the reader and the tag side. This means that we consider that any collision yields the loss of all information. Two different cases will be considered: one in which readers and tags transmit in orthogonal channels, i.e. no collision exists between the transmissions of readers and tags, and the second case in which readers and tags transmit at free will in the same channel, thereby leading to potential collisions. Under these assumptions, the power optimization problem reduces to simply setting the transmit power of the readers to a particular level to ensure an average tag activation and tag detection probabilities. The remaining variables to be optimized are the reader and tag transmission probabilities. To further illustrate the operation of the proposed approach, two different scenarios will be addressed: one in which all elements are statistically identical (symmetrical scenario) and another with asymmetrical features (some tags and readers have different characteristics).

A. Symmetrical case with tags and readers operating in different channels

The first scenario consists of $R = 5$ readers and $J = 15$ tags all with the same channel and queuing states (symmetrical case). Tags and readers are assumed to work in different channels. Fig. 3 displays the results of using the throughput expression in eq.(3) for various values of reader and tag transmission probabilities (p_r and p_t). MATLAB is used for calculating the results and solving explicitly the Markov chain model presented in the previous section and thus obtain the steady state vector and the throughput values. All the cases discussed in this section will use a fixed transmit power that has been set to provide the following tag activation and detection probabilities: $Q_j = 0.7$ and $q_j = 0.95$, respectively. The surface shape of the global tag throughput in Fig. 3 is slightly asymmetrical, which indicates that the optimum probabilities cannot be considered as completely independent. Consider the values of optimum transmission probability without joint design for the ALOHA protocol: $p_{t_{opt}} = 1/15$ and $p_{r_{opt}} = 1/5$, which yields a value of throughput in Fig. 3 of 0.354. By contrast, the true maximum of the system, which can be only found using joint optimization, yields 0.3558 with values $p_{t_{opt}} = 0.085$ and $p_{r_{opt}} = 0.1975$. Regarding stability, Fig. 4 shows the average number of active tags, where it can be observed that joint design can also help in driving the system towards a region with low numbers of backlogged tags. By using joint optimization, the achieved value for the average number of active tags is 9, whereas using the original strategy it would yield an average of 9.8 active tags. Therefore, in this case it has been observed that even in a completely symmetrical scenario where tags and readers operate in different channels, joint design also provides benefits in terms of capacity and stability.

B. Symmetrical case with tags and readers operating in the same channel

Fig. 5 displays the results for the same scenario but this time considering full interference between tags and readers. In this case, the shape is even more asymmetrical, which is consistent with the assumption of full interference between tags and readers and which indicates the increased importance of joint optimization. Consider the values of optimum transmission probability without joint design $p_{t_{opt}} = 1/15$ and $p_{r_{opt}} = 1/5$, which yields a value of throughput in Fig. 5 of 0.1789. By contrast the true maximum of the system, which can only be found using joint optimization, yields 0.2499 with values $p_{t_{opt}} = 0.035$ and $p_{r_{opt}} = 0.085$. This indicates that joint design increases its efficiency when readers and tags operate in the same channel.

C. Asymmetrical case with tags and readers operating in different channel

In the second scenario, we consider that the tag/reader space is divided into two different sets of readers and three different sets of tags (asymmetrical scenario). Readers and tags are working in different channels. The first and second sets of tags can only be reached by the first and second sets of readers, respectively. The third set of tags can be reached by both sets of readers. All tags have the same transmission probability p_t as well as all readers transmit with the same parameter p_r. The results displayed in Fig. 6 show that the throughput has also an irregular shape, which indicates a complex dependency between the transmission probabilities and, hence, an increased advantages of using joint optimization. The results of Fig. 6 have been obtained using three groups of tags with $J_1 = 3$,$J_2 = 5$ and $J_3 = 7$ tags, and two groups of readers with $R_1 = 5$ and $R_2 = 10$ readers. The maximum of the global throughput using joint optimization is 0.4780, whereas using the conventional strategy is 0.4413. Therefore, joint optimization provides even higher gains in this asymmetrical scenario as compared to its symmetrical counterpart. According to these results, joint optimization is well suited for asymmetrical scenarios. However, in an RFID network, accurate tracking of different channels of the tags is a difficult task. A solution to this problem is to use context aware techniques that allow us to estimate tag relative positions with respect to the set of readers. Thus, it is foreseen that joint optimization will be further improved by exploiting context information acquired from different layers. In the future, RFID systems can be based on cross-layer design and then help in the acquisition of all the relevant information to carry out a more efficient optimization.

V. Conclusions

This paper addressed the joint optimization and design of reader and tag anti-collision algorithms for RFID systems. A general framework was developed for cross-layer evaluation and optimization of these two contention resolution schemes. Basic examples using ALOHA protocol have shown that in all the cases joint optimization provides benefits to global system operation in terms of stability and capacity even when readers and tags operate in different channels. The results also indicate that the gains provided by the joint optimization approach increase when the scenario deviates from the symmetrical case, which also means that in a real system deployment context aware information can be used to further improve the joint optimization process. The tools developed in this paper represent a first step towards the full joint design of MAC and PHY layers of RFID systems. The expressions derived in this paper also allow the investigation of advanced signal processing schemes for multi-packet reception which will be addressed in future works.

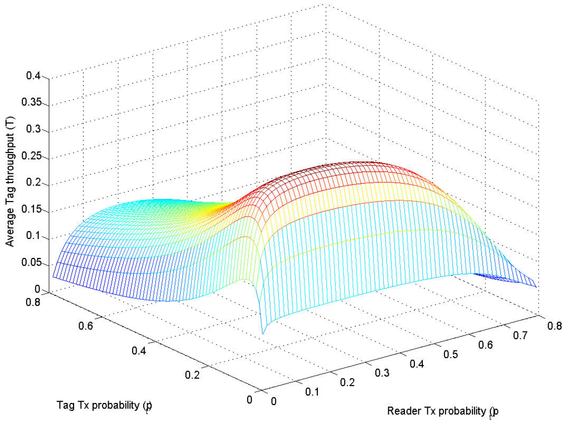


Fig. 3. Throughput (T) vs. reader and tag transmissions probabilities (p_r and p_t) of a symmetrical ALOHA protocol for reader and tag anti-collision assuming no interference between readers and tags.

References

[1] R. Weinstein, "RFID: A Technical overview and its application to the enterprise," *IT profesional,* 2005, Vol. 7, No. 3, pp. 27-33.

[2] L.A. Burdet, "RFID Multiple Access Methods," *ETH Zrich, Summer semester 2004, Seminar "Smart Environments.*

[3] D. Liu, Z. Xang, J. Tan, H. Min, and J. Wang, "ALOHA algorithm considering the slot duration difference in RFID system," *2009 IEEE International Conference on RFID*, pp. 56-63.

[4] S. Birari, "Mitigating the reader collision problem in RFID Networks with mobile readers," Master Thesis.

[5] "EPC Radio-Frequency Identity Protocols. Class-1 Generation-2 UHF RFID. Protocol for communications at 860 MHz - 960Mhz." Version1.2.0", EPC Global, 2008., http://www.epcglobalinc.org/standards/.

[6] "ETSI EN 302 208-1,2 v1.1.1, September 2004. CTAN: http://www.etsi.org"

[7] J. Waldrop, D.W. Engels, and S.E. Sarma, "Colorwave: An anticollision algorithm for the reader collision problem," *IEEE Wireless Communications and Networking Conference (WCNC), 2003.*

[8] S. Birari, "Mitigating the reader collision problem in RFID Networks with mobile readers," *13th IEEE Int. Conf. on Networks, 2005*, pp. 463-468.

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-193-9

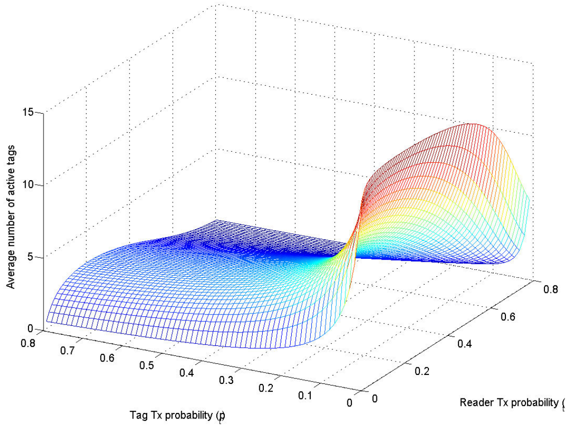


Fig. 4. Average number of active tags ($E[|\mathcal{T}_P|]$) vs. reader and tag transmissions probabilities (p_r and p_t) of a symmetrical ALOHA protocol assuming no interference between readers and tags.

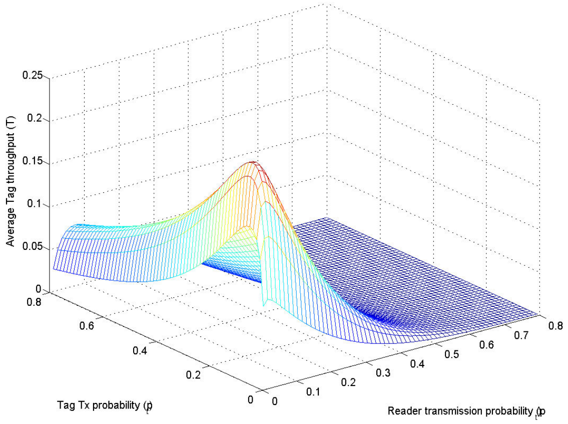


Fig. 5. Throughput (T) vs. reader and tag transmissions probabilities (p_r and p_t) of a symmetrical ALOHA protocol for reader and tag anti-collision assuming full interference between readers and tags.

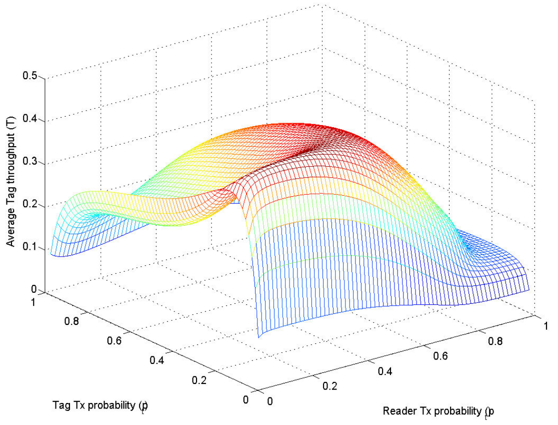


Fig. 6. Throughput (T) vs. reader and tag transmissions probabilities (p_r and p_t) of an asymmetrical ALOHA protocol for reader and tag anti-collision assuming no interference between readers and tags.

[9] K.Ho. Junius, "Solving the reader collision problem with a hierarchical q-learning algorithm," *Master's thesis,*MIT, February, 2003.

[10] S.Y. Kim and J.K. Lee "A study on control method to reduce collisions and interferences between multiple RFID readers and RFID tag," *International Conference on New Trends in Information and Service Science NISS*, 2009, pp.339-343.

[11] V. Srivastaya and M. Montani "Cross-layer design: a survey and the road ahead," *IEEE Commun. Magazine*, Vol. 43, No. 12, December 2005, pp. 112-119.

[12] K. Cha, S. Jagannathan, and D. Pommerenke " Adaptive power control with hardware implementation for wireless sensor and RFID networks," *IEEE Systems Journal,* Vol. 1, No. 2, December 2007, pp. 145-159.

[13] N. Ahmed, R. Kumar, R.S. French, and U. Ramachandran "RF^{2}ID: A Reliable Middleware Framework for RFID Deployment," *IEEE Int. Parallel and Distributed processing Symposium*, March 2007, pp. 1-10.

[14] T. Kriplean, R. Kumar, R.S. French, and U. Ramachandran "Physical Access control for captured RFID Data," *IEEE Pervasive Computing*, Vol. 6, No. 4, 2007, pp. 48-55.

[15] R. Samano-Robles and A. Gameiro "Collision resolution algorithms for RFID applications," *Asia-Pacific Microwave Conference*, 2008, pp. 1-4.

Wireless Network Localization

Optimization Processing

Lukas Klozar, Jan Prokopec
Department of Radio Electronics
FEEC, Brno University of Technology
Brno, Czech Republic
xkloza00@stud.feec.vutbr.cz, prokopec@feec.vutbr.cz

***Abstract*—This paper deals with localization in wireless cellular networks. We performed measurement of the received signal strength in and around Brno city and stored the collected data. The localization approach uses multislope channel models to estimates the propagation distance from the signal strength. Results are processed by two localization techniques. The first one is geometrically based with a triangular constellation of BSs. The second one is independent on the number of connected BSs, however more linked BSs with a triangular constellation refine the localization precision. This technique uses an optimization algorithm and proves to be universal and more accurate.**

Keywords-Channel; localization; modelling; multislope; optimization; propagation; wireless.

I. Introduction

Wireless mobile communication networks are widely spread all around the world and new sites are built contemporary. This brings new opportunities for localization. Many techniques were developed for positioning in wireless networks [1][2][10][11][12][13]. In addition, the dedicated wireless satellite networks like GPS, GLONASS or GALILEO provide localization and navigation services [2].

The motivation of this work is to investigate localization capabilities of an optimization approach in wireless networks. We have measured parameters of GPS and GSM networks in an urban environment of Brno city. With the combined GPS/GSM module XT65 [3] we have collected localization data (latitude, longitude, received signal strength, timing advance, cell identity, signal frequency) and processed them in Matlab.

The measurement of the received signal level in the GSM network is influenced by individual propagation conditions. Therefore, the positioning is ambiguous and additional processing is required to increase precision. We applied the adaptable multislope propagation channel model on the measured values of the signal level to estimate the propagation distance. Next, to determine the final position we propose two localization algorithms. The first one is geometric-based and it is strongly dependent on a network constellation. The second one uses an optimization approach with a mean square error (MSE) estimation to find the Mobile Station (MS) position in two-dimensional space.

At the beginning of this paper is a theoretical description of localization principles in dedicated wireless networks (GPS, GSM). Next, the experimental measurement is introduced. An optimization multislope channel modeling technique is presented [9] to estimate a propagation distance. Next, two localization techniques are described and tested. Results are compared with other techniques described in [11][12]. In conclusion, the results are summarized and further improvements are proposed.

II. GPS Localization

GPS uses the physical model of the Earth called WGS 84 [2]. The localization applied in GPS uses a precise time synchronization, and the GPS receiver measures the Time of Arrival (TOA) [2]. The satellite signals should be received from at least four satellites to achieve sufficient accuracy. Satellite Based Augmentation Systems (SBAS) could provide additional corrections in a receiver. Its accuracy depends on the constellation and the number of visible satellites. With the SBAS [3] is the XT65's Circular Error Probable (CEP) 2 m and Spherical Error Probable (SEP) 3 m [3].

III. GSM Localization

GSM is the wireless network with a cellular architecture (Figure 1). There are many techniques developed for localization [1]. However, the precise position data of BSs (Base Stations) are mandatory for all of them.

A. *Network Based Localization*

The GSM network was not designed for localization, thus this approach requires additional hardware enhancements in network architecture. The Location Measurement Unit (LMU) needs to be involved to perform time based measurements and computation. With the LMU unit the network is capable of using localization techniques TDOA (Time Difference of Arrival), OTD (Observed Time Difference), E-OTD (Enhanced-OTD). Additional capabilities for localization provide the combination with GPS receiver denoted as A-GPS (Assisted-GPS) and also the measurement of AOA (Angle of Arrival). We focus this work on the processing of the received signal strength information by the MS as the part of service communication.

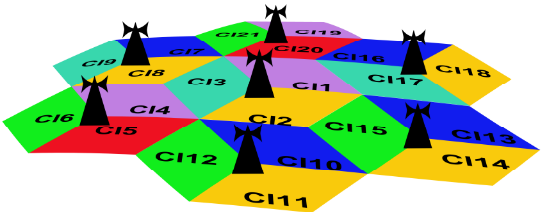


Figure 1. Cell identity of the sectors in a cellular network.

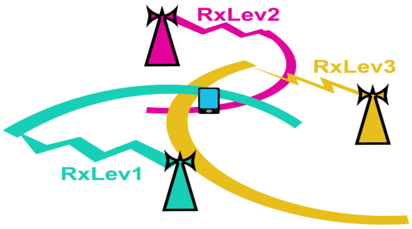


Figure 2. Localization approach based on the measured signal strength

B. Cell Identity Technique

The Cell Identity (CI) code uniquely identifies each sector in the network as illustrated in Figure 1. The Broadcast Control Channel (BCCH) [1] carries the CI code expressed in hexadecimal format as part of the common service communication.

With the database of identification and position information of BSs it is possible to track the MS moving in the network according to the actual CI value i.e., trace connected sectors of BSs. Other codes used for the identification of the MS in the network: Location Area Code (LAC), Mobile Network Code (MNC), Mobile Country Code (MCC) [1].

C. Timing Advance Technique

The Timing Advance (TA) information transmitted in the BCCH corresponds with a propagation delay of a transmitted signal. The TA interval in the GSM network served to avoid an overlapping of the bursts transmitted by different users to a single BS. The TA is expressed as a natural number from 0 to 63. Each value determines the distance from the BS up to 34 km with a 550 m width step [1].

D. Received Signal Level Technique

The Received Signal Level (RxLev) represents the power of the BCCH received by the MS from a network [3]. The RxLev value depends on length and conditions of a propagation channel. The free space path loss formula (1) [10] is

$$PL_{FS} = \left(\frac{4\pi df}{c}\right)^{2} , \qquad (1)$$

where d is the propagation distance in [m], f is the signal frequency in [Hz], c is the speed of light ($3\cdot 10^{8}$ m/s). Figure 2 describes the basic principle of triangulation technique used for the MS position estimation. Character of a propagation environment influences path losses, therefore using the feasible propagation model is mandatory.

IV. Measurement

We designed the Structured Query Language (SQL) database containing position data and identification of BSs (latitude, longitude, CI, LAC, MNC and MCC). Next, we performed measurement in the real network in Brno city with the combined GSM/GPS module [3]. Measured BSs transmitted in the 900 MHz GSM frequency band [1]. Collected service information and position data were sent via GPRS data service on the http server. Next, the PHP script running on this server stored and processed incoming data in to the SQL database. Incoming data were CI, LAC, MNC, MCC, TA, RxLev, Absolute Radio Frequency Channel Number (ARFCN)) for each connected BS, and GPS latitude and longitude for each measuring point. Next, we determined the propagation distance for every connected BS. The path losses depend on a propagation distance and is modeled by propagation models [1][4][5][6]. Figure 3 shows path loss measurement.

Based on the measured parameters of the network, two localization techniques were developed and simulation results were compared.

V. Channel Modelling

Channel modeling is a complex process influenced by individual propagation conditions. We have modeled path losses by some widely used path loss models (COST 231 [4], ECC 131 [6], WINNER II [5]) and adjusted their parameters to fit our measurement [8]. According to the results [8] we propose the adaptable optimization technique [9]. It uses the multislope modeling approach [7][9] and describes the log-distance dependency of path loses (2) [10] as

$$PL_{LD}(d) = PL_{d0} + 10\cdot n\cdot \log_{10}\left(\frac{d}{d_0}\right) , \qquad (2)$$

where n is the path loss exponent setting the slope of the model ($n_{\mathrm{FREE\ SPACE}} = 2$), d is the propagation distance, d_0 is the reference distance (typically $d_0 = 1$ m) and PL_{d0} is the frequency dependent parameter describing the free space path loss (1) at the reference distance $d_0 = 1$ m.

We enhanced the log-distance model with a multislope adaptation as described in [9]. We have optimized the break point positions of a multislope model. According to the MSE estimation, we adjusted the path loss model to fit with the measurement (Figure 4). Our adaptable modeling algorithm uses PSO (Particle Swarm Optimization) to adapt the position of break points (blue and red triangles in Figure 4). The first break point bp0 is static and its position is determined as the free space path loss (1) at a distance of 40 m. This distance represents correction of the BS height

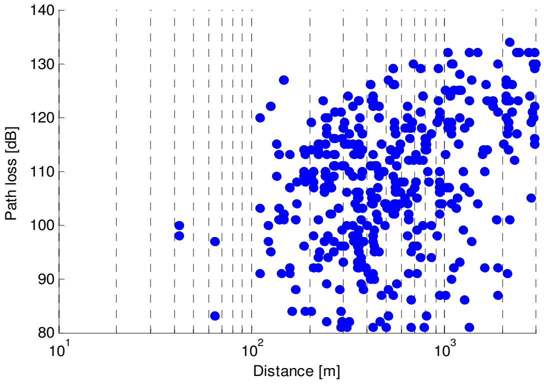


Figure 3. Measured path losses for an outdoor urban scenario in a real GSM network. The distance is determined according to the GPS coordiantes of MS.

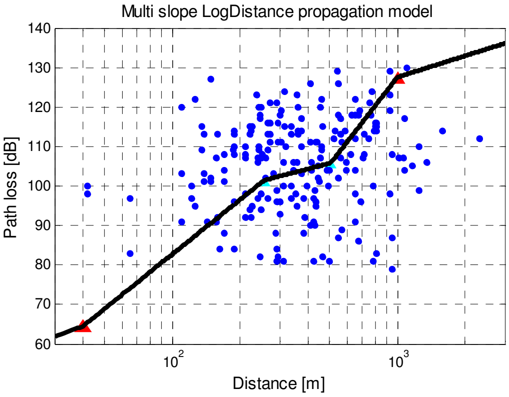


Figure 4. Optimized multislope log-distance model with the deviation error of 12 dB. Break point positions are bp0=[40 m; 64 dB], bp1=[253 m; 100 dB] , bp2=[542 m; 104 dB] , bp3=[1000 m; 127 dB].

for macro cell in an urban area. The PSO algorithm estimates the position of the other three break points (bp1, bp2, bp3) in the range from 40 m to 1000 m The algorithm changes the position (distance and path loss) of the breakpoints and model path losses according to (3) [9].

Path losses of the multislope model are described by (3) for the distances d over the last breakpoint as

$$L_{MS}(d) = PL_{d0} + 10 \cdot n_1 \cdot \log_{10}\left(\frac{h_1}{d_0}\right) + \sum_{i=2}^{n_{BP}} 10 \cdot n_i \cdot \log_{10}\left(\frac{h_i}{h_{i-1}}\right) + 10 \cdot n_{n_{BP}+1} \cdot \log_{10}\left(\frac{d}{h_{n_{BP}}}\right), \qquad (3)$$

where the first part is similar to (2) and describes propagation losses up to the first breakpoint, h_1 is the distance of the first breakpoint, d_0= 1 m is the reference distance. The summation in the second part of (3) describes the propagation losses between the first (bp1) and the last (bp4) breakpoint (the red triangles in Figure 4), n_{BP} is the total number of break points and h_i is the distance of the *i-th* breakpoint. The last part of (3) describes the path losses in the distances d over the last breakpoint h_{nBP-1}.

To estimate the final path loss value in a particular propagation distance it is necessary to determine the correct value of n_{BP} (representing the number of breakpoints) as the number of the last breakpoint previous to the desired distance.

We use the model shown in Figure 3 with parameters: bp0= [40 m; 64 dB], bp1= [253 m; 100 dB], bp2= [542 m; 104 dB], p3= [1000 m; 127 dB]. The standard deviation is 12dB.

VI. LOCALIZATION

Measured signal strength is the initial parameter for the localization. Channel models (COST 231, ECC 131, our optimized multislope log-distance model) were applied to determine the propagation distance. Unfortunately, none of the tested models are precise enough to estimate the exact position applying simple triangulation. Areas like parks, squares, wide streets, and crossroads cause spatial ambiguity and an inaccuracy of the propagation model. Moreover, the relative MS position, reflections and interferences (co-channel, adjacent channel, intersystem) could cause degradation of the measured RxLev value.

The visual presentation is performed in the UTM coordinate system. Displayed by blue circles and circular arcs in Figure 5 and Figure 6 represent the propagation distance In Figure 6 the sectors of the cells are considered.

A. *Geometric Localization Technique*

This technique is possible to apply only in case that the MS have connection with at least three BSs. The triangular constellation of connected BSs is mandatory. The best results were achieved with the constellation conformable to an equilateral triangle (Figure 5).

The basic principle is to link the neighboring BS with a line to create the triangle (red lines in Figure 5). Next, divide those lines according to the ratio of the RxLev value. The perpendicular line (green lines) is led through this dividing point of each side of triangle. Intersections of the green lines create a small triangle. The space limited by this triangle defines the possible MS position. Then the final MS position is estimated in the center of the small triangle (the upper red cross on left). The second red cross on the right is the triangle's centroid. The real GPS position of MS is marked with the green cross.

The localization error for the case in Figure 5 is 165 m and the typical error of this technique was around 300m in an urban environment.

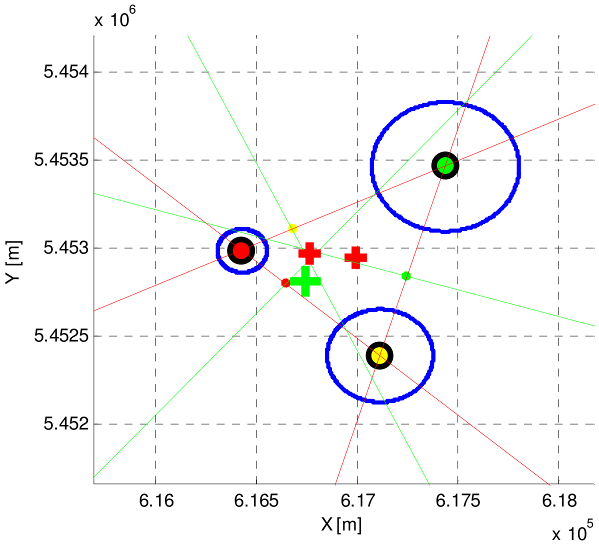


Figure 5. Geometric localization technique (localization error is 165 m). Red lines link BSs and divide them according to the ration of received RxLev. The upper red cross on the left points estimated position, the red cross on the right points the triangle centriod and the green cross points the GPS position.

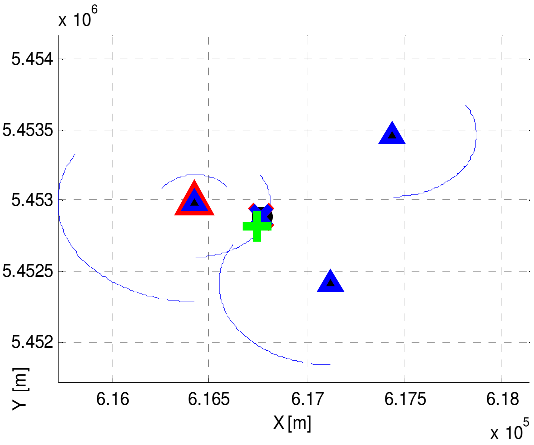


Figure 6. Localization technique with optimization algorithm (localization error is 65 m). The blue arcs describe the modeled propagation distance of received RxLev in connected sectors of BSs. The green cross points real GPS position and the blue cross point the optimized position.

B. *Optimization Localization Techniq*

This localization technique uses PSO (Particle Swarm Optimization) algorithm to estimate the MS position. In dedicated sectors of the linked BSs (blue triangles in Figure 6) the optimized channel model [9] determines the propagation distance (blue circular arcs in Figure 6).

The TA value sets boundaries of the searched space. Twelve PSO agents move inside the defined space with the global scaling factor g= 2.49 and the personal scaling factor p= 1.5. The optimization algorithm has 25 iteration loops. In each loop, the main criteria function computes the criteria value *K* (4) for each of the twelve agents and store one with the minimal value. The stored value is compared with the values received in next loop. A_i represents the distance between the agent position and the modeled propagation distance (the blue circular arcs in Figure 6). A_i is weighted by W_i according to the modeled propagation distance R_i (the higher value of R_i, the higher value of weight W_i). The criteria function *K* for a single agent is described as

$$K = \sum_{i}^{n_{BS}} \left(W_i \cdot A_i \right) = \sum_{i}^{n_{BS}} \left(\left(1 + \frac{1}{(1000/R_i)} \right) \cdot A_i \right), \tag{4}$$

where R_i [m] is the modeled propagation distance between the BS and the MS (blue arcs in Figure 6) and A_i [m] is the Euclidean distance between the agent's position and the modeled propagation distance R_i. The n_{BS} is the number of connected BSs. W_i is the weight describing the dependency on the modeled propagation distance R_i.

C. *Other Localization Approaches*

GSM localization based on measurement of the received signal strength use a fingerprinting approach [11][12]. This approach compares measured or modeled pattern of signal level in the desired area with an actual received value. Instead of determining the MS-BS distance, how it is performed in our approach. Additional improvements of fingerprinting technique are reported in [11] [13]. Achieved localization error was around hundreds of meters. The WLAN (Wireless Local Area Network) localization proposed in [12] has an error of around a few meters.

The localization error (in range of hundreds of meters) of the GSM techniques described in [11][13] is comparable with results obtained by our optimization technique.

VII. CONCLUSION AND FUTURE WORK

We presented the capabilities of localization in cellular wireless networks. We performed measurements of RxLev in the real GSM network in Brno. The post processing of the measured data predicts the propagation distance according to the applied channel model. Propagation models are not capable of involving every individual propagation scenario. Therefore, additional processing is performed to reduce ambiguity. Two localization methods are described and compared.

The first approach is a simple geometric technique. The position is estimated as the ratio of received power from three BSs in the triangle constellation (Figure 5). The positioning error is around 300 m, but it is very strongly dependent on the triangular constellation of BSs. The second technique uses the PSO algorithm (Figure 6). The number and the relative constellation of connected BSs influence the precision of localization. In some cases, the localization error was in ones of meters. Average error was in tens of meters for scenarios with at least three BSs connected.

We created the database of BSs and stored measured and processed data. We use the CI and the RxLev localization technique. For path loss description, we use the multislope propagation model with optimization adaptation [9]. The mean localization error achieved by the geometric technique was around 300 m, the mean error achieved by the PSO technique was around 80 m. We proved the capabilities of the PSO technique in localization. The results were comparable with approaches presented in [11] [13].

We will focus our further work on a comparison of propagation models for serving and neighboring BSs and improving the optimization algorithm to adapt the size of the searched space. The fingerprinting technique will be considered and involved in the ongoing approach.

Acknowledgment

This paper was jointly supported by the Czech Ministry of Education, Youth and Sports through the research program of Brno University of Technology, Electronic Communication Systems and New Generation Technology (ELKOM) MSM0021630513, internal project FEKT-S-11-12 MOBYS and through the project Systems of Wireless Internet Communication (SYWIC) LD11081 in frame of COST IC 0906. The support of the project CZ.1.07/2.3.00/20.0007 WICOMT, financed from the operational program Education for competitiveness, is gratefully acknowledged.

References

[1] T. Halonen, J. Romero, and J. Melero, GSM, GPRS and EDGE Performance : Evolution Towards 3G/UMTS, 2nd ed. England: John Wiley & Sons, Ltd, 2003.

[2] E. D. Kaplan, Understanding GPS: Principles and applications. Norwood: Artech house inc., 1996.

[3] XT65/XT75 - Hardware Interface Description. Siemens, Cinterion, Germany, January, 2007.

[4] E. Damosso and L.M. Correira, Eds. Digital Mobile Radio Towards Future Generation Systems Communications. COST 231 Final Report, Belgium, November, 1999, [Online] Available: http://www.lx.it.pt/cost231.

[5] P. Kyosti, *et al.*, "WINNER II Channel Models," European Commision, IST-WINNER II D1.1.2 V1.2, February, 2008, [Online] Available: http://www.ist-winner.org.

[6] ECC Report 131, June, 2009, [Online] Available: http://www.erodocdb.dk/Docs/doc98/official/pdf/ECCREP131.PDF.

[7] SEAMCAT user manual, European Communications Office, May, 2011, [Online] Available: http://www.seamcat.org.

[8] L. Klozar and J. Prokopec, "Propagation Path Loss Models for Mobile Communication," Proc. of 21st International Conference Radioelektronika 2011, Brno, 2011, pp. 287-290.

[9] L. Klozar, J. Prokopec, and O. Kaller, "Multislope Channel Model Optimization Processing". In 19th Proceedings of Technical Computing Prague 2011. Prague, 2011, pp. 66-66.

[10] T. S. Rappaport, Wireless Communications: Principles and Practice (2nd Edition). USA: Prentice Hall, 2002.

[11] A. Arya, P. Godlewski, and P. Melle, "Performance Analysis of Outdoor Localization Systems Based on RSS Fingerprinting", In Proc. of the 6th International Symposium on Wireless Communication Systems, (ISWCS 2009), pp. 378-382, Tuscany, September 2009.

[12] G., Fuqiang, S. Jianga, and Y. Guizhou, "An Improved Fingerprinting method for localization WLAN-based", In Proc. of International Conference Computer Science and Service System, (CSSS 2011), pp. 2051-2054, Nanjing, June 2011.

[13] M. Ibrahim and M. Youssuef, "CellSence: A Probabilistic RSSI- based GSM Positioning System", In Proc. of the Global Telecommunications Conference, (GLOBECOM 2010), pp. 1-5, Cairo, Egypt, January 2010.

Stereo Video Disparity Estimation Using Multi-wavelets

[1]Pooneh Bagheri Zadeh and [2]Cristian V. Serdean
Department of Engineering, Faculty of Technology,
De Montfort University
Leicester, UK
E-mail: [[1]pbz, [2]cvs]@dmu.ac.uk

***Abstract*—Disparity estimation in stereo video processing is a crucial step in the generation of a 3D view of a scene. In this paper, a multi-wavelet based stereo correspondence matching technique for video is proposed. A multi-wavelet transform is first applied to a pair of stereo frames. Correspondence matching is initially performed at the coarsest level and relies on coarse-to-fine refinement in order to reduce the overall computational costs. Correspondence matching is carried out using a global error energy minimization technique to generate a disparity map for each of the four multiwavelet basebands of the stereo pair. Information in the resulting disparity maps is then combined using an interpolation operator to construct an initial disparity map. The information in the initial disparity map is then progressively propagated to higher resolution levels, on a coarse-to-fine basis, leading to a dense disparity map. Experimental results were generated using two sets of wide baseline convergent multi-view test videos: Breakdancers and Ballet. Results show that multi-wavelets can be a serious contender to scalar wavelets, producing smoother disparity maps with less mismatch errors compared to applying the same global error energy minimization algorithm in the wavelet domain.**

Keywords- Multi-wavelets, Correspondence matching, Disparity estimation, Stereo video.

I. Introduction

Recent years have seen significant advances in multimedia technologies with stereo video and 3D-TV equipment becoming a familiar consumer presence. In stereo video, three-dimensional perception is achieved by simultaneously providing two views of a scene captured by a stereoscopic camera to each corresponding eye, typically with the aid of a pair of active glasses. The brain will then process this stereo information and based on the disparities between the corresponding elements of the two scenes 'convert' it into a meaningful 3D internal representation. The key and the most complex operation in any stereoscopic to real 3D video system is disparity estimation, which needs to accurately find the correspondence points between the two stereo pairs and generate a disparity map for each frame pair. Using these disparity vectors in conjunction with the relevant camera parameters allows one to reconstruct a 3D model of a scene via conventional triangulation techniques.

Many stereo correspondence matching algorithms have been proposed over time, from feature based algorithms, to block-matching, pel-recursive, optical flow, and Bayesian-based approaches. Chien et al. [1] proposed a disparity estimation algorithm for mesh-based stereo images and video featuring a two-stage hybrid approach. In the first stage, an initial disparity map is generated using an iterative block matching algorithm. In the second stage, an iterative octagonal matching algorithm is employed to refine the disparity vectors. Another disparity estimation algorithm for stereo video based on epipolar geometry was reported by Lu et al. [2]. Their theoretical analysis and experimental results showed that their algorithm greatly reduce the search cost, while effectively tracking large and irregular disparities and being less sensitive to epipolar geometry estimation noise. Fan et al. [3] presented a disparity estimation algorithm for stereo video based on edge detection. This algorithm employs the characteristics of human visual system to reduce distortion around the edge regions. They report significant improvement in disparity estimation compared to other state of art techniques. Another disparity estimation technique for stereo video was proposed by Zhu et al. [4], which employed both spatio-temporal correlation and temporal variation of disparity field techniques. By using this technique, they achieved an important reduction in computational complexity compared to full search algorithms.

Over the past years much research has been done to improve the performance of correspondence matching techniques, as well as reducing the computational cost of the search for the best match. Multi-resolution based stereo matching algorithms have received much attentions due to the hierarchical and scale-space localization properties of the wavelets [5],[6]. Correspondence matching can be performed hierarchically, leading to lower computational costs. Yongdong and Guiling [7] proposed a hierarchical multi-resolution based block matching technique for disparity estimation in stereo video. They reported significant improvement in the smoothness of disparity field as well as a reduction in the computational load. Sarkar and Bansal [6] presented a multi-resolution based correspondence matching technique using a mutual information algorithm. They showed that such technique can produce significantly more accurate matching results compared to conventional correlation based algorithms at lower computational costs.

Multi-wavelets offer a number of desirable properties compared to scalar wavelets such as their ability to possess orthogonality, symmetry and high orders of approximation all at once [8]. These properties could increase the accuracy of correspondence matching techniques while still exploiting

ISBN: 978-1-61208-193-9

their hierarchical nature in order to reduce the overall complexity of the correspondence matching algorithms via coarse-to-fine refinement. Bhatti and Nahavandi [9] proposed a multiwavelet based stereo correspondence matching algorithm which makes use of the wavelet transform modulus maxima to generate a disparity map at the coarsest level. This is then followed by a coarse-to-fine strategy to refine the disparity map up to the finest level. Bagheri Zadeh and Serdean [10] provided an evaluation on different types and families of multiwavelets in stereo correspondence matching. They developed an algorithm based on normalized cross correlation. To generate a dense disparity map from the four basebands, a shuffling technique was used in case of balanced multiwavelets and a Fuzzy algorithm was employed in the case of unbalanced multiwavelets. Results showed that the unbalanced multiwavelets produced a smoother disparity map with less mismatch errors compared to balanced multiwavelets.

This paper presents a multi-wavelet based stereo matching algorithm for video which employs a global error energy minimization technique. A multi-wavelet transform is first applied to the input stereo pair to decompose them into a number of subbands. The global error energy minimization algorithm is then employed to generate a disparity map using the coarse subbands. A median operator is then used to combine the disparity maps and generate an initial disparity map. The estimated disparity map is then refined at higher resolution levels, taking advantage of the hierarchical, multi-resolution nature of the multiwavelets to efficiently generate a more accurate final disparity map.

The paper is organized as it follows. Section II presents a brief review of the multi-wavelet transform. The proposed stereo matching technique is discussed in Section III. Experimental results are presented in Section IV while Section V is dedicated to the conclusions.

II. MULTIWAVELET TRANSFORM

Multi-wavelet transforms are similar to scalar wavelet transforms with some key differences. Classical wavelet theory is based on the refinement equations as given below:

$$\begin{aligned}\phi(t) &= \sum_{k=-\infty}^{k=\infty} h_k \phi(m\,t-k)\\ \psi(t) &= \sum_{k=-\infty}^{k=\infty} g_k \psi(m\;\,t-k)\end{aligned} \tag{1}$$

where $\phi(t)$ is a scaling function, $\psi(t)$ is a wavelet function, h_k and g_k are scalar filters and m represents the band number.

In contrast to wavelet transforms, multi-wavelets have two or more scaling and wavelet functions. Scalar wavelets have multiplicity $r = 1$, while multi-wavelets support $r \geq 2$. To date, most multiwavelets have a multiplicity factor of $r = 2$.

The set of scaling and wavelet functions of a multi-wavelet in vector notation can be defined as:

$$\begin{aligned}\Phi(t) &\equiv [\phi_1(t) \quad \phi_2(t) \quad \phi_3(t) \quad \ldots \quad \phi_r(t)]^T\\ \Psi(t) &\equiv [\psi_1(t) \quad \psi_2(t) \quad \psi_3(t) \quad \ldots \quad \psi_r(t)]^T\end{aligned} \tag{2}$$

where $\Phi(t)$ and $\Psi(t)$ represent the multi-scaling and respectively multi-wavelet functions, with r scaling- and wavelet functions. A multi-wavelet with two scaling and wavelet functions can be denoted as [11]:

$$\begin{aligned}\Phi(t) &= \sqrt{2}\sum_{k=-\infty}^{k=\infty} H_k \Phi(m\,t-k)\\ \Psi(t) &= \sqrt{2}\sum_{k=-\infty}^{k=\infty} G_k \Psi(m\,t-k)\end{aligned} \tag{3}$$

where H_k and G_k are $r \times r$ matrix filters and m is the subband number.

Unlike scalar wavelets, multi-wavelets can offer symmetry, orthogonality and approximation orders higher than 1 simultaneously. Similar to wavelet transforms, multi-wavelets can be implemented using Mallat's filter bank theory [5]. A 2D multi-wavelet transform with multiplicity two will produce sixteen subbands: four basebands and twelve high frequency subbands. A visual comparison of the resulting subbands of a 2D wavelet (Antonini 9/7) and respectively 2D multi-wavelet (bat01) is shown in Figure 1.

III. MULTI-WAVELET IN STEREO VIDEO CORRESPONDENCE MATCHING

A block diagram of the multi-wavelet based stereo correspondence matching technique for stereo video based on a global error energy minimization algorithm is shown in Figure 2. A stereoscopic video needs to be input to the system. For the purpose of this paper, two camera views from the multi-view sequences, Breakdancers and Ballet (generated by Microsoft laboratories using eight synchronized PtGrey color cameras) are chosen [12]. As these datasets were captured using convergent cameras, each frame pair needs to be rectified to suppress the vertical displacement. The epipolar rectification algorithm proposed by Fusiello and Irsara [13] has been used in this work to rectify each fame pair of the video input. A multi-wavelet transform is then applied to each rectified frame pair to decompose them into multi-wavelet subbands. The search for the best correspondence points starts at the coarsest level. The corresponding basebands in the two frames are passed to a regional based stereo matching block. The matching algorithm uses a global energy minimization technique [14] to generate a disparity map between the two input subbands.
This global error energy minimization technique is briefly described in Section III.A. The output of the matching process is four disparity maps. These maps are then combined using a median operator to generate an initial disparity map. As the initial disparity map is estimated at the lowest resolution, the information needs to be progressively passed on to higher resolution levels. For this refinement

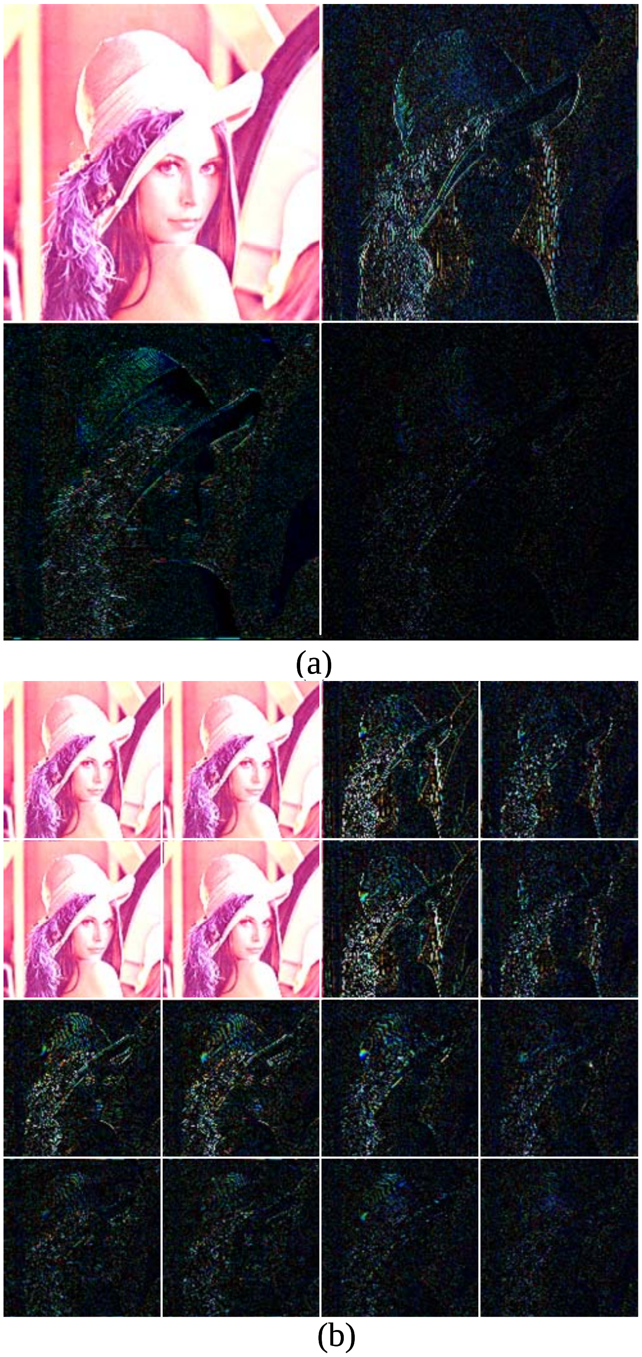


Figure 1. Single level decomposition of Lena test image (a) Antonini 9/7 wavelet transform (b) bat01 multi-wavelet transform.

process, the algorithm presented in [6] is used to propagate information in the coarsest level to the higher resolutions. Finally a median filter is applied to the last processed disparity map to further smooth the final disparity map.

A. Global Error Energy Minimization technique

The Global Error Energy Minimization (GEEM) technique [14] calculates a disparity vector for each pixel. It searches for the best match for each pixel in the correspondence search area of the other image using error minimization criterion. For RGB images, the error energy criterion can be defined as:

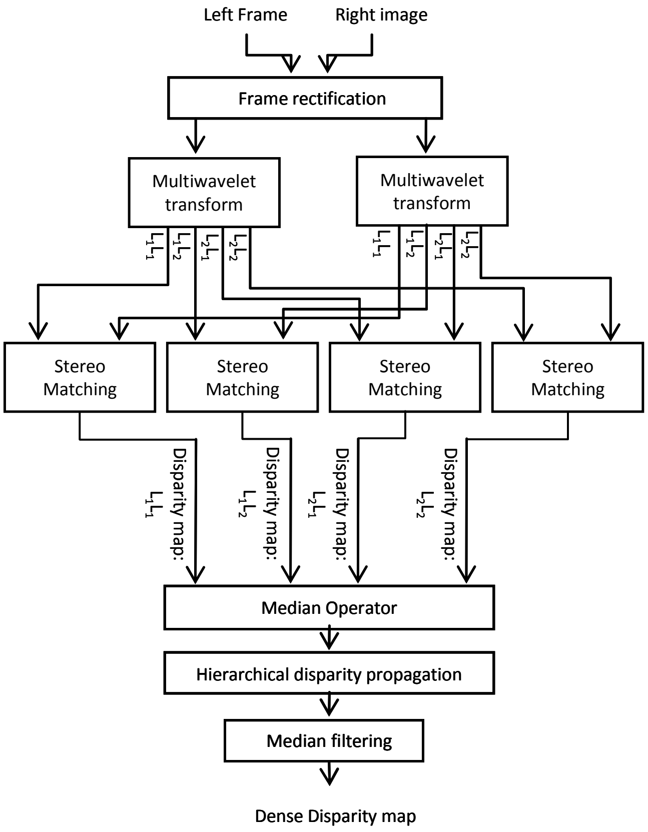
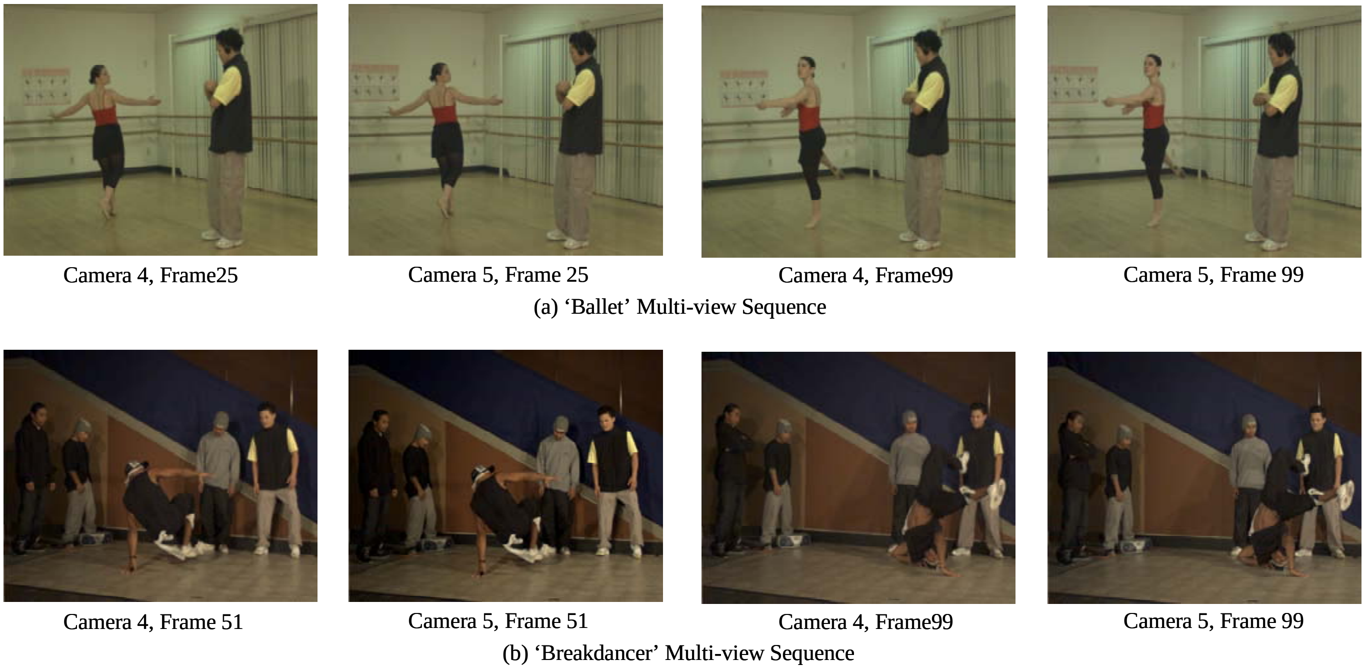


Figure 2. Block diagram of multi-wavelet based stereo matching technique using the global energy minimization algorithm.

$$\begin{aligned} Er_{en}(i,j,w_x,w_y) \; &= \; \frac{1}{3} \sum_{k=1}^{3} (\, I_1(i+w_x, j+w_y, k) - I_2(i,j,k)\,)^2 \\ -d_x \le w_x \le d_x \quad & and \quad -d_y \le w_y \le d_y \\ i \; = \; 1, \; \ldots \; m \quad & and \quad j \; = \; 1, \; \cdots \; n \end{aligned} \tag{4}$$

where I_1 and I_2 are the two input frames, $Er_{en}(i,j,w_x,w_y)$ is the difference energy of the pixel $I_2(i,j)$ and pixel $I_1(i+w_x, j+w_y)$, d_x is the maximum displacement around the pixel in the x direction, d_y is maximum displacement around the pixel in the y direction and m and n are the image size.

In order the GEEM algorithm to determine the disparity vector for each pixel in the current view, it first calculates Er_{en} of each pixel with all the pixels in its search area in the corresponding frame. For every disparity vector (w_x, w_y) in the disparity search area, error energy is calculated using Equation 4 and placed into a matrix. Each of the resulting error energy matrices is first filtered using an average filter


Camera 4, Frame25

Camera 5, Frame 25

Camera 4, Frame99

Camera 5, Frame 99

(a) 'Ballet' Multi-view Sequence

Camera 4, Frame 51

Camera 5, Frame 51

Camera 4, Frame99

Camera 5, Frame 99

(b) 'Breakdancer' Multi-view Sequence

Figure 3. Two views of Multi-view test sequences, (a) 'Ballet' and (b) 'Breakdancers.

to decrease the number of incorrect matches [15]. The disparity index of each pixel is then determined by finding the disparity index of the matrix which contains the minimum error energy for that pixel. In order to increase the reliability of the disparity vectors around the object boundaries, which is the result of object occlusion in images, the generated disparity map undergoes a thresholding procedure as it follows:

$$\tilde{d}(i,j) = \begin{cases} d(i,j) & Er_{en}(i,j) \le \alpha \times Mean(Er_{en}) \\ 0 & Er_{en}(i,j) > \alpha \times Mean(Er_{en}) \end{cases} \qquad (5)$$

where $\tilde{d}(i,j)$ is the processed disparity map, $d(i,j)$ is the disparity map, α is a tolerance reliability factor, $Er_{en}(i,j)$ is the minimum error energy of the pixel (i,j) calculated and selected in the previous stage. Finally a median filter is applied to the processed disparity map, $\tilde{d}(i,j)$ to further smooth the final disparity map.

IV. SIMULATION RESULTS

In order to evaluate the performance of the proposed multi-wavelet technique compared to a similar wavelet based stereo matching technique which employs the same global energy minimization algorithm, both methods are benchmarked using two multi-view sequences, Breakdancers and Ballet [12]. Figure 3 shows frames 25 and 99 of the camera 4 and camera 5 views of Ballet sequence and respectively frames 51 and 99 of the two views (camera 4 and camera 5 views) of Breakdancers video. The resulting disparity maps for the two methods using the GHM unbalanced multi-wavelet and respectively the Antonini 9/7 scalar wavelet for the Ballet sequence (frame 25 and 99 from camera views of 4 and 5) and Breakdancers sequence (frame 51 and 99 from camera views of 4 and 5) are illustrated in Figures 4(a) and 4(b). In these figures areas with intensity 0 represent unreliable disparities. From Figure 4, it is obvious that the disparity map produced by the multi-wavelet based algorithm is more accurate and smoother than that of the wavelet based technique. The different spectral content of the multiwavelet subbands and the greater subband structure flexibility afforded by multi-wavelets enable the global energy minimization algorithm to generate more reliable matches from the multi-wavelet decomposition than from the scalar wavelet decomposition.

V. CONCLUSION

This paper presented a multi-wavelet based stereo correspondence matching technique for video using a global error energy minimization algorithm. A multi-wavelet transform with multiplicity of two decomposes the input rectified frame pairs into four baseband and twelve high frequency subbands. The resulting four basebands of the two views were then employed to generate four disparity maps using the global error energy minimization algorithm. The resulting four disparity maps were then combined using a median operator to generate the initial disparity map, which was then refined by hierarchically propagating it to the finer levels. Results show that multi-wavelets can be a serious

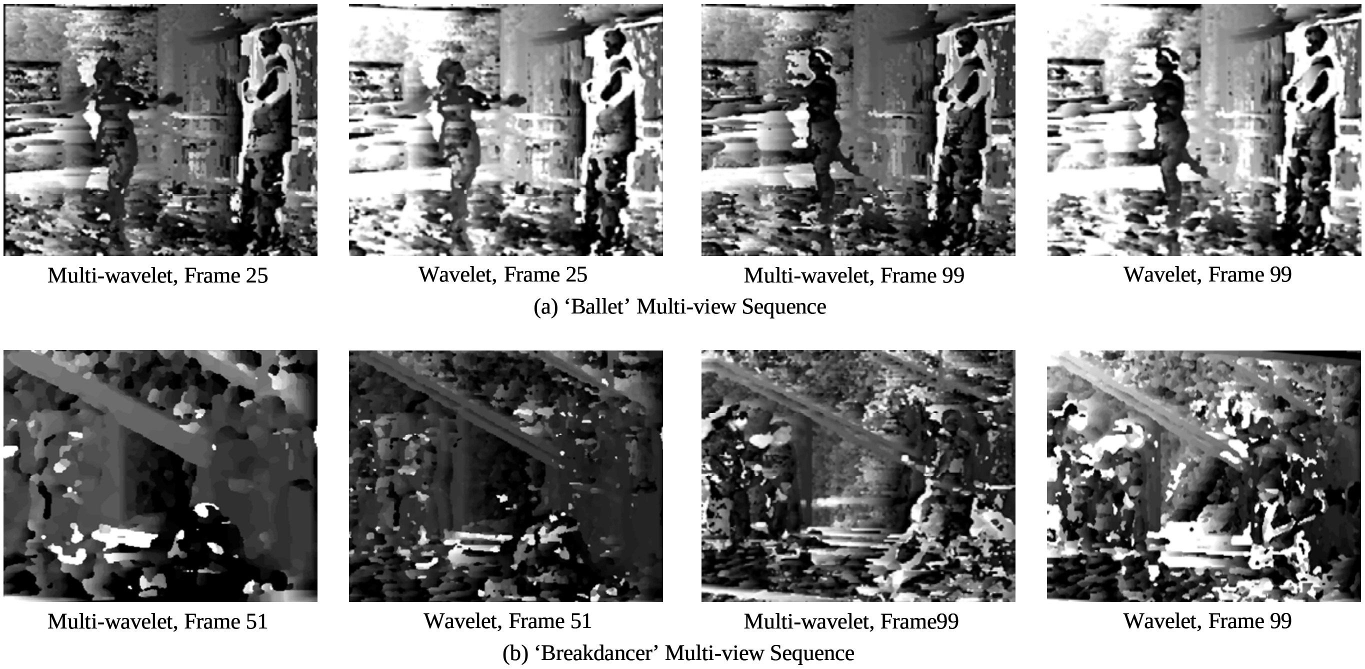


Figure 4. Disparity maps generated using wavelet based algorithm and the multi-wavelet based algorithm for (a) 'Ballet' multi-view sequence and (b) 'Breakdancers' multiview sequence.

contender to scalar wavelets, producing smoother disparity maps with less mismatch errors compared to applying the same global error energy minimization algorithm in the wavelet domain.

ACKNOWLEDGMENT

This work was supported by EPSRC under Grant number: EP/G029423/1.

REFERENCES

[1] S. Chien, S. Yu, L. Ding, Y. Huang, and L. Chen, "Fast disparity estimation algorithm for mesh-based stereo image/video compression with two-stage hybrid approach," Proceedings of SPIE, Vol. 5150, pp. 1521-1530, 2003.

[2] J. Lu, H. Cai, J. G. Lou and J. Li, "An epipolar geometry-based fast disparity estimation algorithm for multiview image and video coding," IEEE Transaction on Circuits System for Video Technology, vol. 17, no. 6, pp.737–750, June 2007.

[3] J. Fan, F. Liu, W. Bao and H. Xia, "Disparity Estimation Algorithm for Stereo Video Coding Based on Edge Detecetion," International Conference on Wireless Communications & Signal Processing, pp. 1-5 , November 2009.

[4] W. Zhu, X. Tian, F. Zhou and Y. Chen, "Fast Disparity Estimation Using Spatio-temporal Correlation of Disparity Field for Multiview Video Coding," IEEE Transactions on Consumer Electronics, vol. 56, no. 2, pp. 957-964, 2010.

[5] S. Mallat, A Wavelet Tour of Signal Processing, Academic Press, 1999.

[6] I. Sarkar, M. Bansal, "A wavelet-based multiresolution approch to solve the stereo correspondence problem using mutual information," IEEE Transaction on system, man, and cybernetics, vol. 37, pp. 1009-1014, Augeust 2007.

[7] Z. Yongdong and L. Guiling, "The research of disparity estimation algorithms in stereo video coding," Journal of Electronic Measurement and Instrumentation, January 2002.

[8] V. Strela and A.T. Walden, "Signal and image denoising via wavelet thresholding: orthogonal and biorthogonal, scalar and multiple wavelet transforms," In Nonlinear and Nonstationary Signal Processing, pp. 124-157, 1998.

[9] A. Bhatti and S. Nahvandi, "Depth estimation using multi-wavelet analysis based stereo vision approach," International Journal of Wavelets, Multiresolution and Information Processing, vol. 6, pp. 481-497, 2008.

[10] P. Bagheri Zadeh and C. Serdean, "An Evaluation of Multiwavelet Families For Stereo Correspondence Matching", The sixth International Conference on Digital Telecommunications (ICDT2011), Budapest, Hungary, pp. 41-45, June 2010.

[11] V. Strela, "Multiwavelets: threory and applications," PhD thesis, MIT, 1996.

[12] C.L. Zitnick, S.B. Kang, M. Uyttendaele, S. Winder, and R. Szeliski, "High-quality video view interpolation using a layered representation," *ACM SIGGRAPH and ACM Trans. on Graphics*, Los Angeles, CA, pp. 600-608, Aug. 2004.

[13] A. Fusiello and L. Irsara, "Quasi-euclidean Uncalibrated Epipolar Rectification," International Conference on Pattern Recognition (ICPR), 2008, Tampare, Finland, 2008.

[14] B. B. Alagoz, "Obtaning depth maps from colour images by region based stereo matching algorithms," OncuBilim Algorithm and System Labs, vol. 08, Art.No:04, 2008.

A LUT Baseband Digital Pre-Distorter For Linearization

[*]Feng Li, [*§]Bruno Feuvrie, [*]Yide Wang, [*§]Anne-Sophie Descamps
[*]*L'UNAM Université - Université de Nantes, UMR CNRS 6164 Institut d'Electronique et de Télécommunications de Rennes (IETR), Ecole polytechnique de l'université de Nantes, Rue Christian Pauc - La Chantrerie BP 50609, 44306 Nantes Cedex 3 France*
[§]*IUT GEII Nantes - Site Fleuriaye, Av. du Prof J. Rouxel, 44475 Carquefou, France*
feng.li@etu.univ-nantes.fr;
Bruno.Feuvrie@univ-nantes.fr;
Yide.Wang@univ-nantes.fr;
Anne-Sophie.Bacquet@univ-nantes.fr

***Abstract*—This paper proposes a Look-Up-Table Digital Pre-Distorter (LUT DPD) for PA linearization with baseband predistorting procedure. Memory effect can be compensated with Memory Polynomial (MP) modelisation for the wideband applications. Measurements are realized on a PA ZFL-2500 driven by a modulated 16QAM signal with 3.84MHz bandwidth at the carrier frequency of 1.8GHz. The proposed solution achieves maximum ACPR reduction of 12.5dB and EVM correction of 3%.**

Keywords-power amplifier; baseband predistortion procedure; nonlinear memory effect; digital predistortion; linearization.

I. Introduction

Modern wireless communication systems aim to provide sevices of high data rates for the applications such as video conference, broadcast TV. Constrained by the limited radio frequence resource, spectrally efficient modulation schemes (OFDM, etc.) are widely used in order to increase system capacity. Unfortunately, the resulted non-constant-envelope signals with high Peak-to-Average Power Ratio (PAPR) become more sensitive to the inherent nonlinearity of Power Amplifier (PA) [3]. Thus, a tradeoff between spectrum efficiency and power efficiency must be made. The memory effect increases the distortions on the output signal. The origins of memory effect can be thermal or electrical. Electrical origin is mainly caused by the mismatch of circuit impendence, due to capacitors and resistances. The thermal origin due to the temperature variation may affect low modulation frequencies up to a few megahertz. Therefore, the nonlinear memory effect is device dependant (bias condition, temperature, et) and signal dependant (signal's PAPR, average power, bandwidth, etc).

Digital Pre-Distortion (DPD) is one of the promising techniques for minimizing these distortions. Its advantages lie in the fact that it is reconfigurable and doesn't require deep knowledge of PA's physic circuits [1]. In order to define the predistorting procedure, we need to find a precise model to exactly describe the nonlinear memory effect behavior of the PA.

Figure 1 illustrates the basic principle of DPD. A digital predistortion circuit is inserted before the PA. Then, the overall system produces linearized characteristic. The objective is to efficiently determine the predistorted signal $e(nT)$, which is also the new input of the PA. With ideal linearization, the output of the DPD-PA cascade $S(nT)$ can be written as:

$$S(nT) = G(F(x(nT))) = G_0 x(nT) \tag{1}$$

where $x(nT)$ is the input signal; G represents the nonlinear behavior of the PA, F the behavior of the DPD, which is also the inversed characteristic of G; G_0 is the desired linear gain.

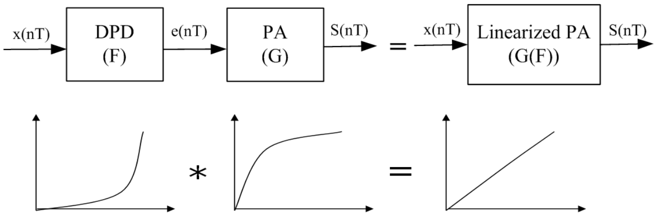


Figure 1. DPD principle

In [5], a baseband DPD based on the Hammerstein model is proposed to linearize the PAs. However this DPD presents two disadvantages. Firstly, the Hammerstein model implicitly separates the memory effect from the nonlinearity. I practice these two effects are often closely related. Secondly, the complex root-finding procedure for finding the amplitude of predistorted signal is not adaptive for the real-time systems. In this paper, we use the MP model to better model the nonlinear memory effect behavior of the PAs and a LUT technique is proposed to avoid the complex root-finding procedure. In section II, the principle of baseband DPD is presented. Based on the baseband signal processing, in section III, the modelisation and the LUT technique are illustrated in detail. In section IV, the measurement is shown. Finally the conclusions and the perspectives are presented.

II. PRINCIPLE OF BASEBAND DPD

Predistorting procedure can be classified into three categories: baseband, Intermediate Frequency (IF) and Radio Frequency (RF). RF predistorter suffers from constrained adaptivity to the variable PA characteristics, due to its high frequency (considered from 800MHz to several GHz for radio communication systems). For IF predistorter, the development of Digital Signal Processing (DSP) still cannot sustain the high sampling frequency to digitize the IF signals. Furthermore, higher power consumption is required for RF/IF predistorters than the baseband predistorters. This affects the flexibility, the size, the mobility, the cost, and the communication quality [2].

Table I
BANDWIDTH FOR DIFFERENT STANDARDS

Standards	Bandwidth
Digital Advanced Mobile Phone System	25kHz
Global System for Mobile Communications	200kHz
Interim Standard 95	1.25MHz
Universal Mobile Telecommunications System	5MHz
Digital Video Broadcasting - Terrestrial	8MHz
3GPP Long Term Evolution and WiMAX	Up to 20MHz
802.11a/g	25MHz
Emerging 4G systems	up to 100MHz

Compared with RF/IF predistorters, baseband DPD presents higher adaptivity to these various parameters. DPD is a DSP-based PD technique. Benefiting from the development of DSP technology, DPD can theoretically to process signal bandwidth greater than 1GHz. But in practice, it is constrained to less than 100MHz [6] for complexity and cost reasons. Baseband predistorting procedure operates on the input signals in baseband frequency, where it is much easier to construct the inversed characteristics [4]. Bandwidths for typical standards are illustrated in Table I. This technique can be applied to different standards and further developed to emerging 4G systems. DPD is added in the stage prior to the DAC and the up-conversion before the PA.

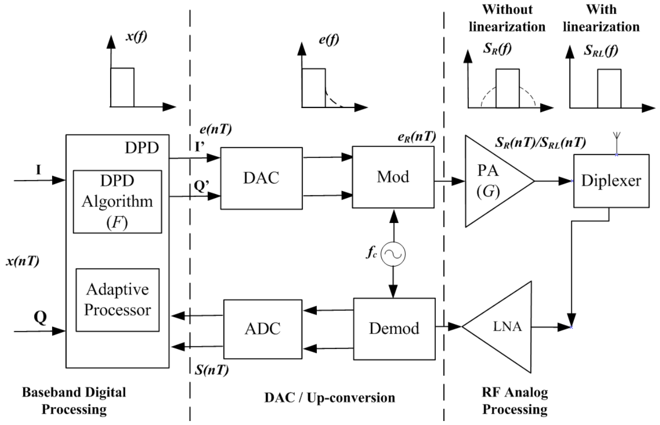


Figure 2. Baseband DPD System Design

Figure 2 presents the typical baseband DPD system design. For the nonlinear system, the gain of the PA G presents device-dependant and signal-dependant characteristics, which can be described as:

$$G = G(B, |x(nT)|, f_c, T \dots) \tag{2}$$

here B is the bandwidth of input signal $x(nT)$, f_c illustrates the carrier frequency, T represents the temperature.

The nonlinear RF output of PA $S_R(nT)$ is down converted to baseband $S(nT)$ to be compared with the baseband input $x(nT)$ and to obtain the inversed nonlinear behavior of PA F:

$$F = \frac{G^{-1}}{G_0} \tag{3}$$

In order to compensate the nonlinear memory effect in wideband systems, the baseband input sample $x(nT)$ is predistorted in DPD (Figure 1):

$$e(nT) = F(x(nT)) \tag{4}$$

The predistorted output signal $e(nT)$ is then directly up converted to RF $e_R(nT)$ (equation (5)) to be amplified by the PA. Finally, the PA's linearized output $S_{RL}(nT)$ (equation (6)) is radiated by the antenna to transmit the messages.

The signal $e_R(nT)$ can be expressed as:

$$\begin{aligned} e_R(nT) =& Re[e(nT)e^{j2\pi f_c nT}] \\ =& I'(nT)\cos(2\pi f_c T) - Q'(nT)\sin(2\pi f_c nT) \end{aligned} \tag{5}$$

The signal $S_{RL}(nT)$ is given by:

$$S_{RL}(nT) = G(e_R(nT)) = G_0 x(nT) e^{j2\pi f_c nT} \tag{6}$$

Note that, with ideal linearization, G_0 is a real constant, presenting no device-dependant or signal-dependant distortions anymore.

As shown in Figure 2, the spectrum of baseband input signal is without spectral regrowth. Due to the nonlinearity or the nonlinear memory effect especially presented by the PA, spectral regrowth appears in the adjacent channels. While with the predistortion, these distortions can be minimized.

III. PROPOSED LUT BASEBAND DPD

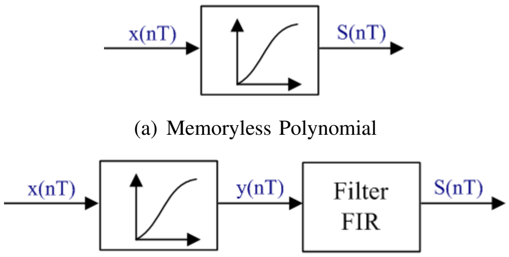
A. PA modelisation

The first step in PA linearization is to provide a mathematical description of PA's nonlinear memory effect behavior.

In the narrowband systems, memoryless polynomial model (Figure 3(a)) exhibits good performances in describing the PA's behavior. With the baseband input sample $x(nT)$, the PA's nonlinear output is:

$$S(nT) = a_1 x(nT) + ... + a_{2j+1} x(nT)|x(nT)|^{2j} + ... \tag{7}$$

where a_{2j+1} $(j = 0, 1, 2 \dots N)$ are the coefficients of the nonlinearity. N represents the order of nonlinearity.



(a) Memoryless Polynomial

(b) Hammerstein model

Figure 3. Model structures

For wideband applications, memory effect should be taken into account. Hammerstein model (Figure 3(b)) consists of a polynomial to model the static nonlinearity and a Finite Impulse Response (FIR) filter to represent the memory effect.

$y(nT)$, the output of the polynomial block is defined as:

$$y(nT) = \sum_{j=0}^{N} b_{2j+1} x(nT)|x(nT)|^{2j} \tag{8}$$

where b_{2j+1} are the coefficients representing nonlinearity.

The output of FIR filter, $S(nT)$, which is also the output of PA is given by:

$$S(nT) = \sum_{i=0}^{P-1} a_i y[(n-i)T] \tag{9}$$

where a_i are the coefficients representing memory effect. P shows the length of memory effect.

The DPD proposed in [5] is based on the Hammerstein model. This model implicitly separates the memory effect from the nonlinearity. However, in practice these two effects are often closely related. In this paper, we use the MP model to better model this nonlinear memory effect behavior. The PA's MP modelisation is given by [7]:

$$S(nT) = \sum_{i=0}^{P-1} \sum_{j=0}^{N} a_{i,2j+1} x[(n-i)T]|x[(n-i)T]|^{2j} \tag{10}$$

where $a_{i,2j+1}$ are the coefficients of the nonlinear memory effect.

B. Look-Up-Table DPD

According to the DPD principle (equation (1)), if the PA is perfectly linearized, we get:

$$\sum_{i=0}^{P-1} \sum_{j=0}^{N} a_{i,2j+1} e[(n-i)T]|e[(n-i)T]|^{2j} = G_0 x(nT) \tag{11}$$

The left member of above equation can be divided into two parts: static nonlinearity $d(nT)$ (equation (12)) and dynamic deviation $P(nT)$ (equation (13)). The first part depends only on the current input at instant nT with $i = 0$. The second part, depending on the previous inputs, is composed of the terms with i varying from 1 to $P - 1$.

$$d(nT) = \sum_{j=0}^{N} a_{0,2j+1} e(nT)|e(nT)|^{2j} \tag{12}$$

$$P(nT) = \sum_{i=1}^{P-1} \sum_{j=0}^{N} a_{i,2j+1} e[(n-i)T]|e[(n-i)T]|^{2j} \tag{13}$$

Seperating the static nonlinearity $d(nT)$ and dynamic deviation $P(nT)$, we get:

$$\sum_{j=0}^{N} a_{0,2j+1} e(nT)|e(nT)|^{2j} = G_0 x(nT) - P(nT) \tag{14}$$

With $P(nT)$ being known at instant nT, the corresponding predistorted signal $e(nT)$ can be found for each $x(nT)$. Taking the absolute value of each side of equation (14), we obtain:

$$|\sum_{j=0}^{N} a_{0,2j+1} e(nT)|e(nT)|^{2j}| = |G_0 x(nT) - P(nT)| \tag{15}$$

In [5], the Hammerstein DPD adopts a complex root-finding procedure to calculate the amplitude of the predistorted signal $|e(nT)|$. Unfortunately, this procedure is too time-consuming to be applicable in the real-time applications. In this paper, the LUT principle [7] is proposed to efficiently find $E(m)$, which is also defined as $|e(nT)|$.

Firstly, we decompose the maximum dynamic range of $|e(nT)|$, function of the input amplitude of saturation point and the maximum magnitude of input signal, into M (table size) intervals of equal length. Each interval corresponds to a quantified value $E(m)$ $(m = 1, 2, M)$. Each $E(m)$ corresponds to a value $f(m)$, according to the left member of equation (15). Thus a LUT (TABLE II) is generated according to the following equation:

$$LUT : f(m) = |\sum_{j=0}^{N} a_{0,2j+1} E(m)^{2j+1}| \tag{16}$$

Table II
LUT

INLUT	OUTLUT
$E(1)$	$f(1)$
...	...
$E(m)$	$f(m)$
...	...
$E(M)$	$f(M)$

For each baseband input sample $x(nT)$, we calculate the right member of equation (15) ($|G_0 x(nT) - P(nT)|$) and compare with the values $f(m)$ in the MP LUT to find the corresponding $E(m)$, which is also the desired $|e(nT)|$.

The corresponding phase $Arg(e(nT))$ is calculated by:

$$Arg(e(nT)) = arg\left\{\frac{G_0x(nT) - P(nT)}{\sum_{j=0}^{N} a_{0,2j+1}|e(nT)|^{2j+1}}\right\} \quad (17)$$

Finally, the predistorted signal $e(nT)$ is given by:

$$e(nT) = |e(nT)|e^{jArg(e(nT))} \quad (18)$$

Table III
LUT DPD ALGORITHM

```
Initialization:
        $n = 0, P(0) = 0$
Generate LUT:
        $f(m) = |\sum_{j=0}^{N} a_{0,2j+1}E(m)^{2j+1}|$
Loop($n = n + 1$)
{
- Calculate:
        $|G_0x(nT) - P(nT)|$
- Compare with values OUTLUT in the table to find the corresponding
$|e(nT)|$ for each x(nT)
- Calculate:
        $Arg(e(nT)) = arg\left\{\frac{G_0x(nT)-P(nT)}{\sum_{j=0}^{N} a_{0,2j+1}|e(nT)|^{2j+1}}\right\}$
- Calculate:
        $e(nT) = |e(nT)|e^{jArg(e(nT))}$
- Calculate:
$P[(n+1)T] = \sum_{i=1}^{P-1}\sum_{j=0}^{N} a_{i,2j+1}e[(n+1-i)T]|e[(n+1-i)T]|^{2j}$
}Goto loop
```

Table III gives the summary of the proposed LUT algorithm.

The proposed MP model based LUT DPD exhibits lower complexity than the Hammerstein DPD [5]. With the simulation in Matlab, the Hammerstein DPD needs $1.24s$, while only $0.32s$ is required for the LUT DPD with a table size of 1000.

IV. MEASUREMENT

The test bench consists of a Vector Signal Generator (VSG), a Spectrum Analyzer (SA) and a PC. It is designed to be fully automatic by using a Matlab toolbox. The baseband data are generated in Matlab and then sent to the VSG. The VSG (Rhode & Schwartz SMU 200A) receives the complex envelope data via an Ethernet cable (TCP/IP) from the PC and uses a direct up-conversion from baseband to RF. Once the data have been sent to the VSG, the VSG will send the corresponding modulated signal to the PA. The RF input and output signals of the PA are then analyzed by the SA (Agilent E4440A). In this case, the signal analysis software (89601A) provided by this instrument can be used to acquire and demodulate the input and output signals separately. It digitalizes each IF signal by using two ADC ($14bits$), each with a frequency of $100MHz$, totally $200MHz$. These signals are then transferred via an Ethernet cable to the PC, and finally processed in the workspace of Matlab.

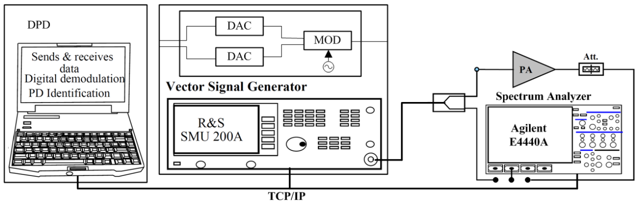
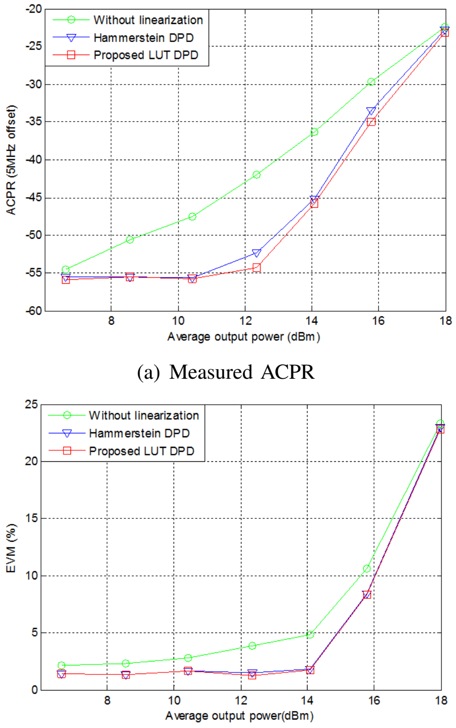


Figure 4. Test Bench

In the PC, the acquired signals by the SA are used to identify the parameters of the PA to obtain the inversed characteristics and then loaded again to the VSG. Finally, the output of the linearized PA is digitized in the SA and sent back to the PC to evaluate the performance of the DPD.

The tested wideband (500-2500 MHz) PA (Mini-Circuits ZFL-2500) has a gain of $28dB(\pm1.5)$ and the 1dB compression point around the output power of $15dBm$. It is saturated at the average output power around $19.7dBm$. The measurement is driven by a 16QAM modulated signal with $3.84MHz$ bandwidth at the carrier frequency of $1.8GHz$. The offset of adjacent channel is set to be $5MHz$. The pulse shaping filters are square-root raised cosine filters with a roll-off factor of 0.35. A sequence of 200 symbols (4000 samples) is sent to the VSG. The table size is 1000 with N of 4 and P of 2.



(a) Measured ACPR

(b) Measured EVM

Figure 5. Measured performances

The measured ACPR performance is presented in Figure

5. The proposed LUT DPD achieves slightly higher ACPR corrections than the Hammerstein DPD, meanwhile with higher simplicity and lower time consumption. The best performance for ACPR reduction is about $12.5dB$ for the proposed LUT DPD and $10dB$ for Hammerstein DPD around the average output power of $12dBm$. For the in-band distortion (EVM), these two DPDs present nearly the same performances with 3% maximum corrections around the average output power of $14dBm$.

V. Conclusion and prospective

In this paper, we proposed a MP model based LUT baseband DPD which presents the ability to linearize the PA with memory effect for the wideband applications. Based on the MP model, a robust modelisation is presented to describe the nonlinear memory effect behavior of the PA. For this DPD, the generated LUT in digital signal processing is used to predistort the baseband signal prior to DAC and up conversion to the desired carrier frequency. The results present maximum ACPR reduction of 12.5dB and EVM correction of 3% with lower complexity compared with the Hammerstein DPD. The LUT is updated continuously so as to enable the DPD to adapt to variations of the transmitter chain characteristics (due to temperature drift, antenna impedances, etc.). We are also planning to consider the nonuniform LUT intervals.

References

[1] F.M. Ghannouchi, and O. Hammi, *Behavioral Modeling and Predistortion*. IEEE Microwave Magazine, Vol. 10, N7, pp. 52-64, Dec 2009.

[2] W.J. Kim, K.J. Cho, S.P. Stapleton, and J.H. Kim, *Baseband Derived RF Digital Predistortion*. Electronics letters, Vol. 42, N8, pp. 468-470, April 2006.

[3] M.A. Hussein, Y. Wang, B. Feuvrie, S. Toutain, and G. Peyresoubes, *Piecewise Complex Circular Approximation of the Inverse Characteristics of Power Amplifiers for Digital Predistortion Techniques*. ICDT ’08, pp. 59-63, Bucharest, June 2008.

[4] E. Cottais, and Y. Wang, *Influence of Instruments Bandwidth in the Power Amplifier Linearization Process*. ICDT ’08, pp. 11-14, Bucharest, June 2008.

[5] E. Cottais, B. Feuvrie, Y. Wang, and S. Toutain, *Experimental results for power amplifier adaptive baseband predistortion linearization*. IEEE Topical Symposium on Power Amplifiers for Wireless Communications, Long Beach, USA, 8-9, January 2007

[6] A. Katz, R. Gray, and R. Dorval, *Truly wideband linearization*. IEEE Microwave Magazine, Vol. 10, N7, pp. 20-27, Dec 2009.

[7] Li F.; Feuvrie B.; Wang Y.; Chen W.; *MP/LUT baseband digital predistorter for wideband linearization*. Electronics Letters, Vol. 47, N19, pp.1096-1098, September 2011.

ISBN: 978-1-61208-193-9

Multiplatform Approaches and Tools for Parallel Computing in Signal Processing Domain

Tomas Fryza, Jitka Svobodova, Roman Marsalek, Jan Prokopec
Department of Radio Electronics
Brno University of Technology
Brno, Czech Republic
{fryza,marsaler,prokopec}@feec.vutbr.cz, xsvobo61@stud.feec.vutbr.cz

***Abstract*—The paper deals with various approaches used for parallel computing in signal processing domain. More precisely, the methods exploiting the multicores Central Processing Units such as Message Passing Interface and OpenMP are taking into account. The properties of the programming methods are experimentally proved in application of two-dimensional Fast Fourier Transform and Discrete Cosine Transform and are compared with possibilities of MATLAB built-in functions and Digital Signal Processors with Very Long Instruction Word architecture. The optimal combination of computing methods in signal processing domain is proposed. Results in the paper prove the possibility of creation of a heterogeneous computing system compounded of CPU and DSP architectures.**

Keywords-transform coding; parallel computing; MPI; OpenMP; DSP.

I. Introduction

There are several approaches for effective parallel programming. The most widely used approach for distributed parallel computing for multicore Central Processing Units (CPUs) is Message Passing Interface (MPI) [1]. The MPI specifies the communication between separate processes, and it was designed for high performance on both massively parallel machines and on workstation clusters. The present-day version of the standard is MPI-2.2 approved by the MPI Forum at September, 2009. MPI library contains functions written in C and Fortran languages and in detail it is described in literature, such as [1], [2], or [3].

A different approach represents OpenMP with shared memory space, where all the cores can access the whole memory space. The OpenMP is an application programming interface for multi-platform parallel programming in C/C++ and Fortran. The current version of the standard is OpenMP 3.1 from July 2011. The specification and detailed tutorials could be found in [4], [5], or [6].

In this paper, both efficiency and limitations of multi-core processing are discussed and the impact in signal processing domain is proved. There are several projects implementing the main algorithms for digital signal processing. This paper deals with possibility of effective implementation of fast Fourier transform and discrete cosine transform. Libraries for fast computing the discrete Fourier transform, which commonly include real and/or complex, multidimensional, and parallel transforms can be found in [7], [8], etc.

Aim of this paper is to provide the comparison of different parallel approaches for signal processing. Two of well-known and widely used signal processing algorithms are implemented using MPI, OpenMP, MATLAB and DSP, then the results are discussed and compared. These algorithms will form a part of the benchmarks set useful for students and researchers interested in radio electronics.

The rest of this paper is organized as follows. Section II presents the chosen algorithms for parallel implementation in both CPU and DSP processors. The considered experiments with implementation of digital signal processing algorithms and achieved results are described in Section III, followed by short conclusion and future plans.

II. Evaluated Algorithms

In this section, two implementation of digital signal processing algorithms are outlined. The algorithms used for evaluation of parallel potentialities are Fast Fourier Transform and Discrete Cosine Transform.

A. Fast Fourier Transform Algorithms

A Discrete Fourier Transform (DFT) complexity grows with the square of the data length (N). Therefore, since the original paper of Cooley and Tukey published in 1965 [9] a tremendous effort has been devoted to the Fast Fourier Transform (FFT) algorithm research. The complexity of the FFT is generally in order of $N \log_2 N$ operations.

Many algorithms for the FFT calculations have been proposed in the past. Their very detailed overview containing the mathematical derivations gives a book [10]. The methods can be basically classified to the Decimation In Time (DIT) or Decimation In Frequency (DIF) families. Further classification of the methods is according the used radix – from the basic radix-2 the algorithms of radix-4 or radix-8 can be derived. It is also possible to use the combinations called split-radix [11] or mixed-radix FFT. A derivation of the basic method – radix-2 DIF is based on the recursive decomposition of the DFT [10]

$$X(r) = \sum_{l=0}^{N-1} x(l)\omega_N^{rl} \tag{1}$$

of the N-point input sequence $x(l)$ into two parts of the same length [10]:

$$X(r) = \sum_{l=0}^{N/2-1} x(l)\omega_N^{rl} + \sum_{l=N/2}^{N-1} x(l)\omega_N^{rl}. \tag{2}$$

After simple manipulations, it can be shown, that the radix-2 DIF FFT of N-sample length sequence $x(l)$ can be computed with the use of two half-size FFT's of sequences $y(l), z(l)$ [10]:

$$Y(k) = \sum_{l=0}^{N/2-1} y(l)\omega_{\frac{N}{2}}^{kl} \quad \text{and} \tag{3}$$

$$Z(k) = \sum_{l=0}^{N/2-1} z(l)\omega_{\frac{N}{2}}^{kl}, \tag{4}$$

with $Y(k) = X(2k), y(l) = x(l) + x\left(l + \frac{N}{2}\right)$, $Z(k) = X\left(2k+1\right), z(l) = \left(x(l) - x\left(l + \frac{N}{2}\right)\right)\omega_N^l$. Note that the twiddle factors ω_N^r are defined as

$$\omega_N^r = e^{jr\theta} = e^{jr\frac{2\pi}{N}}, \quad \text{where} \quad j = \sqrt{-1}. \tag{5}$$

An example of 8-point long FFT calculated using the radix-2 DIF algorithm is shown in Figure 1.

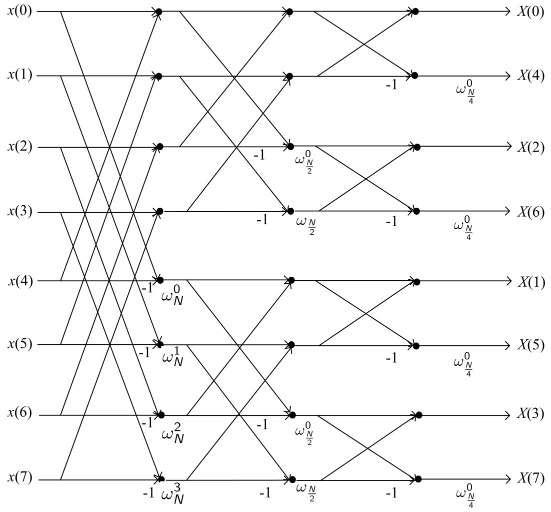


Figure 1. Radix-2 DIF graphical representation for 8-point data sequence.

B. Discrete Cosine Transform

For vector with dimension of N, the forward one-dimensional discrete cosine transform (1-D DCT) is defined in the following way [12]

$$D(u) = \gamma(u) \cdot \sum_{x=0}^{N-1} f(x) \cdot \frac{\pi u(2x+1)}{2N} \tag{6}$$

where $D(u)$ represents 1-D DCT coefficient of a vector item $f(x)$ while $u = 0, \ldots, N-1$. The constant $\gamma(u)$ could be expressed as follows [12]

$$\begin{array}{rcll} \gamma(u) & = & \sqrt{1/N} & : \quad u = 0 \\ & & \sqrt{2/N} & : \quad u \neq 0. \end{array}$$

From the symmetry of DCT base function, the computation load of the DCT can be exploited. There are several known algorithms, such as Arai's [13], Chen's [14], Loeffler's [15], or Vetterli's [16]. For further implementation, the Arai's forward DCT approach was chosen. Let $N = 8$, then according to [13], [17], 5 multiplication and 29 addition operations have to be evaluated in order to calculate eight one-dimensional coefficients. Supposing color block with 8×8 elements, the 1-D transform has to be repeated 48 (8×3 + 8×3) times to obtain 64 two-dimensional frequency coefficients. Therefore, for 8×8 color block, only 720 multiplications and 4 176 additions have to be calculated for transforming a single color block.

III. Experiments

Algorithms were tested via two dimensional transformation of color frame(s) with QSXGA resolution, i.e., with dimensions of 2,560×2,048 pixels. Each pixel is coded in RGB color space by 24 bits. Tested frames were separated into small blocks of $N \times N$ pixels. Those blocks represent input signal for the two-dimensional FFT, or DCT coder. FFT uses complex input/output values, whereas DCT algorithm is adapted for real data only. The proposed implementation of both algorithms (according to Subsection II-A and II-B) uses the common dimension of transform base in signal processing domain, i.e., $N = 8$. Only in MATLAB environment, the built-in functions with dimensions from 8 to 2,048 were used.

For the evaluation of considered parallel computing methods, the several test cases were performed. Mainly, the time consuming of two-dimensional FFT and DCT algorithms with MPI, OpenMP, MATLAB, and Texas Instruments DSP approaches were tested. Two-dimensional transforms were always divided to successive calculation of two 1-D transforms. In general, algorithms could use either fixed-point or floating-point number representation. The most famous open source FFT library FFTW [7] uses double precision floating-point representation in theirs functions, while DSP the library [8] from Texas Instruments (produces of present-day's most powerful DSPs) incorporates both, single and double precision routines. For basic confrontation with mentioned libraries, all data in our tests were represented in single precision floating-point format. Fixed-point releases would be implemented and optimized in the future.

All CPU based parallel computing tests were performed on HP BL465c G5 Blade Server with two quad-core Opteron processors and 32 GB of RAM. The core clock frequency is 2.7 GHz, synchronous DDRII memory was running on 800 MHz.

For the simulation results discussion, we also mention size of CPUs internal cache. Internal L1 cache 256 kB per processor (64 kB for data and 64 kB for instruction), L2 cache is 2 MB (4×512 kB) per processor, L3 cache 6 MB per processor, TLB (Translation Lookaside Buffer) of 4 kB.

The DSP based computing test were performed on Texas Instruments evaluation board with 32-bit floating-point digital signal processor TMS320C6747, with VLIW (Very Long Instruction Word) architecture, and clock frequency $f_{CPU} = 300$ MHz.

A. Implementation Results

Results from first test case are shown in Figure 2. For various QSXGA color frames, the length of MPI message buffer was altered. The buffer contains both the input picture data (from master to slaves communication), and transformed two-dimensional coefficients as well (from slaves to master communication). Average computation times were calculated from sixteen evaluations; 8 cores were used for all calculations. It can be seen, the first fall of the computation time for both transforms, which corresponds with hardware setting of blade server; concretely with TLB size. On the other hand, the second (wider) fall of the computation time corresponds with the L2 cache size. For further computing, the MPI message buffer size of 4 kB would be chosen.

From Figure 2 (a) and Figure 2 (b) it is obvious that the selected implementation of FFT algorithm is slower than implementation of DCT algorithm. For $N = 8$, the implemented FFT algorithm is approximately 1.5-times slower than DCT algorithm. The reason is that FFT needs complex data, while as DCT needs real input and output values. Therefore, thirty two QSXGA color frames could be transformed in 2.2 s by FFT, but only in 1.4 s by DCT method.

Second test case describes parallel implementation of FFT and DCT algorithms with help of OpenMP approach. For transformation of several QSXGA color frames, 1, 2, 4, or 8 cores were used. The number of transformed frames varied between 1 and 32 for FFT algorithm and between 1 and 128 for DCT algorithm. The computation times are shown in Figure 3. With dotted lines, the serial versions of implemented algorithms, as well as ideal curves for parallel versions are expressed. The ideal versions are computed as the portion of serial results. The dashed line in figures represents the results achieved by MPI approach as well.

It can be seen, for lower number of processed data, the OpenMP version is less effective than MPI version. In addition, while a single QSXGA color frame is being transformed, the computation time for serial version is lower that parallel version with 2 cores! Therefore, the beneficial using of simple OpenMP in signal processing domain could be bitrate, which is adequate to 64 QSXGA color frames.

B. Non Standard Implementations

MATLAB's Parallel Computing Toolbox provides running the script in up to 8 threads on a local computer or running it on a cluster machine using MATLAB Distributed Computing Server [18]. The main task is called Job and it is divided into Tasks, which are assigned to the individual workers by scheduler. The default scheduler for MATLAB Distributed Computing Server, MathWorks Job Manager, supports the Platform LSF, Microsoft Compute Cluster Server and Altair PBS Pro. Other schedulers can be integrated by user.

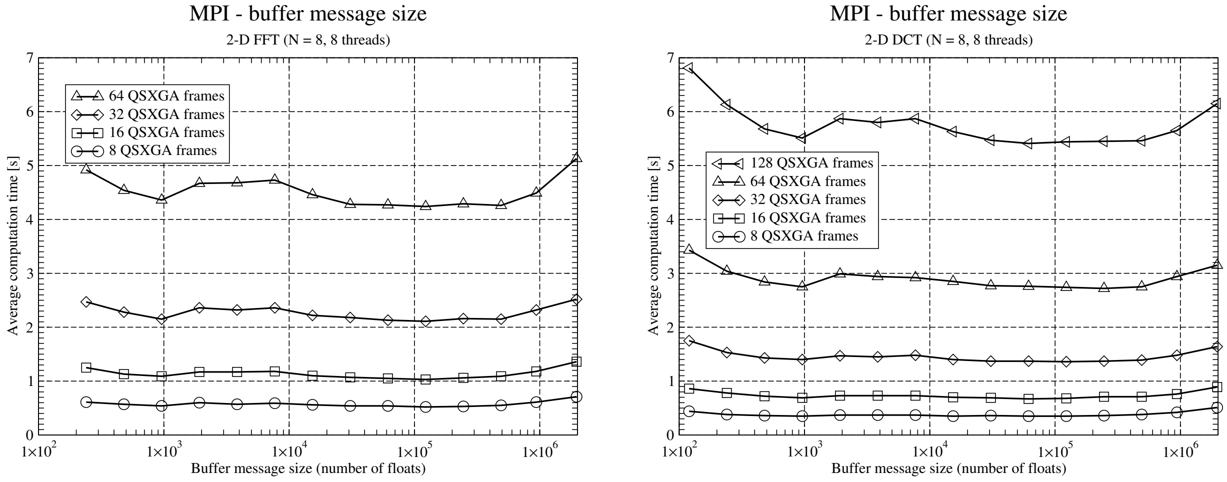
Third test case was performed in MATLAB environment. The MATLAB built-in functions `fft` and `dct` are called in all the individual workers. The computational time measurement starts before the parfor loop and ends after the variables' final reshape after the parfor loop. The results for the FFT and DCT computation from 1 to 8 threads for the blocks of vectors with the length of 8, 16, 32, 64, 128, 256, 512, 1024 and 2048 are depicted in Figure 4.

The application has to be divided into independent tasks which are then processed simultaneously. The most convenient way to solve this particular task uses the MATLAB functions as much as possible, because they are optimized to run fast and to use proper amount of memory. The *fft* and *dct* task is specific because of the use of both MATLAB functions and the parallel expressions. While the length of the array for the *fft* and *dct* function increases, the number of parallel loops decreases because of the total size of the matrix being processed. Thus, the computational time does not decrease constantly with increasing number of threads. This issue can be solved by using simpler algorithm, not the one which is based on two contradictory parts. This problem is to be solved and the new algorithm will be included in the benchmark dataset.

Table I
COMPUTATION TIME FOR TWO-DIMENSIONAL TMS320C6747 IMPLEMENTATION WITH VARYING PROGRAMMING APPROACHES ($f_{CPU} = 300$ MHZ, 1 QSXGA COLOR FRAME: 2,560×2,048 PIXELS)

Algorithm	Programming language	Computation time [s]
2-D FFT	C code	7.08
2-D FFT	Linear assembly	1.65
2-D DCT	C code	3.05
2-D DCT	Linear assembly	1.02

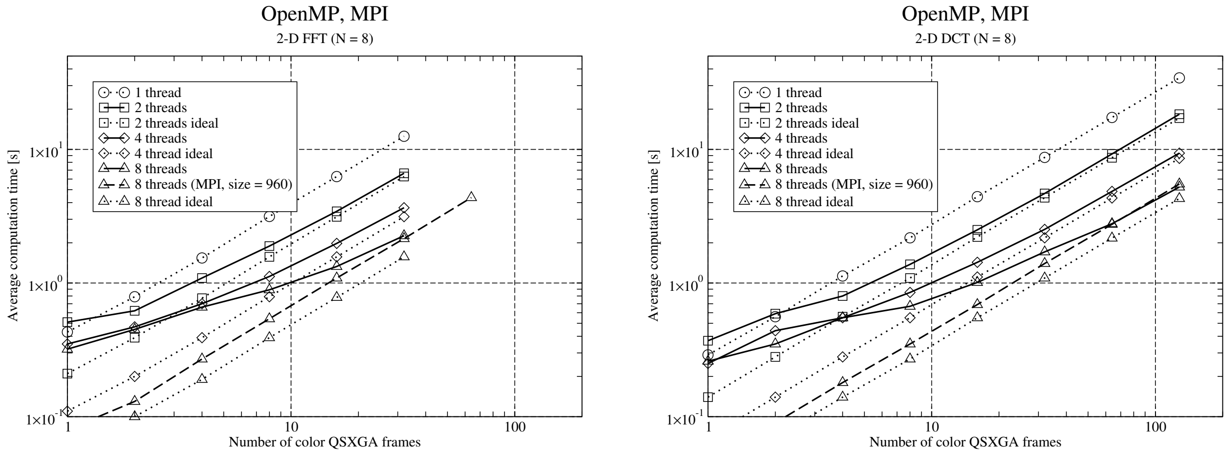
Last considered test case was performed by digital signal processor TMS320C6747, controlled by clock frequency of 300 MHz (9-times lower than CPU based tests). Although, the evaluation board contains only a single core DSP, the VLIW architecture meets the parallel approach. Selected



(a) 2-D FFT

(b) 2-D DCT

Figure 2. Average computation time for two-dimensional MPI implementation with varying buffer message size ($N = 8$, $f_{CPU} = 2.7$ GHz, 8 threads, QSXGA color frames: 2,560×2,048 pixels).



(a) 2-D FFT

(b) 2-D DCT

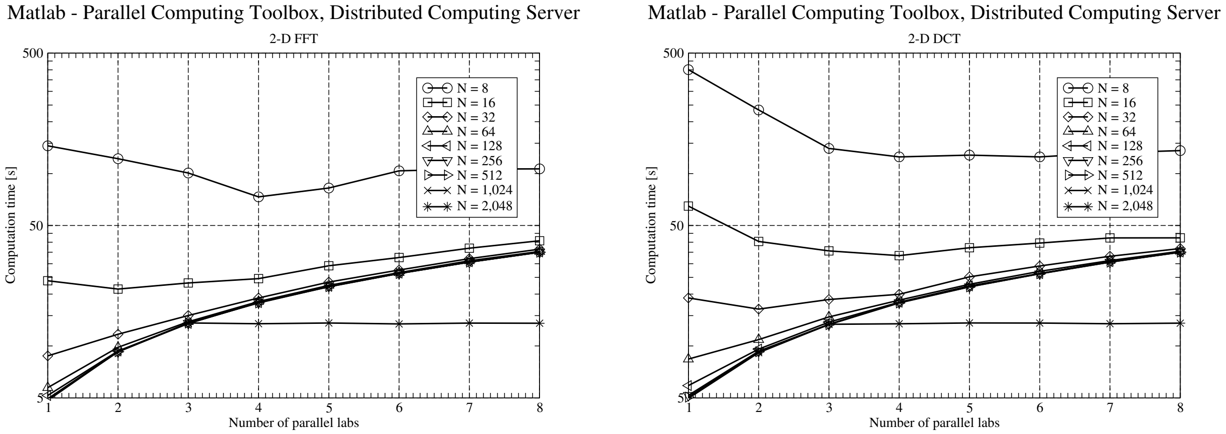
Figure 3. Average computation time for two-dimensional OpenMP implementation with varying transformed frames and threads number ($N = 8$, $f_{CPU} = 2.7$ GHz, QSXGA color frames: 2,560×2,048 pixels).

algorithms were implemented in C language and in linear assembly language. Development tool Code Composer Studio v.3.3 from Texas Instruments was used. All codes were optimized by CCS internal tools as well. The achieved results are shown in Table I.

It can be seen that the general abstraction brought by C code is not useful in this case. The low-level programming of both FFT and DCT algorithms represents outstanding contribution in signal processing. A single QSXGA frame could be transformed in 1.65 s by FFT, and in 1.02 s by DCT method.

IV. Conclusion and Future Work

The paper was focused on the implementation of two transforms, commonly used in signal processing domain. The two-dimensional FFT and DCT were chosen. The outline of currently used methods for parallel computing on CPU was performed as well. The MPI, OpenMP, and MATLAB approach were taken into account. The goal of the paper was also to present a possibility to create an interconnection between CPU based methods and VLIW architecture DSP evaluation boards. The future work would be focused mainly to implementation of digital signal pro-



(a) 2-D FFT

(b) 2-D DCT

Figure 4. Computation time for two-dimensional MATLAB implementation with varying parallel lab number ($f_{CPU} = 2.7$ GHz, 1 QSXGA color frame: 2,560×2,048 pixels).

cessing algorithms to Graphical Processing Units (GPUs) as well as to comparison with other CPUs, such as Intel quad-core Xeon e5640.

ACKNOWLEDGMENT

Research published in this submission was financially supported by the project CZ.1.07/2.3.00/20.0007 WICOMT of the operational program Education for competitiveness and by the Brno University of Technology Internal Grant Agency under project no. FEKT-S-11-12 (MOBYS) and the described research was performed in laboratories supported by the SIX project; the registration number CZ.1.05/2.1.00/03.0072, the operational program Research and Development for Innovation.

REFERENCES

[1] MPI Forum. *Message Passing Interface Forum* (2012-03-11). [online]. Available: http://www.mpi-forum.org/.

[2] A Message-Passing Interface standard. *The International Journal of Supercomputer Applications and High Performance Computing*, 8, 1994.

[3] Marc Snir, Steve Otto, Steven Huss-Lederman, David Walker, and Jack Dongarra. *MPI: The Complete Reference.* The MIT Press, 1998.

[4] *OpenMP* (2012-03-11). [online]. Available: http://openmp.org/wp/.

[5] Barbara Chapman, Gabriele Jost, and Ruud van der Pas. *Using OpenMP – Portable Shared Memory Parallel Programming.* The MIT Press, 2007.

[6] Blaise Barney. *OpenMP* (2012-03-11). [online]. Available: https://computing.llnl.gov/tutorials/openMP/.

[7] *FFTW Home Page* (2012-03-11). [online]. Available: http://www.fftw.org/.

[8] Texas Instruments. *TMS320C67x DSP Library* (2012-03-11). [online]. Available: http://www.ti.com/tool/sprc121.

[9] James William Cooley and John Wilder Tukey. An algorithm for the machine calculation of complex Fourier series, *Mathematics of Computation*. 19, 297–301, 1965.

[10] Eleanor Chin-hwa Chu and Alan George. *Inside the FFT Black Box: Serial and Parallel Fast Fourier Transform Algorithms (Computational Mathematics)*, 1st ed. CRC Press, 1999.

[11] Pierre Duhamel and Henk Hollmann. Split radix FFT algorithm, *Electronics Letters*, vol.20, no.1, pp. 14–16, 1984.

[12] Kamisetty Ramamohan Rao and Patrick Yip. *Discrete Cosine Transform. Algorithms, Advantages, Applications.* San Diego: Academic Press, Inc., 1990.

[13] Yukihiro Arai, Takeshi Agui, and Masayuki Nakajima. A Fast DCT-SQ Scheme for Images. *IEICE Transactions (1976–1990)*, 1988, vol. E71-E, no. 11, pp. 1095–1097.

[14] Wen-Hsiung Chen, Harrison Smith, and Sam Fralick. A fast computational algorithm for the discrete cosine transform. *IEEE, Transactions Commun*, 1977, pp. 1004–1009.

[15] Christoph Loeffler, Adriaan Ligtenberg, and George Moschytz. Practical fast 1-D DCT algorithms with 11 multiplications. *Proc. IEEE ICASSP*, 1989, pp. 988–991.

[16] Martin Vetterli. Fast 2-D discrete cosine transform. In *Proc. ICASSP*, 1985, pp. 1538–1541.

[17] Rafael Gonzalez and Paul Wintz. *Digital Image Processing.* Boston: Addison Wesley Publishing Company, 1987.

[18] MathWorks. *MATLAB and Simulink for Technical Computing* (2012-03-11). [online]. Available: http://www.mathworks.com/.

Estimating Perceived Video Quality from Objective Parameters in Video over IP Services

Pedro de la Cruz Ramos[1], Joaquín Navarro Salmerón[1], Raquel Pérez Leal[2], Francisco González Vidal[1]

[1]Departamento de Ingeniería de Sistemas Telemáticos
Universidad Politécnica de Madrid
Madrid, Spain
{pcruzr, navarro, vidal}@dit.upm.es

[2]Departamento de Teoría de la Señal y Comunicaciones
Universidad Carlos III de Madrid
Madrid, Spain
rpleal@tsc.uc3m.es

Abstract — **In Video over IP services, perceived video quality heavily depends on parameters such as video coding and network Quality of Service. This paper proposes a model for the estimation of perceived video quality in video streaming and broadcasting services that combines the aforementioned parameters with other that depend mainly on the information contents of the video sequences. These fitting parameters are derived from the Spatial and Temporal Information contents of the sequences. This model does not require reference to the original video sequence so it can be used for online, real-time monitoring of perceived video quality in Video over IP services. Furthermore, this paper proposes a measurement workbench designed to acquire both training data for model fitting and test data for model validation. Preliminary results show good correlation between measured and predicted values.**

Keywords - Video over IP, Perceived Quality, Quality Models, Quality of Experience, Quality of Service.

I. INTRODUCTION

User Quality of Experience (QoE) is a determining factor for successful deployment of emerging multimedia services. QoE is easy to understand, but complex to implement in real systems. This complexity is mainly due to the difficulty of its modeling, evaluation and translation into Quality of Service (QoS) parameters.

A complete QoE management procedure should encompass at least: monitoring the user experience when consuming the service; adapting the provisioning of the content to the varying context conditions; predicting potential QoE level degradation; and recovering from QoE degradation due to system changes. In order to have a complete control of the final user experience, all these tasks must be performed in-service and in a coordinated way.

Among multimedia services, Video over IP applications have reached a remarkable market penetration. Furthermore, Video over IP customers expect a QoE comparable to traditional broadcast systems. So the ability to measure, estimate and monitor user perceived quality in near real time and to relate it to network conditions, is critical for Video over IP service providers.

This paper focuses on the perceived video quality aspects of Video over IP streaming and broadcasting services. A model for estimating the Video Quality Metric (VQM) [1] as defined in ITU-T J.144 [2] is proposed.

Subjective quality measurements, as those defined in ITU-T P.910 [3], are undoubtedly the most precise, and constitute the benchmark for any estimation model. However, they are costly, both in time and resources. Thus, our approach has been to estimate an objective perceptual distortion metric, originally defined as a Full Reference (FR) measure, from coding and Network QoS parameters, using a model similar to those suggested in [4], [5], [6] and [7].

The proposed model takes as input easily measurable video coding and Network QoS (NQoS) parameters, and includes some fitting parameters that depend mainly on the information contents of the video sequences. A method for computing them from Average Spatial and Temporal Information content measures (ASI/ATI), similar to those defined in ITU-T P.910 [3], is proposed. All the values required for the estimation can be obtained without reference to the original video sequence, enabling online, real-time evaluation of perceived video quality in Video over IP services.

In the following sections previous work is reviewed; the estimation model is proposed; the method for computing the fitting parameters is described; a measurement workbench is presented; the main conclusions are summarized; and some future work is outlined.

II. RELATED WORK

In [4], a comprehensive model, based on theoretical considerations, is proposed in order to relate several coding and network parameters to the Perceptual Distortion Metric (PDM) of MPEG-2 sequences. The coefficients of this model mainly depend on the complexity (information contents) of the analyzed sequence.

The dependence of VQM on Video Coding Rate (VCR), display format (resolution), codec type and "motion contents", is analyzed for MPEG-2 and H.264 Advanced Video Coding (AVC) sequences in [5]. Although this model takes into account the effects of codec type and coding parameters, it obviates the dependence of VQM on the transmission network parameters.

In the previous models, the variation of the chosen metric follows a negative power function of VCR. Regarding Packet Loss Ratio (PLR), [4] proposes a linear variation while [5] does not consider its effect at all.

Reference [6] estimates the Perceived Video Quality of H.264 sequences combining coding and network QoS parameters (namely VCR and the Packet Loss Frequency, PLF) and codec features in a parametric packet level model. This model states that the variation of the Perceived Video Quality with VCR follows a logistic function while its variation with PLF follows a negative exponential.

In [7], a parametric null reference (NR) model, called “Temporal-Visual (T-V) Model” is proposed. The objective of this model is to estimate the Perceived Video Quality of MPEG-2 and H.264 sequences, using network QoS, coding and other parameters. This model states that the Perceived Video Quality is related to VCR by an exponential function while its variation with PLR follows a logistic function.

One of the key aspects when designing a model is the determination of the “fitting” parameters. In the previous proposals different approaches are followed. In [4], the fitting process is performed for each individual sequence. In [5], the sequences are classified in classes according to their ‘motion contents’ and values are assigned to the parameters for each group of sequences. In [6] and [7], neither the contents nor the spatial or temporal complexity of the sequences are considered.

None of the analyzed models completely fulfill our needs. Some of them are too specific for a particular kind of application or propose forms of variation that do not correspond to our measurements, which rather suggest a (positive or negative) power function. Most of them estimate the subjectively perceived video quality or metrics other than VQM. In [5], VQM is estimated, but it does not take into account the effect of the transmission network. Furthermore, none of the reviewed proposals include the effect of the complexity and/or information contents of the video sequences.

All these reasons lead us to develop a new model for online, real-time estimation of VQM in Video over IP streaming and broadcasting services, using coding and network QoS parameters and the complexity and information contents of the video sequences.

III. Model Description

Different measurements, obtained using the Video Quality Experts Group (VQEG) FR-TV1 test sequences [10] and our Measurement Workbench (described later), confirmed the variation of VQM with coding parameters according to the model of [5]. However, these measurements also showed that the variation of VQM with PLR is far from linear in most of the cases. Figure 1 shows the effects of coding and packet loss.

Figure 1a shows the variation of VQM with VCR for all sequences coded using H.264, prior to transmission (i.e., with no transmission losses). VCR is the actual Average VCR (Video Data Size/Duration).

The relation between VQM and PLR is shown in Figure 1b. In this measurement all the sequences have been coded using H.264 with VCR=5 Mbps. The plotted VQM is the average result of several repetitions with the same PLR in order to attenuate random effects. The value for PLR=0 (no losses) is the value of VQM prior to transmission as given in Figure 1a.

VQM can be split into two parts in order to separate the effects of coding and transmission:

$$VQM = VQM_C + VQM_L \qquad (1)$$

where

VQM_C is the contribution of coding to VQM.

VQM_L is the contribution of packet losses to VQM.

By plotting VQM_C and VQM_L in logarithmic scale, it can be noticed that both curves fit very well to a power function, as they are nearly linear in both cases. They can be expressed as:

$$VQM_C = VQM_{REF} \bullet (VCR/VCR_{REF})^{-K_C} \qquad (2)$$

$$VQM_L = (1\text{-}VQM_C) \bullet (PLR/PLR_1)^{K_L} \qquad (3)$$

where

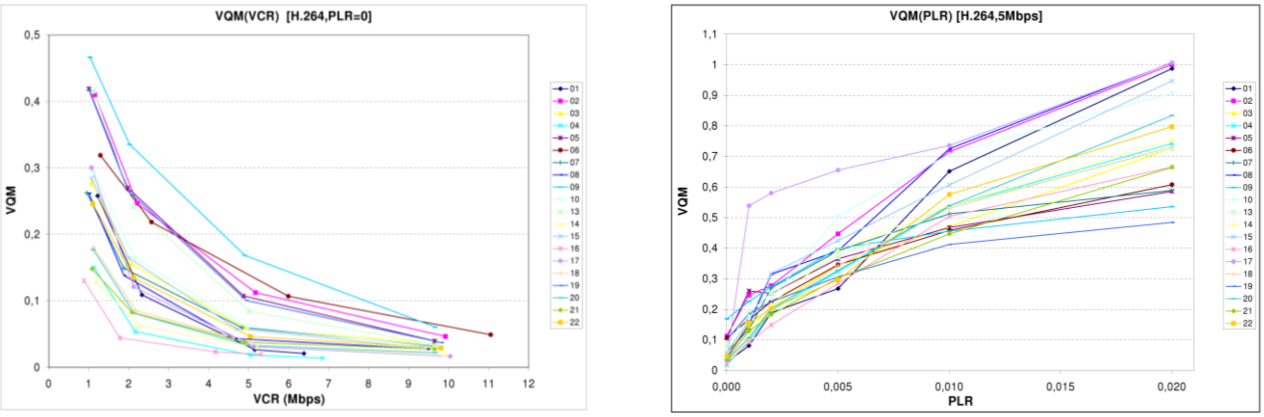
VCR_{REF} is a reference VCR (e.g., 1Mbps).

VQM_{REF} is the value of VQM at the reference VCR.

PLR_1 is the value of PLR for which VQM = 1.

VQM_{REF} and K_C depend on the codec, the coding parameters (except VCR), and the characteristics of the video sequence (e.g., spatial and temporal complexity, information contents, etc.).

PLR_1 and K_L depend on the codec, the coding parameters (including VCR), and the characteristics of the video



(a)

(b)

Figure 1. Variation of VQM with VCR (for PLR=0) and PLR.

sequence. Their variation with VCR fits very well to a function of the form:

$$F(VCR) = A + B \bullet VCR \bullet (1+C \bullet e^{-(VCR/D)^{\wedge}K}) \qquad (4)$$

where A, B, C, D, K are fitting parameters that depend on the codec, the coding parameters (except VCR) and the characteristics of the video sequences (type, format and information contents).

For K≠2 this function approximates to a Weibull curve on top of a linear asymptote. For K=2 it corresponds to a Rayleigh curve, also on top of a linear asymptote. Figure 2 shows the fitting of this function to the values of K_L for a group of sequences coded using H.264.

According to this model, for a given PLR, there is a VCR that minimizes VQM, i.e., maximizes the perceived quality. The consequence is that for higher coding rates, and against the common assumption, the perceived quality decreases due to the increment of packet losses. Therefore, in real systems with transmission errors, increasing the coding rate beyond a certain limit is not only useless (as users don't perceive the difference), but may even be counterproductive. This behavior was already noticed in [4].

IV. Estimation of Model Parameters

As seen in the previous section, the characteristics of each sequence, i.e., its type, complexity and information contents, directly influence the perceived video quality. Therefore, a crucial aspect is how to compute the model parameters for each video sequence, without having to fit the model specifically for each of them. This paper proposes the use of two measures similar to the Spatial/Temporal Information (SI/TI) measures, described in [11]. SI/TI measurements evaluate the spatial/temporal information detail in a way similar to the perception of a human viewer. They are standardized in ITU-T Recommendation P.910 [3]. These measurements are rather easy to obtain using well-known techniques such as the Sobel filter (a simple high-pass, edge enhancement digital filter widely used in image processing) and pixel-wise difference.

However, our preliminary results concluded that SI/TI measurements, as originally described, i.e., based on the maximum SI/TI values of the frames in the sequence, are too sensitive to exceptional SI/TI values of individual frames [8]. Therefore, in order to attenuate this effect, the Average Absolute Spatial/Temporal Information (ASI/ATI) measurements are defined as follows:

1) Use the absolute value of the pixel-wise difference of luminance values of successive frames to compute the Temporal Information values of each frame.

2) Take the average of the SI/TI values of all frames as the ASI/ATI value of the sequence.

ASI/ATI measures will be used as indexes into precomputed "complexity tables". The model parameters for a given sequence will be computed by linear interpolation in these tables. The methods for populating the Complexity Tables and using them to compute the model parameters for arbitrary sequences are described in [8].

The proposed method enables online, real-time monitoring of perceived video quality, because the whole process (ASI/ATI computation, table lookup, interpolation, and model evaluation) takes much less time than the duration of the sequences. In addition, all values required for VQM estimation can be either obtained from the Network Management System (NMS) or measured at the receiving side, so no measurements on the reference sequence are required.

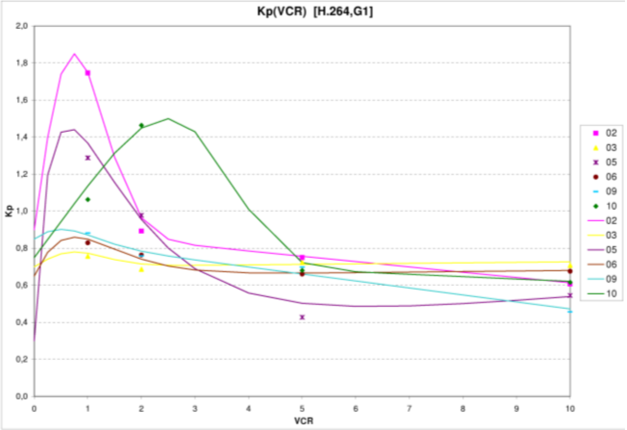


Figure 2. Variation of K_L with VCR for a group of sequences.

V. Measurement Workbench

This section describes the measurement workbench that was implemented in order to obtain training data for model fitting and test data for model validation [9]. Figure 3 shows its functional architecture, which was implemented using the following tools:

- Encoder/Decoder: FFmpeg 0.6.1-2/4 + libX264 [for H.264]
- Transmitter/Receiver: Videolan VLC 1.1.5/7
- Network Simulator: NetEm (Linux Kernel 2.6.35)
- Information Measurement: STIX 0.9
- Distortion Measurement: ITS/NTIA BVQM 1.4
- QoS Measurement: WireShark 1.6.0

Specific tools were developed in order to perform ASI/ATI measurements and frame loss concealment. Frame loss concealment is required because the received and reference sequences must have equal length for BVQM to work adequately. The operation of the frame concealer is based on detecting the lost frames and duplicating the previous one (i.e., freezing).

The workbench comprises four physical nodes. The first one is the emitter station that performs encoding and transmission operations. The second one is the receiver station, responsible for reception, decoding and frame loss concealment. The third node is the network simulator, capable of simulating different network parameters and scenarios. The last one is the measurement workstation, used to perform VQM, QoS and ASI/ATI measurements.

All these nodes were physically implemented using DELL Optiplex 755 PCs with Intel Core 2 Duo processors at 2.66GHz with 3GB of RAM. The emitter and receiver stations and the network simulator run under Ubuntu Linux

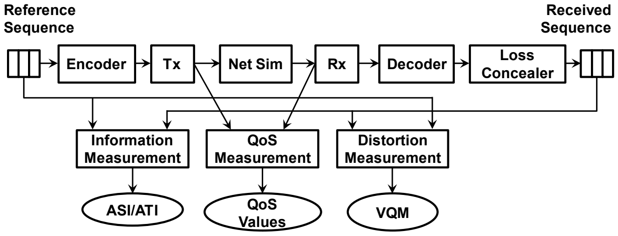


Figure 3. Measurement Workbench Functional Architecture.

10.04.2 LTS, and the measurement workstation under Windows XP Professional SP3. The nodes communicate through a 100Mbps Fast Ethernet LAN.

The test sequence database [10] includes both high and low motion (including static) sequences, spatially simple as well as complex, both natural (filmed) and artificial (animation or computer-generated). All sequences are in ITU-R BT.601 UYVY (Big-YUV) format (either 525lines@60Hz or 625lines@50Hz).

Measurements were made for all sequences coded in MPEG-2, MPEG-4 and H.264 AVC, for several VCR and PLR values. In order to account for random effects of packet losses, these measurements were repeated several times for the same nominal VCR and PLR values. In total more than 6,000 data points were collected. These data will be statistically analised in order to validate the accuracy of the model. Preliminary results show good correlation between measured and predicted values (see Table I).

TABLE I. PRELIMINARY STATISTICAL RESULTS

Codec	Correlation	Avg.Error	RMSE
MPEG-2	0.9519	0.0339	0.0703
MPEG-4	0.9471	0.0551	0.0704
H.264	0.9462	0.0549	0.0749
ALL	0.9511	0.0487	0.0722

VI. CONCLUSION AND FUTURE WORKS

This paper proposed a new model for online, real-time estimation and monitoring of perceived video quality in Video over IP streaming and broadcasting services, using the Video Quality Metric (VQM) as objective measure. This model is based on video coding and Network Quality of Service (NQoS) parameters. Our model shows that the contributions to VQM from coding (VQM_C) and packet losses (VQM_L) follow power functions of the Video Coding Rate (VCR) and Packet Loss Ratio (PLR) respectively.

Additionally, the model includes fitting parameters that depend mainly on the complexity (information contents) of the video sequence. These parameters are estimated using the Average Absolute Spatial and Temporal Information (ASI/ ATI) contents of the sequence.

A measurement workbench was implemented. It comprises several nodes, such as emitter and receiver stations, a network simulator and a measurement workstation. This workbench was used with a public test sequence database in order to obtain training data for fitting the model parameters. Preliminary results show good correlation between measured and predicted values.

The following points remain open: sequence classification based on features other than ASI/ATI, and use of different complexity tables for each group of sequences; influence of coding parameters other than VCR; effect of NQoS parameters other than PLR (e.g., Packet Error Ratio (PER) and/or Bit Error Ratio (BER)); influence of error/loss patterns (distribution), in particular the Average Burst Length (ABL); effect of extreme variation of the ASI/ATI values of received (distorted) sequences with respect to that of original sequences, in the computation of fitting parameters from the complexity tables; definition of spatial and temporal information measures based on chrominance values, and inclusion of them in the estimation model.

ACKNOWLEDGMENT

This research was partially supported by the Spanish Ministry of Science and Innovation grant TEC2008-06539 (ARCO Project).

REFERENCES

[1] M. H. Pinson and S. Wolf, “A New Standardized Method for Objectively Measuring Video Quality,” IEEE Transactions on Broadcasting, Vol. 50, No. 3, pp. 312-322. September 2004.

[2] ITU-T J.144, “Objective Perceptual Video Quality Measurement Techniques for Digital Cable Television in the Presence of a Full Reference,” International Telecommunication Union, March 2004.

[3] ITU-T P.910, “Subjective Video Quality Assessment Methods for Multimedia Applications,” International Telecommunication Union, April 2008.

[4] P. Frossard and O. Verscheure, “Joint Source/FEC Rate Selection for Quality-Optimal MPEG-2 Video Delivery,” IEEE Transactions on Image Processing, Vol. 10, No. 12. December 2001.

[5] J. Joskowicz, J. C. López Ardao, M. A. González Ortega, and C. López García, “A Mathematical Model for Evaluating the Perceptual Quality of Video,” Second International Workshop on Future Multimedia Networking (FMN 2009), Coimbra, Portugal, June 2009. In Lecture Notes on Computer Science (LNCS) 5360, pp. 164-175, Springer Verlag, 2009.

[6] K. Yamagishi and T. Hayashi, “Parametric Packet-Layer Model for Monitoring Video Quality of IPTV Services,” IEEE International Conference on Communications ICC 2008, May 2008.

[7] M. N. Garcia, R. Schleicher, and A. Raake, “Impairment-Factor-Based Audiovisual Quality Model for IPTV: Influence of Video Resolution, Degradation Type, and Content Type,” EURASIP Journal on Image and Video Processing, Volume 2011, Article ID 629284, pp. 14. 2011.

[8] P. de la Cruz Ramos, F. González Vidal, and R. Pérez Leal, “Perceived Video Quality Estimation from Spatial and Temporal Information Contents and Network Performance Parameters in IPTV,” Proc. of the Fifth IARIA International Conference on Digital Telecommunications (ICDT 2010), pp. 128-131, Athens, Greece, June 2010.

[9] E. Álvarez Villacé, “Design and Implementation of a Measurement Workbench for Estimation of Perceived Video Quality in IPTV,” Master Thesis, Polytechnical University of Madrid, July 2011.

[10] VQEG, “Final Report from the Video Quality Experts Group on the Validation of Objective Models of Video Quality,” ITU-T SG 9,Contribution COM 9-8, June 2000.

[11] A. A. Webster, C. T. Jones, M. H. Pinson, S. D. Voran, and S. Wolf, “An Objective Video Quality Assessment System Based on Human Perception,” SPIE Conference on Human Vision, Visual Processing, and Digital Display, 1993.

A Radio-Resource Switching Scheme in Aggregated Radio Access Network

Xuan-Dat Trinh*, Gahee Jo*, Jaewon Lee*, Jee-Hyeon Na**, Woogoo Park**, and Ho-Shin Cho*
*Electronics Engineering Department, School of IT Engineering
Kyungpook National University
Daegu, Republic of Korea
** Electronics and Telecommunications Research Institute
Daejeon, Republic of Korea
E-mail: *{xuandat-trinh, jghv, jwlee, hscho}@ee.knu.ac.kr, **{jhna,wgpark}@etri.re.kr

***Abstract* – In an aggregated radio access network (A-RAN), wherein multiple radio access technologies (RAT) coexist, different radio resource utilization in each RAT may result in decreasing quality of service (QoS). In this paper, we propose a radio resource switching scheme in A-RAN coexist to provide multi-modal mobile users with the best service using a cloud base-station concept. The proposed scheme is manipulated to optimize radio utilization and QoS.**

Keywords: radio-resource; cloud; radio-access-network.

I. INTRODUCTION

The widespread and increasing use of smart-phones and broadband based services such as high-quality video and peer-to-peer services has caused the explosion of data traffic in mobile networks. To cope with this surge of data traffic, new technologies such as the WiMAX [1] and LTE [2] families have been implemented in a short period. Therefore, these newly deployed systems are likely to coexist with legacy systems, with each owning a separately their radio access network, resulting in high capital expenditure/operating expenditure (CAPEX/OPEX). Moreover, dynamic variations in traffic load may cause lower average utilization of a base station. To settle these challenges, a cloud-conceptual base station system has been introduced. Through virtualization using cloud technologies, any user equipment (UE) is able to access one of the common cell-sites, behind which multiple radio access technologies (RATs) are used to service a user-specific traffic in the best possible manner. We call such a network with multi-RATs, an aggregated radio access network (A-RAN). In A-RAN, it is also anticipated that software-defined radio equipment for signal processing of each radio access technology (RAT) provides the capability of sharing radio resources between different RATs to optimize frequency usage. In this paper, we propose a scheme to switch radio resources between the different radio access technologies used within an aggregated radio access network (A-RAN) to increase spectrum utilization.

The rest of this paper is organized as follows. The system to which our proposed scheme is applied is described in Section II. The proposed scheme to switch radio resources is described in Section III. The current simulation system, discussion, and further works to fulfill our study are presented in Section IV. Finally, Section V summarizes our conclusions.

II. SYSTEM MODEL

A. Aggregated radio access network architecture.

A-RAN has been developing to reduce network implementation and maintenance costs as well as to increase efficiency of hardware usage. Some A-RAN models have been introduced such as KT CCC (Cloud Communication Center) [3], Alcatel-Lucent lightRadio [4] or C-RAN of China Mobile Research Institute [5]. Fig. 1 shows the generalization of A-RAN architectures mentioned above. The A-RAN consists of Radio Units (RUs) and Digital Units (DUs), as shown in Fig. 1. A DU performs all of the functions performed by a traditional base station transceiver (BTS) and base station controller (BSC) in a legacy system, including baseband signal processing of user data, radio-resource allocation, and various control functions. On the other hand, an RU only has a physical radio interface comprising an antenna and power amplification. By centralized processing and control functions in DU, DU can be built on cloud computing environment to optimize hardware utilization as well as using software-defined radio equipments to have flexible capability of processing any baseband signals.

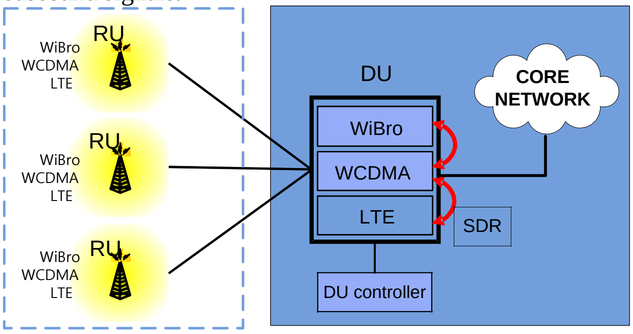


Figure 1. A-RAN architecture.

B. System model

In our study, we assume three different wireless access systems: A, B, and C. These systems are characterized by the carried traffic types such as packet or circuit-switched and the availability of carrier aggregation, a functionality introduced for broadband services in the LTE-Advance standard [6]. The descriptive parameters of the three above mentioned systems are summarized in Table 1.

TABLE I. SYSTEM PARAMETERS

Parameters	System A	System B	System C
System bandwidth (MHz)	40	40	20
Channelization (MHz)	20, 10, 5, 5	20, 10, 5, 5	5, 5, 5, 5
Carrier aggregation	Yes	No	No
Traffic type	Packet traffic	Packet traffic	Circuit traffic
Frame length (ms)	10	10	N/A

Systems A and B have LTE-like orthogonal frequency division multiple access (OFDMA) channel architectures [7], in which the time axis is divided into 10-ms frames that are further subdivided into ten 1-ms slots. In the frequency domain, the subcarrier spacing is 1 MHz. In the 2-dimensional channel domain, a resource-allocation unit called a resource block (RB) is defined as 1 ms × 1 MHz [time × frequency]. There are three UE types: mode-A-only, which is capable of only accessing system A; mode-B-only, which is for system B only; and multi-modal UE, which is capable of using both systems A and B. On the other hand, system C is accessible to all UE for circuit traffic. Moreover, in terms of the channel handling capability, the UE is categorized as having 5, 10, 20, and 40 MHz capability.

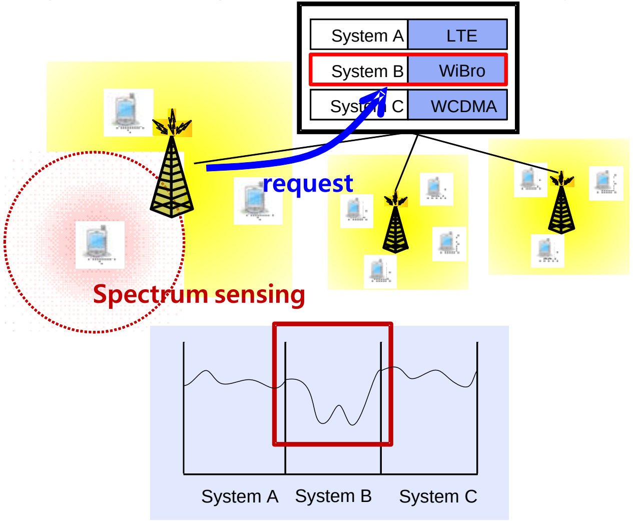


Figure 2. Spectrum map in UE

Before making a call request, a multi-modal UE performs a kind of preliminary preparation called early spectrum sensing (ESS). It measures the power levels over the entire frequency band of Systems A, B, and C and builds a spectrum map based on these measurements, as shown in Figure 2. Then, the UE tries to send a call request to the system that contains the lowest power level, which means that the frequency availability is the highest. When the target system receives UE request, if it has enough available resource and its average packet loss ratio over N frames is lower than packet loss threshold, it will admit the call. Otherwise, it consults DU controller about other available system. By checking remaining resource and packet loss ratio of other systems, DU controller informs target system of a new system that has enough free resource and average packet loss ratio over N frames is lower than packet loss threshold, and then target system signals to UE to send call request to the new one. If other systems also do not meet these two conditions, UE call request will be dropped, with real-time traffic, or put into buffer, with delay tolerance traffic.

III. PROPOSAL OF RADIO RESOURCE SWITCHING

DU is able to monitor the entire radio resource utilization for Systems A, B, and C, and control the amount of radio resources belonging to each system if needed. On that basis, we propose a scheme for switching radio resources between systems to ensure better system performance and higher resource utilization. The proposed scheme is illustrated using the flow chart shown in Figure 3.

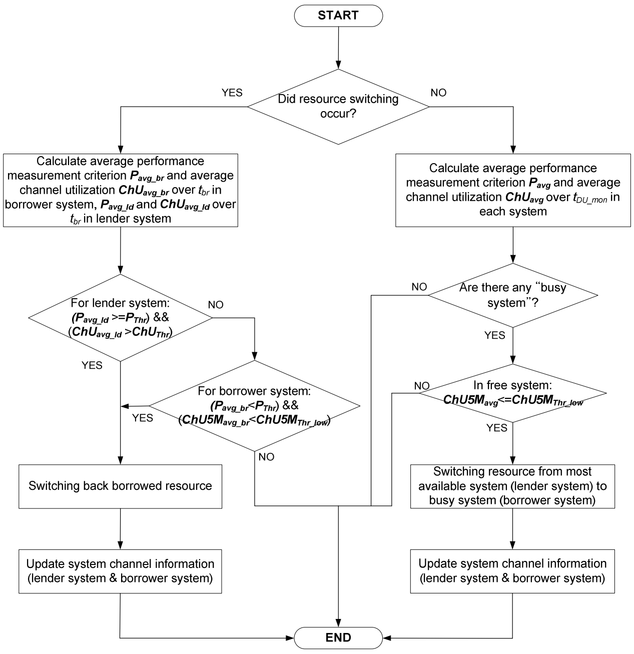

START
Did resource switching occur?
YES
NO
Calculate average performance measurement criterion P_{avg_br} and average channel utilization ChU_{avg_br} over t_{br} in borrower system, P_{avg_ld} and ChU_{avg_ld} over t_{br} in lender system
Calculate average performance measurement criterion P_{avg} and average channel utilization ChU_{avg} over t_{DU_mon} in each system
For lender system: $(P_{avg\_ld} >= P_{Thr})$ && $(ChU_{avg\_ld} > ChU_{Thr})$
NO
YES
For borrower system: $(P_{avg\_br} < P_{Thr})$ && $(ChU5M_{avg\_br} < ChU5M_{Thr\_low})$
YES
NO
Are there any "busy system"?
NO
YES
In free system: $ChU5M_{avg} <= ChU5M_{Thr\_low}$
NO
YES
Switching back borrowed resource
Switching resource from most available system (lender system) to busy system (borrower system)
Update system channel information (lender system & borrower system)
Update system channel information (lender system & borrower system)
END

Figure 3. Proposed radio resource switching scheme

In our proposed scheme:

- If radio resource switching (RRS) has not yet occurred, DU calculates the performance measures for each period, t_{DU_mon}
- If RRS has occurred, DU calculates the performance measures for each period, t_{br}, in the borrower system to which the free radio resources are switched, and for each t_{ld} in the lender system from which free radio resources are switched.

ISBN: 978-1-61208-193-9

We define a "busy system" as a system in which both the average packet loss rate over period t_{DU_mon}, P_{avg}, is higher than the packet loss threshold, P_{Thr}, and the average channel utilization over period t_{DU_mon}, ChU_{avg}, is higher than the channel utilization threshold ChU_{Thr}, which is 95 percentage capacity of all channels in system. If these two conditions do not occur simultaneously, we consider the system to be a "free system".

The operation of our scheme is described below. If radio resource switching has not yet occurred, the DU continuously monitors the average packet loss rate, P_{avg}, and average channel utilization, ChU_{avg}, in each system for each period, t_{DU_mon}. When it finds a busy system, it then checks the utilization of 5MHz channels ChU5M in the free systems: if it is lower than a threshold called $ChU5M_{Thr\_low}$:

$$ChU5M_{Thr\_low} = (num\_5M-1)/\ num\_5M \qquad (1)$$

whereas num_5M is the number of 5MHz channels in free system, the DU will switch resources from the free system with the most available resources to the busy system. Then, each system will update its channelization information. If radio resource switching has already occurred, the DU calculates the average packet loss ratio, P_{avg_br}, and average channel utilization, ChU_{avg_br}, over t_{br} in the borrower system, along with the average packet loss ratio, P_{avg_ld}, and average channel utilization, ChU_{avg_ld}, over t_{ld} in the lender system. In the lender system, if both P_{avg_ld} and ChU_{avg_ld} are higher than thresholds P_{Thr} and ChU_{Thr}, respectively, the DU restores the lent resources to the system to which they belong. This means the lender system becomes busier and needs to recall its lent resources. In the case of the borrower system, if the following condition happens

$$P_{avg\_br}<P_{Thr}\ \&\ ChU5M_{avg\_br}<ChU5M_{Thr\_low} \qquad (2)$$

the DU restores the lent resources to the original system. This indicates that the load on the borrower system is decreasing and it may no longer need the additional resources. When monitoring system C, packet lost ratio is replaced by call block probability.

In A-RAN, channel processing modules of system can be software-defined radio equipments. So it is able to switch radio channel from one system to other system by re-configure the software of channel processing module.

IV. ONGOING SIMULATION

A. *Simulation environment*

We start the computer simulation under a single-cell condition with a static UE number, but will eventually consider multi-cell environments and user mobility with handoffs between them. UEs are distributed uniformly in cell coverage. DU manages a finite-length buffer for each user in order to store data coming from the core network side. If the buffer overflows, the packets are lost. In this paper, the packets are categorized as real-time packet and delay-tolerant packet, with tolerances ranging from one to three frames. The traffic generation of user *n* is modeled as a Poisson arrival process with rate λ_n, $n\in\{1,2,\ldots N\}$. Then, the total arrival rate is

$$\lambda_{all} = \sum_{n=1}^{N} \lambda_n \qquad (3)$$

The traffic volume is normalized by the number of RBs needed to carry it.

B. *Initial simulation result*

In this part, we will discuss our initial simulation result.

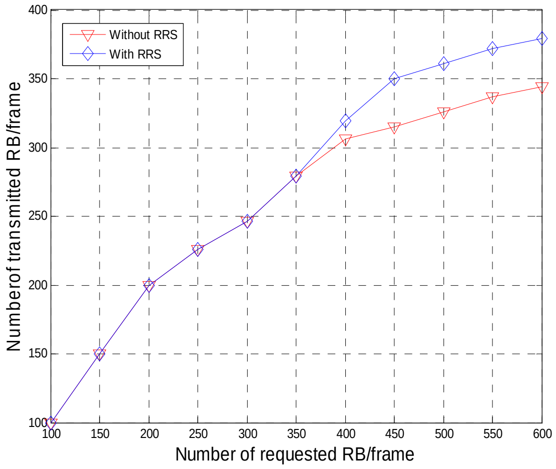


Figure 4. System performance

Fig. 4 compares the system throughput when radio resource switching is used and when it is not used. RRS helps to improve system throughput. Additional simulations will help us ensure that packet lost ratio with RRS is lower than without RRS.

V. CONCLUSION AND FUTURE WORKS

In this paper, a radio resource switching scheme was proposed for application to an A-RAN based on the deployment of a cloud base-station in order to improve frequency utilization, increase system throughput, and enhance the QoS. We expect that after the system simulations are completed, the simulation results will validate the performance enhancement of the proposed scheme. In the next step, we will consider this scheme in an environment closer to reality, i.e., system parameters are same as defined in standards.

ACKNOWLEDGMENT

We would like to thank the Electronics and Telecommunications Research Institute (ETRI) in Daejeon, Republic of Korea for their support to conduct this study.

REFERENCES

[1] WiMAX http://ieee802.org/16/tgm/ <retrieved: Feb 1st 2012>

[2] 3GPP LTE http://www.3gpp.org/LTE <retrieved: Feb 1st 2012>

[3] Hyun Pyo Kim "KT smart network world – Mobile Network Technology and Advanced Plans" (in Korean) in *The 21st High Speed Work Shop*, Gyeongju, Republic of Korea Jan. 19th -21st 2011. Available: http://plum.hufs.ac.kr/hsn2011/pdf/6-1.pdf <retrieved: Feb 2nd 2012>

[4] lightRadio Portfolio: Technical overview, Alcatel-Lucent white paper. Available: http://www.alcatel-lucent.com/wps/portal/Locator?LMSG_CABINET=Docs_and_Resource_Ctr&LMSG_CONTENT_FILE=White_Papers/LightRadio_WP1_Technical_Overview.pdf&UNIQUE_NAME=&lu_lang_code=en_WW <retrieved: Feb 2nd 2012>

[5] China Mobile Research Institute "C-RAN – Road Towards Green Radio Access Network" in *C-RAN International Workshop*, Beijing, China, Apr. 23rd 2010. Available: http://labs.chinamobile.com/article_download.php?id=63069 <retrieved: Feb 2nd 2012>

[6] 3GPP TS 36.300 v10.4 Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description; Stage 2

[7] 3GPP TS 36.201 v10.0 Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); LTE physical layer; General description

ISBN: 978-1-61208-193-9

Comparison of Steroscopic Display Systems by Subjective Testing

Ondrej Kaller, Libor Bolecek, Martin Slanina, Tomas Kratochvil
Department of Radio Electronics
Brno University of Technology
Brno, Czech Republic
e-mail: xkalle00@stud.feec.vutbr.cz, xbolec01@stud.feec.vutbr.cz, slaninam@feec.vutbr.cz, kratot@feec.vutbr.cz

Abstract — **This short paper compares 3D video image quality and perceived 3D video image depth of three present-day stereoscopic displays for home entertainment. These stereoscopic displays are represented by the commercially available plasma display panel (PDP) with active shutter glasses, digital light processing (DLP) projection also with active shutter glasses and liquid crystal display (LCD) with passive polarization glasses. Subjective tests and assessment of 3D video image quality and stereoscopic effects have been organized with help of 128 respondents in various age categories and 32 various 3D video or image sequences. The paper presents results of subjectively evaluated 3D video image depth and determination of the viewing conditions impact on perceived 3D depth.**

Keywords – stereoscopic display; 3D passive display; 3D polarization display; shutter glasses; circle polarization glasses, 3D video image quality; subjective testing; viewing angles

I. INTRODUCTION

Today’s development of stereoscopic imaging and 3D video image quality evaluation is divided mainly into three branches. Testing methodology for 3D video image quality evaluation is the first one, which tries to define testing conditions and processing of data from subjective tests. These tests could be classical deliberation or behavioral evaluation [1], where the quality and fidelity of 3D imaging is evaluated by biological responses of tested subject. These responses are produced automatically such as postural responses, skin conductance or heart rate. The second way is to find possibilities how to describe projection of objective video image parameters to the space of subjective test results and to define metrics for their evaluation [2]. The third type of contemporary research in this area is focused on Quality of Service (QoS) determination in concrete application such as 3D IPTV [3] or wireless transmission is [4].

This paper describes subjective tests that we have recently organized and brings preliminary and partial results with their discussion. These tests have been quite complex and intended to investigate the subjective 3D video image quality and stereoscopic effect related to different display technologies, content parameters, light conditions, viewing angles and characteristics of respondents. This contribution selects only a fragment of our results. It focuses only on the influence of the viewer position on the Quality of Experience (QoE) for three present-day 3D display technologies.

We have tested three different technologies, currently widespread on the market for 3D home entertainment (Table I). These present Plasma Display Panel (PDP) and Digital Light Processing (DLP) projection, both with active shutter glasses, and Liquid Crystal Display (LCD) with passive polarization glasses. Our analysis aims at the comparison of the technologies in terms of perceived quality of stereoscopic content and in terms of naturalness of the perceived 3D video image depth.

The rest of the paper is organized as follows. Section 2 summarizes relevant information about the mentioned display technologies and it needed for understanding the subjective test adjustment. Section 3 contains description of individual technical equipment used for testing and subjective tests arrangement. It also mentions the parameters of the tested 3D image video content. Next, Section 4 provides some information about our respondents and also shows the results. Finally, in Section 5, the results are discussed and a brief outlook for the future work is given.

II. 3D DISPLAY TECHNOLOGY

Time multiplexing is the most extended technique for stereo pair discrimination today. The display itself can be a classical 2D panel with higher video frame rate. The most important part of the system is synchronization of active shutter glasses, which switch light sequentially in time multiplex to both eyes. This approach has theoretically limitless horizontal and vertical viewing angles, in fact limited just by the display itself. However, in reality, the manufacturers admit some limitation in light separation, because of the directional characteristic of active LCD glasses [5] [6].

For this purpose, it seems very convenient to use a PDP, which has in principle no problem with fast refresh rate (e.g., 0,001 ms - Panasonic TX-P42GT20E [6]). That is because the gas discharge ignition is practically immediate. In spite of this potential parameter, the current systems use only 120 Hz frame rate for 3D. It leads to 60 frames per eye, which is a lower rate compared to what the classical 2D systems use today. Also it is not a problem for concurrent products based on LCD panels to reach the same parameters.

Due to its simple configuration, the DLP technology is widely used for home 3D projection with time multiplexing of both halves of stereo pair. DLP technology can reach higher frame rate in comparison with LCD projectors. Frame rate is the same (low) as in previous system (e.g., BenQ W710ST, frame rate 119 Hz), but we can remark, that it is going to be increased in next generation of projectors.

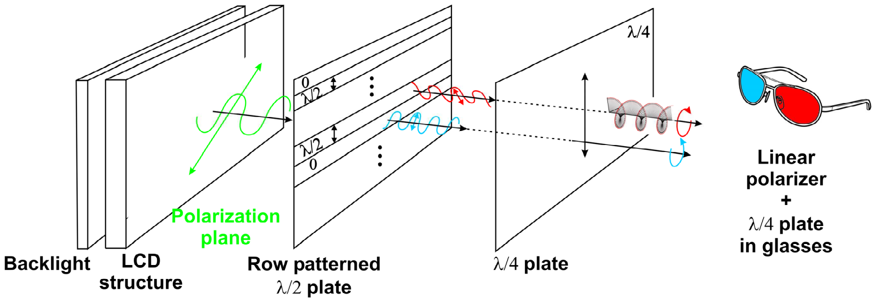


Figure 1. Principle of polarizated discrimination display demonstrated on cross section.

Generally, a characteristic property of projectors is the higher diagonal dimension of the image.

In case of 3D utilization, it brings a higher parallax and consequently larger 3D effect in the same viewing conditions. This "advantage" should also cause problems, because the available content is usually calculated for smaller diagonal dimension and due to adaptive parallax [7] the 3D effect can be higher and perceptual depth can leave the comfortable zone. This could cause so called "dizziness".

Besides time multiplexing, the new implementation of the old known polarization technology celebrates success at present. Its novelty lies in using a patented system for circular light polarization called Film-type Patterned Retarder (FPR), which decreases production costs. Demonstrational cross section is shown in Fig. 1 [5]. Unpolarized light from Cold Cathode Fluorescent Lamp (CCFL) or Light Emitting Diode (LED) is in principle linearly polarized in the system of LCD panel. A half-wave row slice structure rotates the light polarization plane by π/2 radians in case of odd rows of image. In this plane of the structure, the information for the left and right eye is separated spatially and by linear polarization. Circular polarization, which is used for the intra-eye crosstalk minimization, is then obtained by quarter-wave plate.

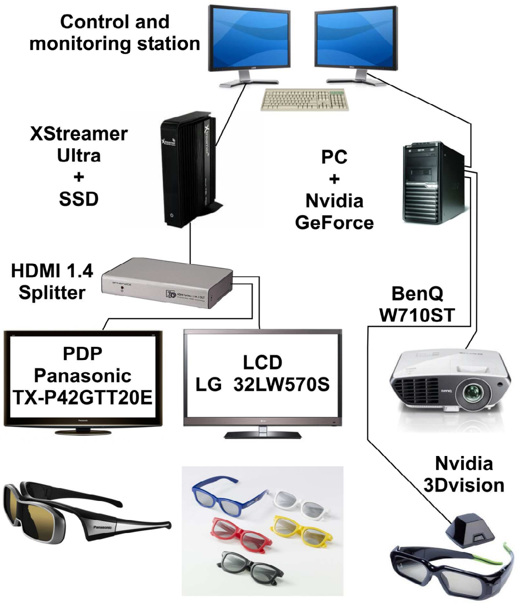


Figure 2. Illustrative scheme of laboratory arrangement for testing.

Separation of polarized light in glasses uses a reverse mechanism [8]. One advantage of passive system is of course the weight of glasses, which achieves 15 g in case of polarized discrimination glasses (LG) compared to 50 g or 28 g for active shutter glasses (NVIDIA and Panasonic, respectively). The design of Panasonic glasses has been criticized for wearing discomfort by respondents.

III. SUBJECTIVE TESTING

A. *Technical Equipment*

The laboratory equipment (Fig. 2) consists of two sources of 3D video signal, HDMI 1.4 splitter, 3 different stereoscopic displays (Table I) and control and monitoring displays.

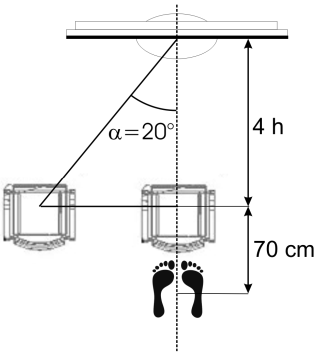


Figure 3. Floor projection of test arrangement.

TABLE I. PARAMETERS OF THE TESTED 3D SYSTEMS

3D System	Display Technology	Stereo Pair Discrimination	Native Displayed Resolution in 3D	Diagonal [cm]
LG 32LW570S	LCD	Polarization	960 x 540	82
BenQ W710ST (NVIDIA)	DLP	Time multiplex	960 x 720	196
Panasonic TX-P42GT20E	PDP	Time multiplex	960 x 1080	106

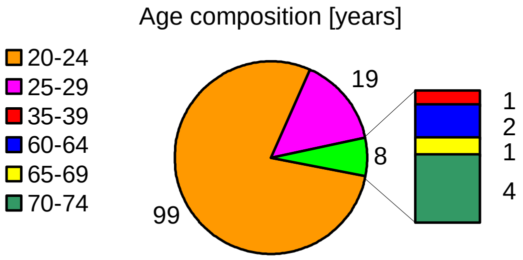


Figure 4. Age composition of the respondents.

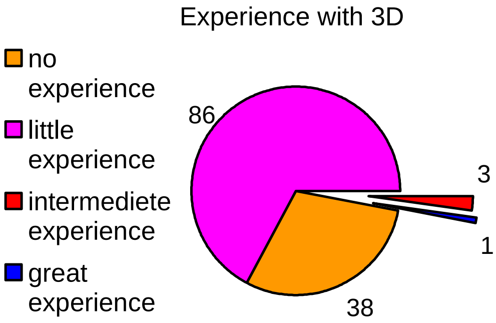


Figure 5. Previous experience with 3D television.

Two independent signal sources were used, because it is necessary to generate separate video signals. One of them is time multiplexed for 3D projector – a PC with graphic card NVIDIA GeForce 8000 has been used. To control the other two displays, the home theater PC XStreamer Ultra with built-in SSD hard drive has been used.

B. Laboratory Arrangement

Fig. 3 shows a floor projection of the testing site. In all three cases, the same viewing conditions have been defined, especially the horizontal and vertical viewing angles.

The viewing distance has been calculated as four times of the picture height (4h) in case of ideal viewing condition [9]. While horizontal angular displacement α has been set directly by the seat position, vertical displacement β depends on the tested subject height. We have asked for it in the evaluation form.

The average height has been 180.2 cm (values from 166 to 196 cm). These values lead to the mean vertical viewing angle β of 14° (from 11° to 18°) in case of LCD display, assuming the average distance of eyes (optical axis) from the top of the head is 12 cm. For PDP, the same value was 11° (from 8° to 14°). For projection, the mean vertical angle was 2.5° (from 0.5° to 4.5°) in case of positive vertical displacement (standing observers) and 3.2° (from 1.5 to 4.3) in case of negative vertical displacement.

C. 3D Video Image Content

We have utilized three sources of content to evoke impression of standard home usage: Blu-ray disc, 3D satellite digital television broadcasting and amateur 3D camera captured sequences. Four sets of sequences have been prepared for testing. Three from the previously mentioned content types with an additional set containing static images, created from the sequences in the other sets.

The sources used had variable native 3D format, compression and resolution. The original content was coded with Multiview Video Coding (MVC) format (for Blu-ray sequences), spatially compressed side-by-side and H.264 encoded (for satellite television broadcasting) and spatially compressed and Advanced Video Coding High Definition (AVCHD) encoded (for amateur capture).

For playback during the test session, we have converted all the sequences to spatially compressed (side by side) “Full HD” format 1920x1080/25p. The sequences were stored in YUV raw video format and played back with no compression. Native pixel resolution in Table I is calculated for one half of stereo pair of this input format.

D. Test Session

Structure of each set is done according to ITU recommendation ITU-R BT.710 [9], where 10 to 15 s of video sequence/static image is followed by 5 s of mid gray color. The sequences were played back in random order, 8 sequences per format, resulting in the total of 32 sequences.

After viewing the set of sequences on one 3D system, the observers were asked to perform the test on another system. The order of the sequences was different for this following test. We repeated the same procedure for all the three 3D systems under test. At the end of the subjective test session, the observers were asked to fill in a simple questionnaire including several personal questions and an overall judgment comparing the three systems in terms of 3D effect quality and depth naturalness. The task was to select which of the three systems performed 3D display and according QoE the best [10].

E. Observers

We have tested a sample of 128 people of age between 20 and 74 years (Fig. 4). The yield of test has been over 95 %. In total 74 respondents compared all three 3D technologies in their subjective tests.

We have not done any training of our respondents; we have just allowed them some time to read the test form. The tested sample of people has consisted mainly of students (93%), which have no experience with video image and multimedia subjective quality tests at all.

At the beginning of the test we have asked the subjects about personal information, including gender, age and employment. We have investigated for how long they watch TV per week, whether they suffer from eye defect and what their experience with 3D technology is. We have asked in particular about 3D home television/cinema systems, not about 3D cinema (Fig. 5). Scale has been four-level, where intermediate/great experience means, that the subject owns a 3D display and watches 3D content sporadically/regularly.

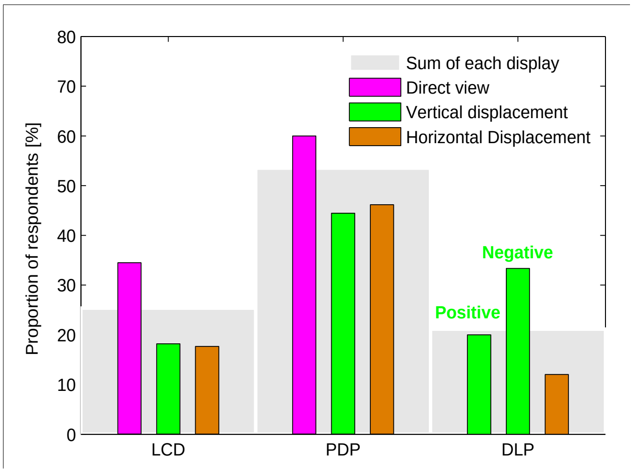


Figure 6. The best 3D video image quality in dependence on 3D system and actual viewing position of the respondent.

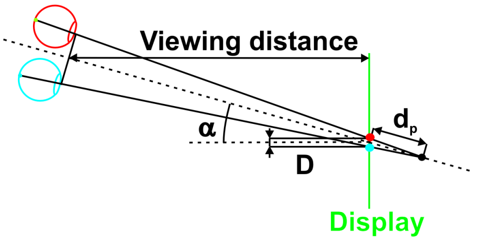


Figure 7. Illustration for the perceived depth ratio evaluation.

The eye defects we distinguish among corrected ones as myopia/astigmatism, in that case were questioner classically processed, then serious defects as amblyopia, or disorder of spatial perception, where respondents has been discarded and daltonism, where only questions about perceived 3D video image depth has been taken into account when processing the results.

IV. Results

In this paper, the answers for two questions from our complex questionnaire are only presented, associated with the technological aspects of the used 3D displays.

A. *Perceived Quality*

The first test question deals with 3D video image quality evaluation (Fig. 6). Percentage of respondents evaluating video image quality of a particular system as the best is shown by a gray bar. The colored columns show the proportion of respondents, who decided for a concrete horizontal or vertical viewing angle. This percentage is calculated among the respondents, who participated in the subjective test in a particular position [7].

The best 3D video image quality is given by, according to test results, the PDP system. This fact could be associated with the highest native resolution of the one displayed image from the stereo pair (Table I).

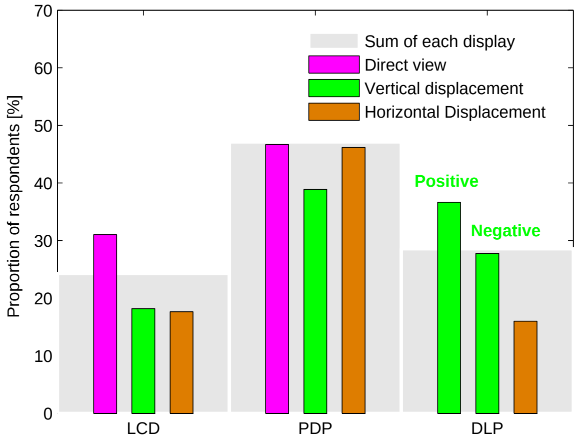


Figure 8. The highest 3D depth perception in dependence on 3D system and actual viewing position of the respondent.

TABLE II. Theoretical Percieved Depth Ratio for Various Viewing Conditions

Technology	Viewing position	α [°]	β [°]	Perceived depth ratio
Polarization LCD	Horizontal displacement	20	0	**1.00**
	Direct View	0	0	**1.02**
	Vertical displacement	0	14,3	**1.03**
Time-multiplex PDP	Horizontal displacement	20	0	**1.47**
	Direct View	0	0	**1.53**
	Vertical displacement	0	10,9	**1.53**
Time-multiplex DLP	Horizontal displacement	20	0	**5.66**
	Vertical displacement +	0	2.5	**6.67**
	Vertical displacement -	0	3.2	**6.67**

B. *Perceived Depth*

The second test question discussed in this paper is which system and configuration provides the highest depth perception. The answer for this could be, unlike the previous, theoretically calculated from the known viewing position under the condition that the same 3D content is displayed on the compared displays (Table II).

The perceived depth ratio, which we have defined for this purpose as our own and original measure, gives an objective comparison of the 3D systems under given observation conditions. It is computed as follows (Fig. 7). At first, the perceived distance (d_p) of static stereoscopic parallax is calculated. The value depends on pixel disparity (D), defined viewing distance (4h) and horizontal displacement (α). The pixel disparity increases with the 3D display diagonal size. Perceived depth ratio is then defined as normalized value of the d_p to the viewing distance. From the results, the rows and their order in the Table II show, that the DLP projection should provide the highest perceived depth and the best stereoscopic effect to the viewer.

Unfortunately, the described calculation is in contradiction with the test results (Fig. 8). The tested subjects consider PDP depth perception the highest. One of the aspects, where test results correspond to theoretical computations, is the lowest depth of LCD polarization system. How to explain the general difference? One hypothesis says that the stereo effect of DLP system may be so strong, that the brain of some part of respondents can not process it. We may also suppose that level of light, which has been changed during the subjective test, but intentionally not discussed in this paper, degrades the results of DLP system. Lighting conditions influences the quality of experience for sure and they are important topic for forthcoming investigation. In fact, variety of illumination during the test was set from 10 lx to 500 lx, but light conditions have not been strictly complied with ITU recommendation [9]. They have been specified as a most common and comparable with home environment and scenario.

V. Conclusions and Future Work

In this short paper, we presented a comparison of performance of present-day stereoscopic display systems by subjective testing. The aim of this short paper was not to bring complete study and present all the results from our subjective tests, but only describe the present commercial 3D display technologies and then our methods, technical equipment, laboratory arrangement, definition of 3D video image content and group of observers. The results are very brief and evaluate just answers to the two questions from our complex questionnaire for the 3D video image quality and its subjective evaluation related to QoE.

Within the evaluation and subjective testing of 3D systems that was discussed in this paper, we have also measured some objective parameters of the individual displays. The technological limitations of the used 3D systems were taken, such as the maximum useable displacement or crosstalk between the halves of stereo pair.

The aim of our future work is to find and quantify all technological aspects of 3D video image quality and image depth to improve these parameters. Of course, our findings could have some discrepancies with theory and the data will be statistically processed in a more complex manner to find hidden dependences.

Acknowledgment

This paper was jointly supported by the grant project of the Czech Science Foundation no. 102/10/1320 “Research and modeling of advanced methods of image quality evaluation (DEIMOS)” and internal project FEKT-S-11-12 MOBYS. The described research was performed in laboratories supported by the SIX project; the registration number CZ.1.05/2.1.00/03.0072, the operational program Research and Development for Innovation. The support of the project CZ.1.07/2.3.00/20.0007 WICOMT, financed from the operational program Education for competitiveness, is gratefully acknowledged.

References

[1] W. Ijsselstein, H. de Ridder, J. Freeman, S. E. Avons, and D. Bouwhuis, ”Effect of stereoscopic presentation, image motion, and screen size on subjective and objective corroborative measures of presence,” Presence, vol. 10, no. 3, June 2001, pp. 298–311.

[2] A. Benoit, P. Le Callet, P. Campisi, and R. Cousseau, “Quality assessment of stereoscopic images,” EURASIP Journal on Image and Video Processing, vol. 2008, Article ID 659024, 2008, pp. 1-13, doi:10.1155/2008/659024.

[3] M. Barkuowsky, K. Wang, R. Cousseau, K. Brunnstrom, R. Ollson, and P. Le Callet, “Subjective quality assesment for error concealment strategies for 3DTV in the presence of asymmetric transmission errors,” Proceedings of the 2010 IEEE 18th International Packet Video Workshop, Dec. 2010, pp. 193-200, doi: 10.1109/PV.2010.5706838.

[4] S. L. P. Yasakethu, W. A. C. Fernando, B. Kamolrat, and A. Kondoz, ”Analyzing Perceptual Attributes of 3D Video,” IEEE Transaction on Consumer Electronic, vol. 55, no. 2, May 2009, pp. 864-872, doi: 10.1109/TCE.2009.5174467.

[5] H. Urey, Chellappan, K. V., E. Erden, and Surman, P., “State of the art in stereoscopic and autostereoscopic displays,” in Proceedings of the IEEE. Feb. 2011, pp. 540-555, doi: 10.1109/JPROC.2010.1098351

[6] Datasheet and manuals of panasonic PDP panel TX-P42GT20E [14. 2. 2011 available at Panasonic.cz]

[7] K. Ide and T. Sikora, Adaptive parallax for 3D television, in “3DTV-Conference: The TrueVision – Capture”, in Transmission and Display of 3D Video (3DTV-CON). Tampere, 2010. s.1-4. ISBN: 978-1-4244-6377-0

[8] Y. Yosihihara, H. Ujike, and T. Tanabe, “3D Crosstalk of stereoscopic (3D) display using patterned retarder and corresponding glasses”, in IDW ´08, pp.1135-1138.

[9] The ITU Radiocommunication Assembly, “Recommendation ITU-R BT.710-3 Subjective assesment for image quality in high-definition television,” 1997.

[10] J. Gutiérrez, P. Pérez, F. Jaureguizar, J. Cabrera, and N. García, “Subjective assessment of the impact of the of transmission errors in 3DTV compared to HDTV,” 3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON), May 2011, pp. 1-4, doi: 10.1109/3DTV.2011.5877209.

Robust Digital Video Watermarking in the Spatial and Wavelet Domain

Radu Ovidiu Preda, Cristina Oprea, Ionuţ Pirnog, Lucian Andrei Perişoară
Faculty of Electronics, Telecommunications and Information Technology
Politehnica University of Bucharest
Bucharest, Romania
radu@comm.pub.ro, cristina@comm.pub.ro, ionut@comm.pub.ro, lperisoara@yahoo.com

Abstract — **This paper presents two blind video watermarking techniques in the spatial and wavelet domain proposed by the authors and compares the two approaches. The original watermark and the original, unwatermarked videos are not required for the watermark extraction process. The two methods are combinations of spread-spectrum and quantization based techniques. The watermarks used are binary images, containing the copyright information. The watermark is protected against singular bit errors with a Hamming error correction code. The spatial domain technique embeds a watermark bit by spreading it in a luminance block. The actual embedding is done using a quantization based approach. The wavelet based technique embeds the same watermark bit into a number of chosen detail wavelet coefficients of the middle wavelet sub-band. The resilience of the schemes is improved by redundantly embedding the same watermark in a number of video frames. We have tested the perceptual quality of the watermarked videos and the resilience of our schemes to eight different attacks in the spatial, temporal and compressed domain, for different quantization step sizes and different number of redundant frames. The test results show that our wavelet domain technique achieves better video quality and robustness to attacks than the spatial domain method.**

Keywords - Digital Video Watermarking; Copyright Protection; Spatial Domain; Wavelet Domain; Comparison; Perceptual Quality; Robustness to Attacks.

I. INTRODUCTION

Video watermarking techniques are characterized by the domain that the watermark is being embedded or detected, their capacity, the perceptual quality of the watermarked videos and their robustness to particular types of attacks. They can be divided into three main groups according to the domain, in which the watermark is embedded and extracted: spatial domain, frequency domain and compressed domain watermarking. We will focus here on spatial and frequency domain watermarking.

Spatial domain algorithms embed the watermark into the pixel values and no transforms are applied to the host signal during the embedding process. The most common techniques to insert the watermark into the host data in the spatial domain is via Least Significant Bit (LSB) modification, Spread Spectrum Modulation (SS) and Quantization Index Modulation (QIM). The LSB techniques are not robust to attacks because the LSB plane can be easily replaced by random bits, removing the watermark.

Spread spectrum methods view watermarking as a problem of communication through a noisy channel. As a means to combating this noise or interference, spread-spectrum techniques are employed to allow reliable communication in such noisy environments. In this case, the watermark data is coded with a pseudorandom code sequence to spread its power spectrum in the image or video, thus increasing its robustness to attacks. One of the first methods was the one-dimensional spread spectrum approach [1]. Here, the watermark is a pseudo-random sequence spread over the video frames by direct spatial domain addition. The watermark is repeatedly embedded throughout the video in a sequential manner. Other more complicated spread-spectrum methods were proposed in [2][3].

Quantization Index Modulation (QIM) refers to a class of data hiding schemes that exploit Costa's [4] findings by embedding information in the choice of quantizers. Over the past few years, QIM-based data hiding has received increasing attention from the data hiding community because it is more robust than techniques such as spread spectrum and LSB modification. State of the art proposed QIM schemes include Chen and Wornell's QIM and dither modulation [5], Eggers et al's scalar Costa scheme (SCS) [6], Jie and Zhiqiang's color image QIM scheme [7] and Kalantari and Ahadi's logarithmic QIM scheme [8].

In frequency domain watermarking, the most common transforms being used are the Discrete Cosine Transform (DCT), Discrete Fourier Transform (DFT) and Discrete Wavelet Transform. The main advantage offered by transform domain techniques is that they can take advantage of special properties of the alternate domains to address the limitations of pixel-based methods or to support additional features. Also, they have better resistance to compression based attacks. Generally, the main drawback of transform domain methods is their higher computational requirements.

Lately, algorithms in the Wavelet domain have gained more popularity due to their excellent spatial localization, frequency spread, and multi-resolution characteristics [9-14].

A lot of research has been done lately in developing new and improved watermarking techniques, but there is a difficulty in comparing the research results, because independent researchers use very different watermarks, watermark capacity, test videos, parameters for watermark embedding and extraction and attacks with different parameters to test the robustness of their schemes. There is a need to compare the watermarking methods in different domains. Our paper addresses this issue by proposing two approaches in the spatial and Wavelet domain that have similar specifications, like watermark, watermark capacity, test videos, attacks with the same parameters. Both approaches embed the same watermark (binary image) with spatial and temporal redundancy and use a blind method for watermark extraction.

The rest of this paper is organized as follows: Sections II and III describe the proposed video watermarking techniques in the spatial and DWT domain, respectively, providing detailed diagrams and description of the watermark embedding and extraction strategies. Section IV contains the experimental results and a detailed comparison of the proposed methods in terms of perceptual quality and robustness to different attacks. Finally, Section V presents the conclusions of our work.

II. THE PROPOSED VIDEO WATERMARKING SCHEME IN THE SPATIAL DOMAIN

The watermark embedding process, illustrated in Fig. 1, is described in the following:

First, the original video is partitioned into groups of k frames. Every frame of the group is converted to the YC$_b$C$_r$ color space.

The binary image matrix is transformed into a binary row vector w of size $P = h \times v$. To protect the watermark against bit errors, a Hamming error correction code (m,n) with codeword length of m bits and data-word length of n bits is applied to the vector w. The size of the resulting watermark vector w_c is:

$$P^{'} = P\frac{m}{n} \tag{1}$$

The binary sequence w_c is partitioned into a number of $\frac{F}{k}$ sequences $w_c(j)$ of size $P'\frac{k}{F}$, where $j = \overline{1, \frac{F}{k}}$, F is the number of frames of the video and k is the number of redundant frames. The dimensions h and v of the watermark are chosen so that $P'\frac{k}{F}$ is an integer. The same sequence $w_c(j)$ will be inserted into every frame of a group j of k frames.

The size l of a square bloc of $l \times l$ luminance values is calculated in (2) to embed a bit of the watermark:

$$l = \left[\sqrt{\frac{MNC}{P'k}}\right], \tag{2}$$

where [.] is the integer part operator.

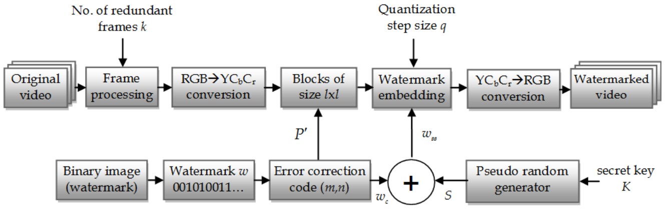


Figure 1. Block diagram of the spatial watermark encoder

A spread-spectrum technique is used to spread the power spectrum of the watermark data, thus, increasing its robustness against attacks. First a binary pseudo-random sequence $S = \{s_r \,|\, s_r \in \{0,1\}, r = 1,...,l^2\}$ of size l^2 with equal number of zeros and ones is generated using the Mersenne-Twister algorithm proposed in [15] with the use of the last 64 bits of the secret key K as seed for the generator. This method generates numbers with a period of $(2^{19937} - 1)/2$. For every bit of the watermark $w_c(j)$, the corresponding spread spectrum sequence is:

$$w_{ss} = \begin{cases} [s_1, s_2, ..., s_{l^2}], \text{ if } w_c = 0 \\ [\overline{s}_1, \overline{s}_2, ..., \overline{s}_{l^2}], \text{ if } w_c = 1 \end{cases} \tag{3}$$

A sequence S (representing one bit of the original watermark) is embedded in every bloc of $l \times l$ luminance values. A bit of S is embedded into the luminance value of the pixel of the same index by rounding its value to an even or odd quantization level. Rounding to an even quantization level embeds a “0”, while rounding to an odd quantization level embeds a “1”, as shown in (4):

$$L_w(i,j) = \left[\frac{L}{2q}\right] \cdot 2q + q \cdot w \cdot sign\left(L(i,j) - \left[\frac{L(i,j)}{2q}\right] \cdot 2q \right), \tag{4}$$

where $L(i,j)$ is the original luminance value, $L_w(i,j)$ is the watermarked luminance value, q is the quantization step size and $sign$() is defined as:

$$sign(x) = \begin{cases} -1, & if \;\; x \le 0 \\ 1, & if \;\; x > 0 \end{cases} \tag{5}$$

The video is converted back to the RGB format, obtaining the watermark video.

The choice of the quantization step size q is a tradeoff between the perceptual quality of the watermarked video (q must have a small value) and the resilience of the watermarking scheme to attacks (q must have a big value).

The watermark extraction process, shown in Fig. 2, implies the following steps:

The watermarked video is partitioned into groups of k frames. Every frame of the group is converted to the YC$_b$C$_r$ format. Every luminance frame is partitioned into square blocks of $l \times l$ luminance values. A bit of the spread spectrum sequence w_{ss}' of size l^2 is extracted from every luminance value of a block of size $l \times l$ using (6):

$$w' = \text{mod}\,2\left(round\left(\frac{L_w(i,j)}{q} \right) \right), \tag{6}$$

where w' is the extracted watermark bit, $L_w(i,j)$ is the luminance value of the pixel at position (i,j), q is the quantization step size and *mod2* is the modulo2 function.

Using the 64 bit seed from the secret key K the binary sequence S is generated locally. The extracted watermark bit for the corresponding block is:

$$w_b' = \begin{cases} 0, & \text{if } \sum_{r=1}^{l^2} \left| w_{ss,r}' - s_r \right| \le \frac{l^2}{2} \\ 1, & \text{if } \sum_{r=1}^{l^2} \left| w_{ss,r}' - s_r \right| > \frac{l^2}{2} \end{cases} \tag{7}$$

A binary sequence $w'_{c,i}(j)$ is extracted from every frame of a group of k frames, where $i = \overline{1,k}$. The sequence $w'_c(j)$ is computed from $w'_{c,i}(j)$ using (8):

$$w'_c(j) = \begin{cases} 0, & \text{if } \sum_{i=1}^{k} w'_{c,i}(j) \le \frac{k}{2} \\ 1, & \text{if } \sum_{i=1}^{k} w'_{c,i}(j) > \frac{k}{2} \end{cases}, \quad j \in \{1,2,\ldots,P'\} \tag{8}$$

The resulting watermark bit stream w'_c of size *P'* is error corrected and the watermark w' of size P is obtained. The extracted binary image is obtained by reshaping the vector w' to a matrix of size $h \times v$.

The choice of the quantization step size q is a tradeoff between the perceptual quality of the watermarked video (q should have a small value) and the resilience of the watermarking scheme to attacks (q should have a big value).

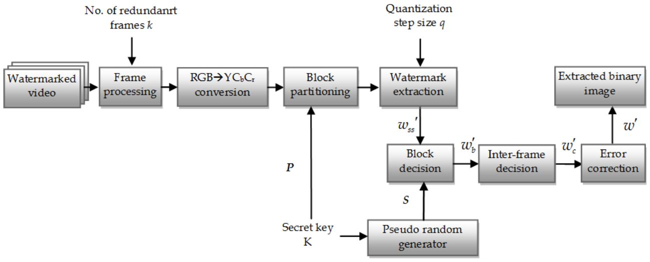


Figure 2. Block diagram of the spatial watermark decoder

III. The Proposed Video Watermarking Scheme in the Wavelet Domain

The watermark is embedded in the selected wavelet coefficients of the luminance Y of every frame of the video. The wavelet decomposition of the luminance is done using the 2D Discrete Wavelet Transform. We have chosen a Wavelet decomposition on L=3 resolution levels. The watermark is embedded in the wavelet coefficients of the LH, HL and HH sub-bands of the second Wavelet decomposition level. The choice of the second decomposition level is a tradeoff between the invisibility of the watermark and the resilience to attacks. A watermark embedded in the wavelet coefficients of the LH_1, HL_1 and HH_1 sub-bands is very sensitive to attacks, because these sub-bands contain the finest details of the frame. On the other hand, if we embed the watermark in the LH_3, HL_3 and HH_3 sub-bands, the perceptual quality of the video will be significantly altered. For these reasons, the best choice for watermark embedding is the second wavelet decomposition level.

For videos of resolution $M \times N$, the number of selected wavelet coefficients for a frame is:

$$C = 3\frac{MN}{2^{2(L-1)}} \tag{9}$$

The maximum capacity of the watermarking scheme is $C' = FC$ where F is the number of video frames and can be achieved by embedding a watermark bit in every selected wavelet coefficient. For example, for CIF videos of resolution 352x288 and 30 frames/s, the maximum capacity is 556kb/s. This maximum capacity is not needed in most applications, thus we will reduce it to improve the robustness of the scheme. Fig. 3 shows the block diagram of our Wavelet based watermark embedding scheme and is described in the following steps:

The binary image matrix is transformed into a binary row vector w of size $P = h \times v$. To protect the watermark against bit errors, a Hamming error correction code with codeword length of m bits and data word length of n bits is applied to vector w, resulting in a watermark vector w' of size P'.

A same spread-spectrum technique is used to spread the power spectrum of the watermark data. First the binary sequence $S = \left\{ s_j \middle| s_j \in \{0,1\}, j = 0,1,\ldots,G \right\}$ with equal number of zeros and ones is generated using the Mersenne-Twister algorithm with the use of 64 bits of the secret key K as seed for the generator. For every bit of the watermark *w'*, the corresponding spread spectrum sequence is:

$$w_{ss}(i) = \begin{cases} [s_1, s_2, \ldots, s_G], & \textit{if } w'(i) = 0 \\ [\overline{s}_1, \overline{s}_2, \ldots, \overline{s}_G], & \textit{if } w'(i) = 1 \end{cases}, \quad i = 1, \ldots, P' \tag{10}$$

Every sequence $w_{ss}(i)$ (representing one bit of the original watermark) is embedded into a number G of wavelet coefficients, every bit of $w_{ss}(i)$ in a wavelet coefficient.

ISBN: 978-1-61208-193-9

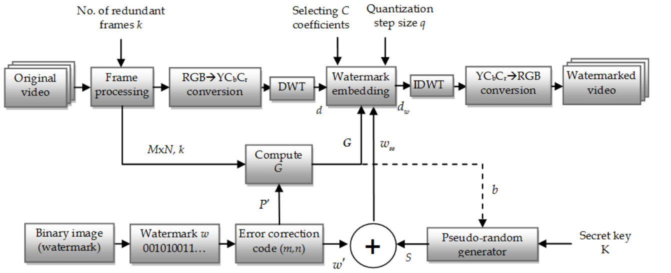


Figure 3. Block diagram of the wavelet watermark encoder

The number G depends on the number C of the selected wavelet coefficients, the number of frames F of the original video and the size P' of the watermark:

$$G = \left[\frac{C \cdot F}{P'} \right] \tag{11}$$

A bit of the binary sequence S is embedded in the selected wavelet coefficient by rounding its value to an even or odd quantization level. Rounding to an even quantization level embeds a “0”, while rounding to an odd quantization level embeds a “1”, as shown in (12):

$$d_w = \left[\frac{d}{2q} \right] \cdot 2q + q \cdot w \cdot sign\left(d \; - \left[\frac{d}{2q} \right] \cdot 2q \right), \tag{12}$$

where d is the original wavelet coefficient, d_w is the watermarked wavelet coefficient and q is the quantization step size.

After the entire watermark has been embedded, the 2D Inverse Discrete Wavelet Transform is computed for every frame to obtain the watermarked video.

The watermark extraction process, shown in Fig. 4, is explained in the following:

First, the wavelet decomposition of the watermarked, possibly attacked video is performed, then the wavelet coefficients used for embedding are selected. Parameter G is computed using the information about the size of the watermark provided by the secret key K. From every selected coefficient a bit is extracted according to (13), resulting in a sequence $w'_{ss}(j)$ of G bits from every group.

$$w' = \mathrm{mod}_2\left(round\left(\frac{d_w}{q} \right) \right), \tag{13}$$

where d_w is the watermarked wavelet coefficient.

Using the 64 bit seed from the secret key K the binary sequence S of size G is generated. The extracted watermark bit $w''(i)$ corresponding to a group of G wavelet coefficients is computed in (14).

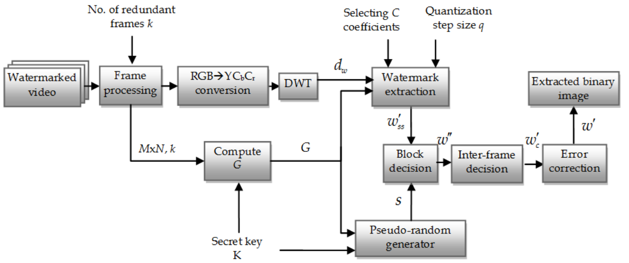


Figure 4. Block diagram of the wavelet watermark decoder

$$w''(i) = \begin{cases} 0, if \ \sum_{j=1}^{G} [w'_j(i) - s_j] \le \frac{G}{2} \\ 1, if \ \sum_{j=1}^{G} [w'_j(i) - s_j] > \frac{G}{2} \end{cases}, \; i = 1, ..., P' \tag{14}$$

The resulting watermark bit stream of size P' is error corrected and the watermark w' of size P is obtained. The extracted binary image is obtained by reshaping the vector w' to a matrix of size $h \times v$.

To improve the resilience of the algorithm against temporal attacks we embedded the same watermark redundantly in every k frames. Thus, the number of wavelet coefficients used for embedding a watermark bit is decreased from G to G/k.

IV. Comparison of the Proposed Techniques

Our algorithms were tested using the first 27 frames of the videos “stefan”, “forman” and “bus” in RGB uncompressed avi format, of resolution 352x288, 24 bits/pixel and frame rate of 30 frames/s. The binary image used as watermark is a copyright logo containing the name of one of the authors. The resolution of the image depends on the error correction code used, the number of redundant frames and the resolution of the initial video. The size of the watermark used is rather big, in order to better compare the two approaches. Using watermarks with smaller payload would improve the robustness, with BER values very close to zero for both methods, making it harder to compare them.

We have conducted the experiments for both methods using the quantization step sizes $q = 2$, $q = 4$, embedding of the same watermark in $k = 3$ and $k = 9$ frames, without using an error correction code and using a Hamming (7,4) error correction code.

To compare the perceptual quality of the watermarked video with the original one, we have computed the mean Peak Signal to Noise Ration (PSNR) of all frames of the video. The PSNR results are shown in Fig. 5. We can see that the best quality is obtained using the Wavelet approach. The PSNR results for the spatial watermarking scheme are quite low for quantization with bigger quantization step sizes (for $q = 4$ and $q = 8$ below the accepted value of 40 dB). For $q = 8$ only the wavelet based technique achieves a PSNR value higher than 40 dB.

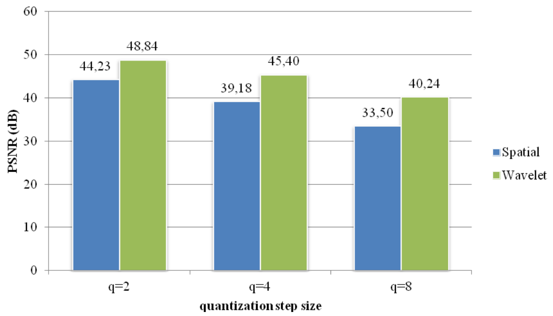


Figure 5. PSNR values for the proposed methods for different quantization step sizes

Next, we wanted to test the robustness of the proposed watermarking schemes. For this purpose we have carried out a range of eight attacks on the watermarked videos: (a) blurring of 2x2 pixel blocks, (b) brightening, adding *Y*=6 to the luminance of every pixel, (c) addition of Gaussian noise with mean 0 and variance 0,0003, (d) median filtering using a 3x3 pixel neighborhood, (e) addition of "salt and pepper" noise with density 0,3%, (f) frame averaging of 20% of the frames, where the current frame is the mean of the previous, current and next frame of the video, (g) JPEG compression of every frame using a quality factor Q=60 and (h) MPEG-2 compression at 4 and 2 Mbps. The parameters of the attacks were chosen in such a manner, that the visual degradation of the attacked videos is acceptable, because, by attacking a watermarked video, an attacker wants to destroy the watermark, but not the video quality.

To evaluate the robustness objectively, we have calculated the mean values of the decoding BER for the watermarks extracted from all test videos after they were attacked and plotted 6 different graphs (Fig. 6 - 8), where we represent the mean decoding BER for every method and every attack. The variables are the quantization step size *q* (chosen 2, 4 and 8) and the number of frames *k* used for embedding the same watermark (chosen 3 and 9). For $q = 2$ no error correction code was used, because the corresponding BER values are quite high and the Hamming (7,4) error correction would not work for such high bit error rates. For $q = 4$ and $q = 8$, with lower BER values, we used the Hamming (7,4) code, which can correct single bit errors.

The method working in the spatial domain is vulnerable to the brightening attack. For example by adding *Y*=6 to every luminance value, the decoding BER is 100% for every combination of parameters. We didn't represent this value on the graphs, because we didn't want to scale all BER values to 100%. On the other hand, the spatial embedding method has the best resilience to median filtering attacks. The weakness of the wavelet-based method to 3x3 median filtering can be improved by embedding the watermark in the third level wavelet subbands instead of the second. Because of the lower computational complexity, the spatial method could be used for real time processing.

The best overall resilience is achieved by the method working in the wavelet domain, with perfect decoding of the watermark for $q = 8$, $k = 9$ and Hamming (7,4) error correction.

V. Conclusion

In this paper we have compared our two proposed, blind video watermarking techniques in the spatial and wavelet domain. The original watermark and the original, unwatermarked videos are not required for the watermark extraction process. The methods are combinations of spread-spectrum and quantization based techniques. The watermarks used are binary images, containing the copyright information. The watermark is protected against singular bit errors using a Hamming error correction code.

The spatial domain technique embeds a watermark bit by spreading it in a luminance block. The actual embedding into a luminance value is done using a quantization based approach. The wavelet based technique embeds the same watermark bit into a number of detail wavelet coefficients of the middle wavelet sub-bands.

The resilience of the schemes is improved by redundantly embedding the same watermark in a number of *k* video frames. We have tested the perceptual quality of the watermarked videos and the resilience of the schemes to eight different attacks in the spatial, temporal and compressed domain, for different quantization step sizes and different number of redundant frames.

The experimental results show, that the wavelet domain technique achieves better video quality and better robustness to most attacks. The spatial domain method is most vulnerable to the brightening attack. The wavelet based technique achieves very good overall scores, being the better candidate for robust video watermarking.

References

[1] F. Hartung and B. Girod, "Watermarking of uncompressed and compressed video", Signal Processing, vol. 66, no. 3, pp. 283–301, May 1998.

[2] H. O. Altun, A. Orsdemir, G. Sharma, and M. F. Bocko, "Optimal Spread Spectrum Watermark Embedding via a Multistep Feasibility Formulation", IEEE Transactions on Image Processing, vol. 18, no. 2, pp. 371-387, Feb. 2009.

[3] S. P. Maity and S. Maity, "Multistage Spread Spectrum Watermark Detection Technique Using Fuzzy Logic", IEEE Signal Processing Letters, vol. 16, no. 4, pp. 245-248, April 2009.

[4] M. H. M. Costa, "Writing on dirty paper", IEEE Transactions on Information Theory, vol. IT-29, no. 3, pp. 439–441, May 1983.

[5] B. Chen and G. W. Wornell, "Quantization index modulation: A class of provably good methods for digital watermarking and information embedding", IEEE Transactions on Information Theory, vol. 47, pp. 1423–1443, May 2001.

[6] J. J. Eggers, R. Bauml, R. Tzschoppe, and B. Girod, "Scalar Costa scheme for information embedding", IEEE Transactions On Signal Processing, vol. 51, no. 4, pp. 1003–1019, 2003.

[7] N. Jie and W. Zhiqiang, "A new public watermarking algorithm for RGB color image based on Quantization Index

Modulation”, International Conference on Information and Automation, ICIA '09, pp. 837-841, June 2009.

[8] N.K. Kalantari and S.M. Ahadi, “A Logarithmic Quantization Index Modulation for Perceptually Better Data Hiding”, IEEE Transactions on Image Processing, vol. 19, no. 6, pp. 1504-1517, June 2010.

[9] M. Barni, F. Bartolini, and A. Piva, “Improved wavelet-based watermarking through pixel-wise masking”, IEEE Transactions on Image Processing, vol. 10, no. 5, pp. 783-791, 2001.

[10] D. Zou, Y.Q. Shi, Z. Ni, and W. Su, “A Semi-Fragile Lossless Digital Watermarking Scheme Based on Integer Wavelet Transform”, IEEE Transactions on Circuits and Systems for Video Technology, vol. 16, no. 10, pp. 1294-1300, Oct. 2006.

[11] G. S. El-Taweel, H. M. Onsi, M.Samy, and M.G. Darwish, “Secure and Non-Blind Watermarking Scheme for Color Images Based on DWT”, GVIP Special Issue on Watermarking, 2007.

[12] L.E. Coria, M. R. Pickering, P. Nasiopoulos, and R. K. Ward, “A Video Watermarking Scheme Based on the Dual-Tree Complex Wavelet Transform”, IEEE Transactions on Information Forensics and Security, vol. 3, no. 3, pp. 466-474, Sept. 2008.

[13] R. O. Preda and N. Vizireanu, “Robust wavelet-based video watermarking scheme for copyright protection using the human visual system”, Journal of Electronic Imaging 20, 013022, 2011.

[14] R. O. Preda and N. Vizireanu, “Quantization-based video watermarking in the wavelet domain with spatial and temporal redundancy”, International Journal of Electronics, Vol. 98, Issue 3, pp. 393-405, 2011.

[15] M. Matsumoto and T. Nishimura, “Mersenne Twister: A 623-Dimensionally Equidistributed Uniform Pseudorandom Number Generator”, ACM Transactions on Modeling and Computer Simulation, vol. 8, no. 1, pp. 3-30, 1998.

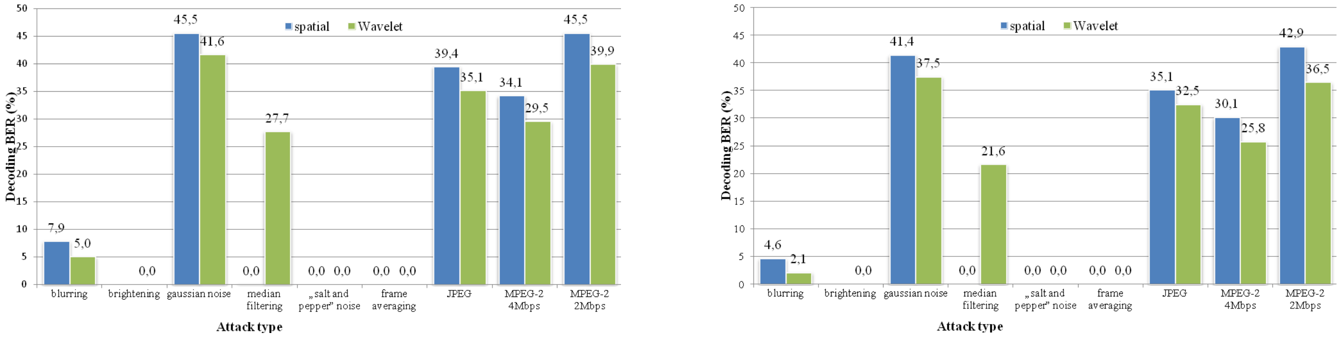


Figure 6. Decoding BER (%) for the proposed methods using $q = 2$, no error correction code and (a) k=3 and (b) k=9 redundant frames

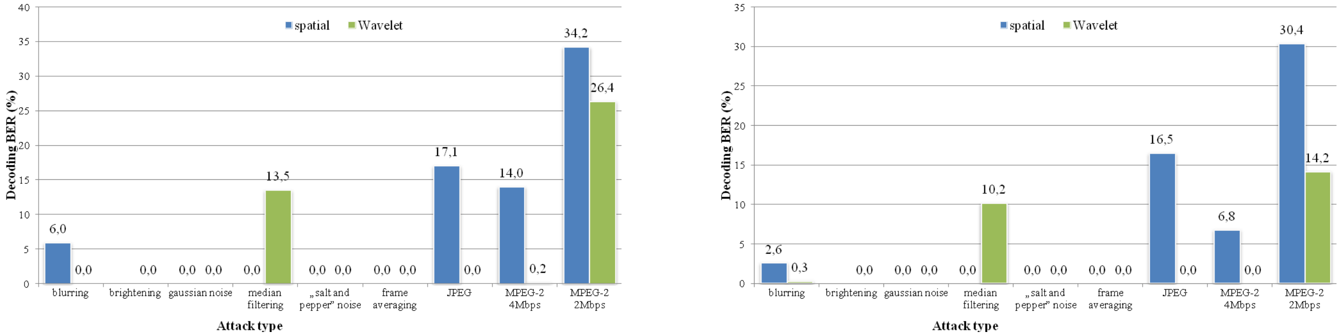


Figure 7. Decoding BER (%) for the proposed methods using $q = 4$, no error correction code and (a) k=3 and (b) k=9 redundant frames

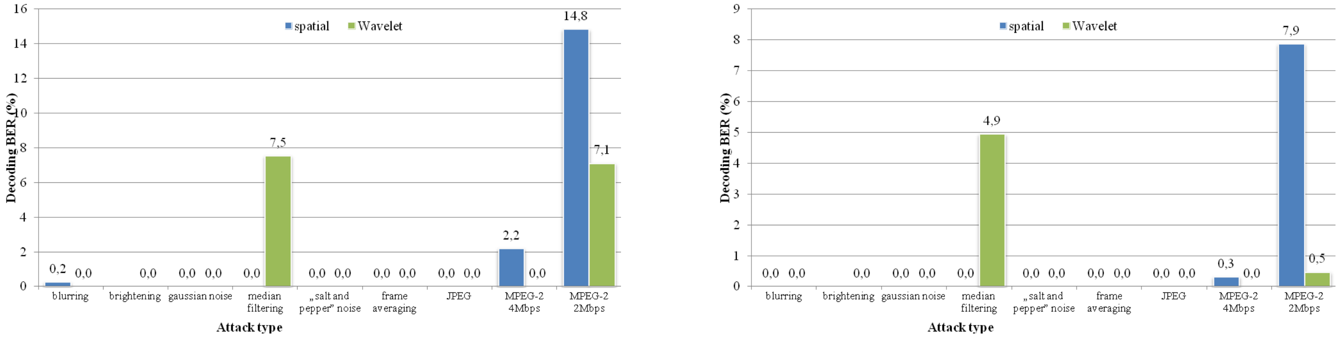


Figure 8. Decoding BER (%) for the proposed methods using $q = 8$, no error correction code and (a) k=3 and (b) k=9 redundant frames

Performance of the Rotated Constellation in DVB-T2

Ladislav Polak, Tomas Kratochvil
Department of Radio Electronics
Brno University of Technology
Brno, Czech Republic
xpolak18@stud.feec.vutbr.cz, kratot@feec.vutbr.cz

***Abstract*—This short paper deals with the performance of the rotated constellation, which is one of the main innovations in the DVB-T2 (Second Generation Digital Terrestrial Television Broadcasting) standard. Rotated constellation is an optional feature to improve performance in frequency selective channels. This paper contains the present and progress results of the rotated constellation performance. For the determination of differences between the non-rotated and rotated constellation, a 0 dB Echo channel model was used. Graphical dependences of the BER before and after LDPC coding are given. Finally, achieved results are evaluated and discussed with promising expectations of a very good performance of rotated constellation technique in DVB-T2.**

Keywords-DVB-T2; LDPC coding and decoding; rotated constellation; 0 dB Echo channel; BER

I. Introduction

Nowadays, DVB-T2 (2nd Generation Digital Terrestrial Television Broadcasting) standard is definitely the world's most advanced DTT (Digital Terrestrial Transmission) system, which offers robustness and high efficiency for terrestrial broadcasting [1]. It is built on the technologies used as part of the first-generation system (DVB-T), developed over a decade ago. DVB-T2 extends the possibilities of all parameters of DVB-T and significantly reduces overhead to build a system with a throughput close to theoretical channel capacity [2], [3].

DVB-T2 specification includes many innovations in system parameters. The combination of LDPC (Low Density Parity Check) and BCH (Bose-Chaudhuri-Hocquengham) codes give a very robust channel coding. Moreover, several options are available in areas such as the number of carriers, GI (Guard Interval) sizes and pilot signals. Therefore, the overheads can be optimized for any target transmission channel [1]- [3]. The advanced coding and interleaving techniques [2] offer good performance in so called non-selective channels. However, frequency selective channels (with deep fadings) need extra redundancy, previously given by a lower-rate code.

DVB-T2 also includes a novel technique of constellation, so called rotated constellation, which is one of the main innovations in DVB-T2 system configuration. It is an optional feature to improve performance even for very frequency selective channels [3]. The technique of the rotated constellation and the idea of its use in communication systems are not new. This method has been studied since 1997, when Giraud et al. presented the lattice constellations for the Rayleigh fading channel [4].

After then, in [5] an alternative diversity technique for Rayleigh fading channel was presented, which was achieved by the high diversity modulation schemes (rotation of constellation points). Now, the exploring of the possibilities of rotated constellation technique was used for the improving of the DVB-T2 system performance. In [6] and [7] the rotated constellation technique has been analyzed in terms of BER (Bit Error Ratio) evaluation. For the exploring of its performance a Rayliegh (P1 channel) and RME (Rayleigh Memoryless Erasures) [2] fading channel models were used. On the other hand, the appropriate design of the rotated demapper on the receiver side is very important [8]. When rotated constellation is not used, the LLRs (Log Likelihood-Ratios) soft decision metrics can be derived in the normal, one-dimensional way [2], well known from DVB-T. In case of rotated constellation the 2D LLR demapping is used. In [9] a very promising solution of this problem was presented, where a novel detection method (QAM detector) also reduces the demand of the computational resources.

In this work in progress paper, the investigation is focused on the performance of the rotated constellation in DVB-T2 standard from the perspective of transmission distortion in frequency selective channels. To demonstrate simply transmission distortions, a two path, 0 dB Echo channel was used [2]. Moreover, this article proposes and evaluates a configuration that has been optimized for the mentioned channel conditions.

The rest of the paper is organized as follows. After the introduction and related state-of-the-art works review in the area of rotated constellation, a brief description of the main differences between the non-rotated and rotated constellation techniques is presented in Section II. The parameters and short description of used channel model for the analysis and simulation are presented in Section III. Section IV contains the graphical dependences of the BER before and after LDPC decoding on C/N (Carrier-to-Noise Ratio) ratio for both, non-rotated and rotated constellation modes. Finally, the results are evaluated and discussed in Section V.

II. Non-Rotated and Rotated Constellations

A. *Non-Rotated Constellation*

A constellation diagram is a representation of a signal modulated by a digital modulation scheme, such as QAM. In DVB-T2 standard, it is a selection from QPSK, 16QAM, 64QAM or 256QAM modulation. In a classical, non-rotated

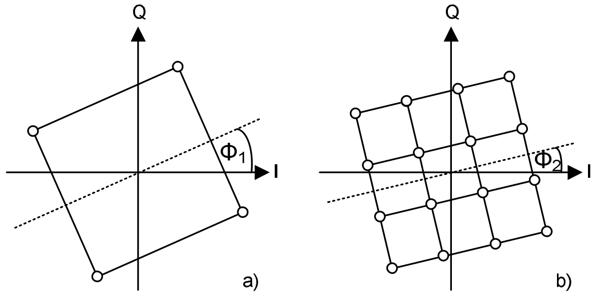


Figure 1. A theoretical rotated constellation diagram for a) QPSK with Φ_1 and b) 16QAM with Φ_2.

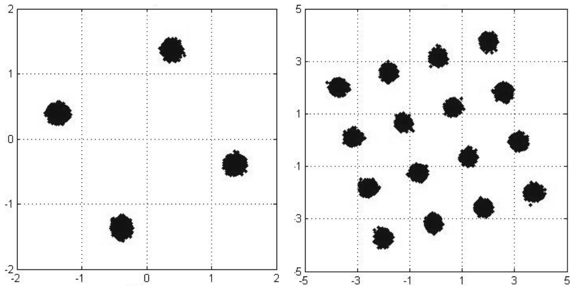


Figure 2. A simulated rotated constellation diagram for a) QPSK with Φ_1=29.0 and b) 16QAM with Φ_2=16.8 in AWGN channel with C/N = 20 dB.

constellation, the receiver needs both I (In-Phase) and Q (Quadrature) components of one constellation point to identify, which information was transmitted. The reason is that the estimation of I component does not give information about a Q component [6].

B. Rotated Constellation

As it was mentioned before, DVB-T2 standard introduces a novel technique to improve performance in selective fading channels. In case of the rotated constellation (see Fig. 1), a certain rotation angle is applied in the complex plane to a classical signal constellation. Then each component (I or Q), has enough information by its own to guess, which was the transmitted symbol [6]. Of course, the performances gain, in case of this technique, depends on the rotation angle. These angle values are different for each type of modulations in DVB-T2 and their exact value (see Fig. 2) can be found in [2].

Of course, only the rotation of constellation points is not enough for achieving a good performance. The additional innovation, and also the trick, is that the rotated constellation comes with Q-delay (after the constellation mapping). Delay means in this context that the Q components are shifted to the next COFDM (Coded Orthogonal Frequency Division Multiplexing) cell. This cyclical delay is realized on the level of individual FEC (Forward Error Correction) blocks [7], [10]. Thank to the combination of rotated constellation and Q-delays, I and Q components are now separated by the interleaving process (in cell, time and frequency) so that in general they are transmitted on different frequencies, different carriers and at different time. Therefore, if the channel destroys one of the components (I or Q) the other component (Q or I) can be used to recover the information [2].

III. Simulation Parameters

A brief description of two constellation techniques for DVB-T2 broadcasting standard was presented in the previous section. Due to the innovations in constellation mapping, DVB-T2 standard enables to improve the performance of data reconstruction on the receiver side, when the transmission conditions contains a lot of fadings. For the comparison of differences between the mentioned constellation techniques, we used a special type of fading channel, 0 dB Echo, well known from DVB-T.

The 0 dB Echo channel profile has been defined by Motivate partners [2]. Its composition has been largely influenced by the nature of the DVB-T/T2 signal. Concretely, it is defined by the following parameters:

- Spread spectrum technique - introducing ICI (Inter Carrier Interference) sensitivity to Doppler spread,
- Guard Interval - introducing IS (Inter Symbol) sensitivity to the echo delays.

This profile is made of two paths, having the same power (0 dB). These echoes are delayed by half of the GI value and they are presenting a pure Doppler characteristic [2]. The general graphical representation of the impulse and frequency response of 0 dB echo channel is shown in Fig. 3 and Fig. 4. The *Tg* is representing the value of the GI.

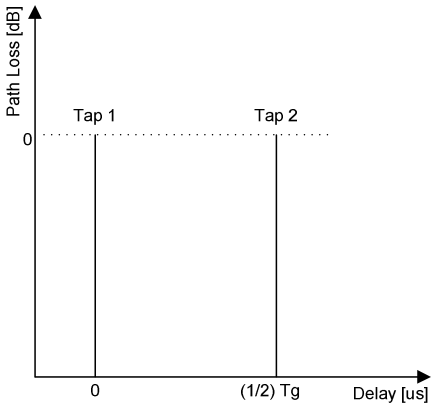


Figure 3. Impulse response of a 0 dB Echo channel.

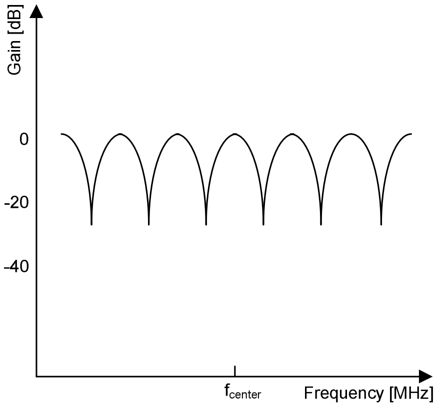


Figure 4. Frequency response of a 0 dB Echo channel.

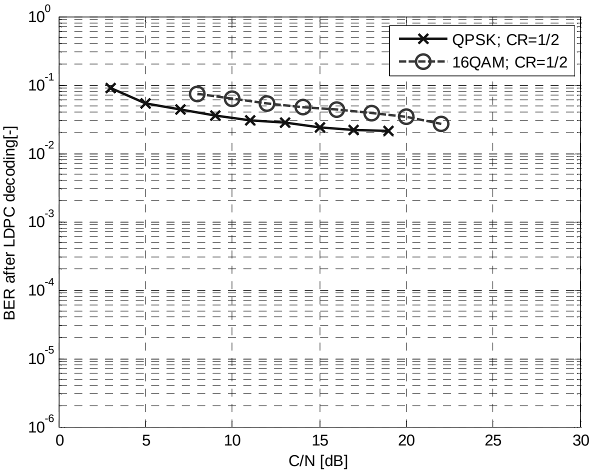


Figure 5. BER before LDPC decoding as a function of C/N ratio in the “0 dB Echo” channel (QPSK and 16QAM – non-rotated constellation, 2k mode, CR 1/2 and GI 1/16).

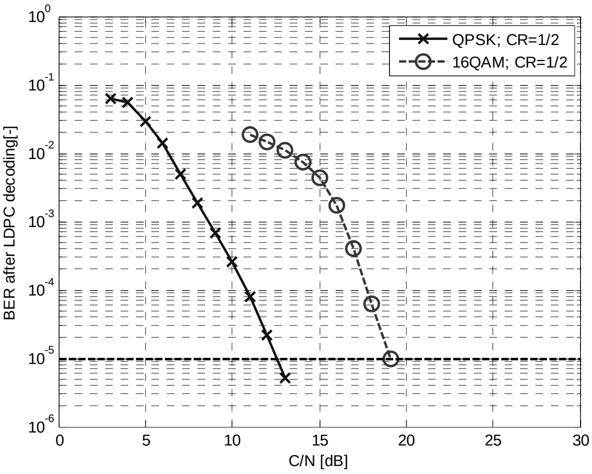


Figure 6. BER after LDPC decoding as a function of C/N ratio in the “0 dB Echo” channel (QPSK and 16QAM – non-rotated constellation, 2k mode, CR 1/2 and GI 1/16).

The implementation of functional model for the simulation of DVB-T2 transmission in MATLAB was done as it is recommended in [2]. For the simulation of the DVB-T2 transmission the following settings were used:

- mode: 2k (mobile reception),
- LDPC code ratio: 1/2 (robust protection),
- modulation: QPSK and 16QAM,
- constellation: non-rotated and rotated,
- rotation angle [Φ]: 29.0 (QPSK) and 16.8 (16QAM),
- Guard Interval: 1/16 (mid SFN - Single Frequency Network),
- decoding method: LDPC + BCH (with 50 iteration, as recommended in [2]).

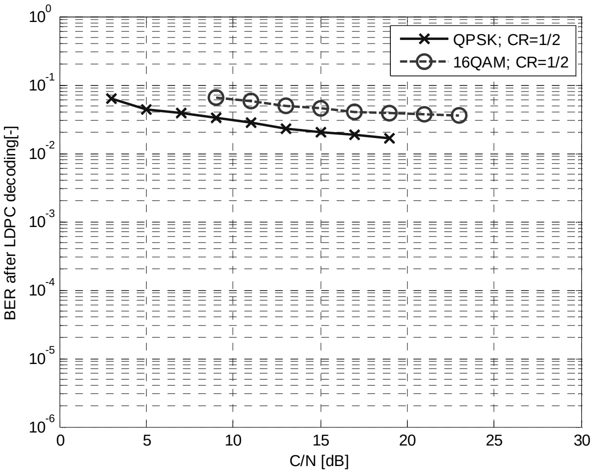


Figure 7. BER before LDPC decoding as a function of C/N ratio in the “0 dB Echo” channel (QPSK and 16QAM – rotated constellation, 2k mode, CR 1/2 and GI 1/16).

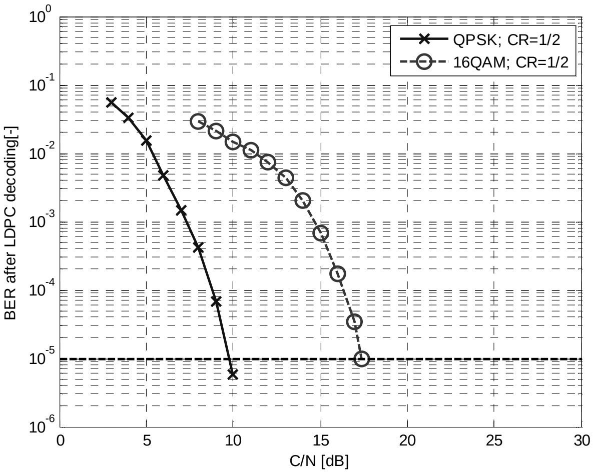


Figure 8. BER after LDPC decoding as a function of C/N ratio in the “0 dB Echo” channel (QPSK and 16QAM – rotated constellation, 2k mode, CR 1/2 and GI 1/16).

IV. SIMUALTION RESULTS

Simulation results of the DVB-T2 transmission for various C/N ratios in the 0 dB Echo fading channel were obtained. The simulation was done for two types of constellation technique: rotated and non-rotated.

DVB-T2 uses concatenated LDPC + BCH coding, the same as in DVB-S2 (2nd Generation Satellite DVB). These codes assure better protection, allowing more data to be transported in a given channel. It means that, for achieving a good signal quality (low BER); a lower C/N ratio is needed. In this paper, the limit of the error-free reception is considered as C/N for which BER is equal to 1.10^{-5} after LDPC decoding, as it is used in [6]. The number of iterations in LDPC decoding is depending on the hardware complexity of the receiver. In this paper, the number of iterations is equal to fifty (50) as recommended in [2].

TABLE I. COMPARISON OF THE SIMULATED RESULTS *C/N* FOR THE *BER* EQUAL TO10^{-5} IN 0 DB ECHO CHANNEL

Modulation	Configuration	Non-Rotated Constellation C/N [dB]	Rotated Constellation C/N [dB]
QPSK	2k mode CR_{LDPC} = 1/2 GI = 1/16	13.2	9.8
16QAM		19.3	16.5

As mentioned above, for the comparison of the performance of the non-rotated and rotated constellation 0 dB Echo channel was used. This type of channel is the worst case channel, which consists of two paths, with equal level and the second arriving later than the first as shown in Fig. 3. In this paper, we used configuration for the mobile scenario (2k mode) with mid size (GI = 1/16) of the SFN network. Therefore, the delay of the second path is equal to 7 us. Moreover, for the increase of the fadings in the channel, the speed of the receiver was set to 50 km/h.

Fig. 5 and Fig. 7 illustrate the BER before the LDPC decoding for QPSK and 16QAM modulations, when non-rotated and rotated constellation techniques were used. In these figures the effect of 0 dB echo fading channel can be observed. The BER decreases with the increased C/N ratio only slightly. The BER before the LDPC decoding (not corrected data) in both cases of non-rotated and rotated constellation are very similar. As can be seen, at this point, the performance of the non-rotated and rotated constellation is almost the same and any significant differences can not be found.

On the other hand, visible differences in achieved BER can be seen after the LDPC decoding, which are shown in Fig. 6 for non-rotated constellation, for rotated constellation in Fig. 8. In case, when we are using non-rotated constellation, for achieving a 1.10^{-5} BER it is needed high value of C/N ratio: 13.2 dB for QPSK and 19.3 dB for 16QAM modulation (see Tab. I). In case of rotated-constellation these values are 9.8 dB for QPSK and 16.5 dB for 16QAM modulation.

It should be noted that the maximum gain is obtained, when QPSK modulation with rotated constellation was used. This can be easily explained, since this modulation is the most robust to fadings. On the other hand, in comparison with classical constellation technique, the gain was better by 3 dB for both types of modulations, when the rotated-constellation technique was used.

V. CONCLUSIONS AND FUTURE WORKS

In this paper, the performance of the rotated constellation for DVB-T2, with comparison of non-rotated constellation, was explored. It has been shown that for fading channels with very bad conditions, a good performance can be obtained with rotated constellation. On the other hand, only the features of rotated constellation are not allowed for achieving a good signal quality. Unconditionally, the mentioned innovation of FEC coding/decoding, which is used in DVB-T2, has a significant role. Thank to the number of decoding processes (50 iterations in this simulation), which is used in this paper, the results in the special 0 dB Echo fading channel are much better. This advantage of DVB-T2 standard also improves the BER ratio in the fading channel [11]. The additional robustness can be used to increase the data rate by choosing a higher code rate while keeping the same minimum field strength.

This work will continue in the future by improving the rotated constellation technique for the analysis and simulation the transmission distortions in all possible scenarios, which can occur in DVB-T2 system configurations [12]. The work will be also focused on the investigation of the performance of rotated constellation in, so called, erasures channels. In case of DVB-T2 this is the RME channel [2]. Finally, the MER (Modulation Error Ratio) for both constellation techniques should be investigated deeper.

ACKNOWLEDGMENT

This paper was supported by the grant projects of the Czech Science Foundation no. 102/10/1320 “Research and modeling of advanced methods of image quality evaluation (DEIMOS)”, MEYS no. CZ.1.07/2.3.00/20.0007 “Wireless Communication Team (WICOMT)”, financed from the operational program Education for competitiveness. The described research was also performed in laboratories supported by the SIX project; no.CZ.1.05/2.1.00/03.0072, the operational program Research and Development for Innovation.

REFERENCES

[1] DVB Fact Sheet. DVB-T2 – *2nd Generation Terrestrial* (2011-08).

[2] TR 102 831 V0.10.4 (2010-06). *Digital Video Broadcasting (DVB); Implementation guidelines for a second generation digital terrestrial television broadcasting system (DVB-T2)*, Technical Report ETSI, 2010.

[3] L. Vangelista, et al., “Key Technologies for Next-Generation Terrestrial Digital Television Standrad DVB-T2,” IEEE communication Magazine, 2009, vol. 47, no. 10, pp. 146-153.

[4] X. Giraud, E. Boutillon and J. C. Belfiore, “Algebraic tools to Build Modulation Schemes for Fading Channels,” IEEE Transactions on Information Theory, vol. 43, no. 3, 1997, pp. 938–952.

[5] J. Boutros and E. Viterbo, “Signal Space Diversity: A Power- and Bandwidth-Efficient Diversity Technique for the Rayleigh Fading Channels,” IEEE Transactions on Information Theory, vol. 44, no. 4, 1998, pp. 1453–1467.

[6] D. P. Calderón, C. Oria, J. García, P. López, V. Baena and I. Lacadena, “Rotated constellations for DVB-T2, ” DCIS´09, Spain, 2009, pp. 41-45.

[7] CH. A. Nour, C. Douillard, “Rotated QAM Constellations to Improve BICM Performance for DVB-T2,” ISSSTA´08, Italy, 2008, pp. 354-359.

[8] D. P. Calderón, V. B. Lecuyer, A. C. Oria, P. López and J. G. Doblado, “Rotated constellation demapper for DVB-T2,” IEEE Electronic Letters, vol. 47, no. 1, 2011, pp. 31–32.

[9] M. Li, CH., A. Nour, CH. Jégo and C. Douillard, “Design of rotated QAM Mapper/Demapper for the DVB-T2 Standard, ” SiPS´09, Finland, 2009, pp. 18-23,.

[10] T. Vieracker, “Planning DVB-T2 – Advance and challenge (White Paper)”, LStelcom, June 2010, pp. 1-9.

[11] L. Polák and T. Kratochvíl, “DVB-SH-A and DVB-T2 Performance in Mobile TV Environment,” ISWCS´11, Germany, 2011, pp. 1-5.

[12] L. Polák and T. Kratochvíl, “Simulation and Measurement of the Transmission Distortions of the Digital Television DVB-T/H Part 3: Transmission in Fading Channels,” Radioengineering, 2010, vol. 19, no. 4, pp. 703-711.

Estimation of Perceived Quality in Convergent Services

Pedro de la Cruz Ramos[1], Mario Cao Cueto[1], Raquel Pérez Leal[2], Francisco González Vidal[1]

[1]Depto. de Ingeniería Telemática
Universidad Politécnica de Madrid
Madrid, Spain
{pcruzr, mcao, vidal}@dit.upm.es

[2]Depto. de Teoría de la Señal y Comunicaciones
Universidad Carlos III de Madrid
Madrid, Spain
rpleal@tsc.uc3m.es

Abstract— **Triple-Play (3P) and Quadruple-Play (4P) services are being widely offered by telecommunication services providers. Such services must be able to offer equal or higher quality levels than those obtained with traditional systems, especially for the most demanding services such as broadcast IPTV. This paper presents a matrix-based model, defined in terms of service components, user perceptions, agent capabilities, performance indicators and evaluation functions, which allows to estimate the overall quality of a set of convergent services, as perceived by the users, from a set of performance and/or Quality of Service (QoS) parameters of the convergent IP transport network.**

Keywords- Quality of Experience, Perceived Quality, Quality of Service, Network Performance, Quality Models.

I. INTRODUCTION

Customers of convergent Triple-Play (3P) and Quadruple-Play (4P) services expect a Quality of Experience (QoE) comparable to that obtained with traditional broadcast systems. Consequently, it is of utmost importance for 3P and 4P service providers to be able to measure, estimate and/or monitor user perceived quality in near real time, especially for the most demanding services such as broadcast IPTV.

User QoE in 3P/4P services depends on many factors, among other:

1) Perceived quality of each of the individual services, which in turn depends on:
   a) Perceived quality of each of the service components.
   b) Relationships, interactions and/or dependencies between the components.
2) Service availability and reliability.
3) System responsiveness, user-friendliness, etc.
4) Customer service.

For instance, 3P QoE depends on the perceived quality of the IPTV service, which in turn depends on audiovisual quality, which depends on audio quality, video quality and audio-video synchronization (lip sync), and so on.

This paper focuses on those elements of perceived quality that can be estimated, directly or indirectly, from performance or Quality of Service (QoS) parameters of the convergent IP transport network, i.e., parameters which can be measured at easily accessible reference points [1] or obtained from the Network Management System (NMS). These parameters include:

1) IP Packet Error Ratio (PER) and Packet Loss Ratio (PLR).
2) End-to-end IP packet delay.
3) Delay variation (jitter).

For instance, MPEG Video Quality can be estimated from QoS parameters such as PLR [2][3].

In order to estimate user perceived quality we propose the use of a matrix-based model, which allows to estimate the overall quality of a set of convergent services, as perceived by one or more types of users, from a set of performance and/or Quality of Service (QoS) parameters of the convergent IP transport network.

In the following sections the model is presented; its application to convergent services is described; the quality evaluation process is detailed; the main conclusions are summarized; and some future work is outlined.

II. PRESENTATION OF THE MODEL

The model is schematically depicted in Figure 1. Its elements are succinctly described in the following sections. It is thoroughly described in [4][5][6], where its application to a 3P (data + voice + video) service offering is also explained. The video service (VoD), however, is considered of little importance and thus dropped. In the case of domestic users, the voice service (VoIP) is also dropped, so that the global service reduces to a data service (Internet access + On-Line Gaming).

In this paper, instead, a full Triple-Play service offering, including data, voice and video services, is covered.

For the purposes of this paper, a "user" is anyone who "consumes" some service included in a 3P/4P service offering. Typically, they are unaware of the internal mechanisms used to provide the service, and of its composing elements. They are only interested in the "experience" delivered to them by these services, and judge the quality of this experience (QoE) by means of their subjective perceptions, and not by technical criteria.

Usually, we will not be interested in the individual QoE of specific users, but in the "average QoE" of a community or category of users with similar characteristics, i.e., of a "user type". Sometimes, we will even consider the average QoE of a wide sample of customers whith quite dissimilar characteristics, i.e., that of the "average user".

ISBN: 978-1-61208-193-9

A. Services and User Perceptions

In Figure 1, the upper side of the matrix corresponds to the services and how the users experience those services as sets of user perceptions.

Services. For the purposes of this paper, a service is defined as a set of functionalities whose purpose is to satisfy certain needs and which are perceived as a whole by the users. We will distinguish between:

1) Overall or Global Services: they are offered by providers as a whole, but composed of more elementary services. An example is the 3P Service offered by a provider as a whole.
2) Final Services: components of Global Services. They are not offered independently by providers, but are perceived as independent services by users. Examples: Internet Access, IPTV, VoIP, etc.
3) Elementary or Basic Services: components of Final Services. They are not offered independently by providers. Users perceive them as separate, but not independent, services. Examples: Web Browsing, Electronic Mail, File Transfer, etc.
4) Support Services: they support the Final and/or Basic Services. They are not offered independently, and users are often unaware of their existence. Examples: ADSL access, DNS, DHCP, etc.

In this paper, “Service” will mean “Final Service” unless otherwise stated.

User Perceptions. A user perception is a factor that influences the evaluation of the service quality as perceived by users, i.e., the Quality of Experience (QoE).

It may be quantified by means of a valuation, similar to that obtained by subjective methods such as MOS, DSCQM, SSCQE, etc. [7][8]. In this paper, user perceptions will be quantified using the Standard MOS Scale (from 1=unacceptable to 5=excellent) [9].

For each user perception, **Global Valuation Factors (GVF)** should be defined as objective, quantifiable parameters which determine (or at least influence) the subjective perception of quality. They are the result of the performance achieved by providers and obtained by the users of the service. They provide a clean separation between technical performance parameters and user quality evaluations. For example a GVF for the perception of “Download Speed” may be the “Page Download Time”.

B. Agents and Agent Capabilities

In the left side of Figure 1, the agents and their capabilities are depicted.

Agents. An agent is any component of a system which has individual, separate existence and provides an identifiable set of functionalities with the purpose of providing some service to the users. Examples: Content providers, Carriers, Access providers.

Agent Capabilities. These are the different functionalities provided by the agents, contributing to the provision of a service to the users. Examples: Connectivity, processing, data storage, data transfer.

For each agent there are **(Internal) Performance Parameters**, which are internal elements or factors that an agent may control or manage and that contribute to the performance of a capability. Typically, they are magnitudes related to the internal infrastructure or operation of the agent. Some typical examples are: throughput, bit error rate, MTBF.

C. Matching Points and Performance Indicators

As previously mentioned, the matrix-oriented quality estimation model tries to identify the dependencies between services and the performance and/or quality parameters related to the agents and their capabilities.

Matching Points. They represent the relationships or dependencies between user perceptions and agent capabilities, such that the capability affects or influences the perception. For instance, the data transfer capability of the transport network influences the loading speed perception in the web browsing service.

Performance Indicators. These are measurable magnitudes, associated to matching points, whose values determine or affect the user valuation of the corresponding perception. We will distinguish between:

1) Elementary Performance Indicators, which model the contribution of a single capability of an agent to a perception.
2) Local Performace Indicators, which model the contribution of all capabilities of an agent to a perception.
3) Global Performance Indicators, which model the contribution of all agents to a perception.

Some examples are: bandwith, delay, jitter.

D. Quality Evaluation Process

The Quality Evaluation Process comprises a set of sub-processes and functions. Figure 2 shows the information flow of these evaluation functions and processes, where the output of each step is the input to the next one. We will distinguish the following evaluation functions:

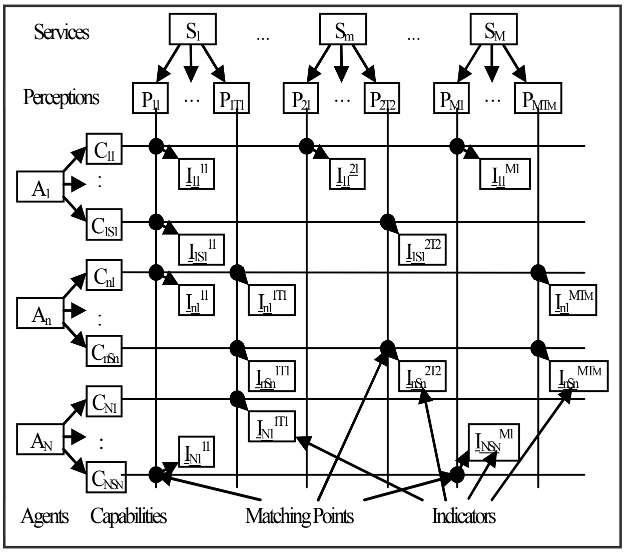


Figure 1. Matrix-oriented quality estimation model.

1) Performance Functions: compute the Elementary Performance Indicators from the Internal Performance Parameters.
2) Local Weighting Processes: compute the Local Performance Indicators from the Elementary Performance Indicators.
3) Global Aggregation Process: computes the Global Performance Indicators from the Local Performance Indicators.
4) Parameterization Functions: compute the Global Valuation Factors from the Global Performance Indicators.
5) Valuation Functions: compute the valuations of perceptions from the Global Valuation Factors.
6) Global Evaluation Process: computes the evaluations of the perceived quality of each Final Service and the overall evaluation of the Global Service, for each User Type and/or for the Average User. For this purpose, the Analytic Hierarchy Process (AHP) method [10] is used.

III. Application of the Model to Convergent (3P) Services

We will follow the application methodology described in [4]. We will try to estimate the average QoE of so-called “Residential Users”, i.e., domestic (home), non-enterprise users, whose interests are mainly the leisure and pastime opportunities given by broadcast IPTV, and the information access possibilities offered by Internet Access, and specifically, Web Browsing. These users also seek cost saving opportunities offered by VoIP.

For broadcast IPTV, these customers expect a QoE comparable to that of traditional broadcast systems (i.e., terrestrial or satellite TV). For VoIP, they will also expect a quality similar to that of POTS, but will very likely accept a quality similar to that of mobile telephony, if the cost savings are substantial.

Non-residential users, i.e., enterprise and SOHO (Small Office/Home Office), are usually not interested in TV services, and so in 3P services, and are thus not considered in this model.

A. Identification of Components

Following the model presentation described above, this section aims to identify the model components and define its corresponding parameters.

For the purposes of this paper, we will consider a 3P Global Service offering composed of the following **Final Services**:

1) Internet Access, including Web Browsing, Electronic Mail, File Transfer and File Sharing (P2P).
2) IP Telephony: Voice Call.
3) IPTV: Digital Video Broadcast (DVB).

As we are specifically interested in the estimation of user perceived quality from performance and/or QoS parameters of the underlying convergent IP transport network, we will deliberately ignore the Customer Service, Pricing and Marketing aspects of these services, as they cannot be estimated from network parameters, and will concentrate only in the technical quality aspects.

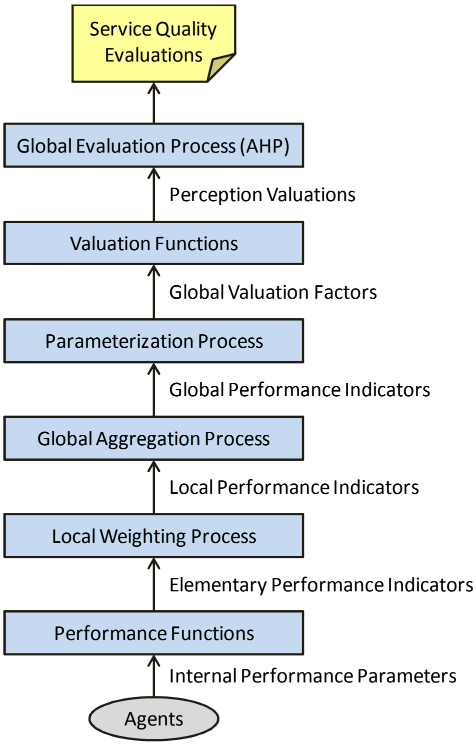

Service Quality Evaluations
Global Evaluation Process (AHP)
Perception Valuations
Valuation Functions
Global Valuation Factors
Parameterization Process
Global Performance Indicators
Global Aggregation Process
Local Performance Indicators
Local Weighting Process
Elementary Performance Indicators
Performance Functions
Internal Performance Parameters
Agents

Figure 2. Matrix-oriented quality estimation model.

The **User Perceptions** that we consider relevant for each service are shown in Table I. Some perceptions are common to all services. The **Global Valuation Factors** for each user perception are also shown in Table I.

The **Global Performance Indicators** (GPI) for each user perception of a representative service (Digital Video

Table I. User Perceptions and Global Valuation Factors

Service	**Perception**	**GVF**
Web Browsing	Download Speed	Page Download Time
Electronic Mail	Response Speed	Response Time
File Transfer	Download Speed	Download Rate
	Upload Speed	Upload Rate
File Sharing	Download Speed	Download Rate
Voice Call	Voice Quality	R-Factor Codec Parameters
	Response Speed	Response Delay
	Call Setup Speed	Call Setup Delay
Digital Video Broadcast	Video Quality	Video Quality Metric
	Audio Quality	PEAQ Metric
	Lip Sync	Audio-Video Delay
	Channel Change Speed	Channel Change Time
All Services	Availability	Successful Connection Percentage
	Reliability	Interrupted Connection Percentage

ISBN: 978-1-61208-193-9

Broadcast) are shown in Table II. The detailed relationships between indicators and GVFs (parameterization functions), and the process for deriving Global Performance Indicators (GPI) from Local and Elementary Performance Indicators (LPI/EPI) (i.e., the Local/Global Weighting Processes) will be given in Section IV.

For the identification of **Agents** we will use the network model and reference points recommended in [1]. The main agents are:

1) Content Provider(s), which are the ultimate responsibles of the delivery of the final services.
2) Service Provider(s), that we further subdivide into:
    a) Internet Service Provider (ISP), which provides Internet Access Services to end users. It may also integrate some additional services.
    b) Network Services Provider (NSP), which provides some of the Support Services (such as DNS, DHCP, etc.).
    c) Service Centers, which host the Final Services and provide connectivity between the ISPs and the Content Providers.
3) Network Provider(s), that we further subdivide into:
    a) Access Network Provider, which transports the information between the end-user and the ISP.
    b) Core (Transport) Network Provider (Carrier), which includes all those elements which connect the ISP that hosts the final service to other ISPs, e.g., neutral points, international accesses, inter-ISP accesses, etc.
4) End-User, including the User Plattform and Customer Premises Equipment (CPE).

The most relevant capabilities for each agent are shown in Table III.

TABLE II. GLOBAL PERFORMANCE INDICATORS (GPI) FOR A REPRESENTATIVE SERVICE (DIGITAL VIDEO BROADCAST)

Perception	Indicators
Video Quality	Packet Loss Ratio Video Coding Rate Image Size (Resolution) Image Rate Codec Parameters
Audio Quality	Packet Loss Ratio Audio Coding Rate Codec Parameters
Lip Sync	Audio-Video Delay
Channel Change Speed	IGMP Leave Time IGMP Join Time Key Acquisition Time Program Decoding Time Key-Frame Acquisition Time Frame Reordering Time Error Correction Time Processing Time Buffering Delay

B. Definition of Matching Points.

The relationships (Matching Points) between User Perceptions and Agent Capabilities depend on the precise information flows. We may distinguish four cases, depending on whether the content server is internal or external to the ISP, and whether or not the content is "cached" (stored) in the ISP or in the user platform.

For the purposes of this paper, we will consider the case where the content server is external to the ISP and there is no content caching outside the content provider. The resulting matching points between capabilities and perceptions are shown in Table III.

IV. QUALITY EVALUATION PROCESS

In this section, we will describe in detail:

1) The local and global weighting and/or aggregation processes (including weighting matrixes and/or metrics), valuation functions and quality models for each service.
2) The process for computing the Global Perceived Quality from the valuations of the perceptions for each service.

TABLE III. PERCEPTION-CAPABILITY MATCHING POINTS.

	Services	Web Browsing	Electronic Mail	File Transfer		File Sharing	Voice Call	Digital Video Broadcast				All Services	
	Perceptions	Download Speed	Response Speed	Download Speed	Upload Speed	Download Speed	Voice Quality	Video Quality	Audio Quality	Lip Sync	Channel Change Speed	Availability	Reliability
Agents	Capabilities	Matching Points											
User Platform	Processing	X	X	X	X	X	X	X	X	X	X	X	X
	Transfer	X	X	X	X	X	X	X	X	X	X	X	X
CPE	Processing	X	X	X	X	X	X	X	X	X	X	X	X
	Transfer	X	X	X	X	X	X	X	X	X	X	X	X
Access Network	Upstream C.		X		X		X				X	X	X
	Downstr. C.	X	X	X		X	X	X	X	X	X	X	X
Transport Network	Upstream C.		X		X		X				X	X	X
	Downstr. C.	X	X	X		X	X	X	X	X	X	X	X
ISP	Internal C.	X	X	X	X	X	X	X	X	X	X	X	X
	External C.	X	X	X	X	X	X	X	X	X	X	X	X
NSP	Connectivity	X	X	X	X	X					X	X	X
	Processing	X	X	X	X	X					X	X	X
Service Centers	Upstream C.		X		X		X				X	X	X
	Downstr. C.	X	X	X		X	X	X	X	X	X	X	X
	Processing	X	X	X	X	X	X	X	X	X	X	X	X
Content Provider	Upstream C.		X		X		X				X	X	X
	Downstr. C.	X	X	X		X	X	X	X	X	X	X	X
	Processing	X	X	X	X	X	X	X	X	X	X	X	X

ISBN: 978-1-61208-193-9

A. *Performance Functions*

The Performance Indicators corresponding to the capabilities of each agent will be measured directly (or obtained from the Network Management System). Thus Performance Functions are not required in this case.

In the most general case, where Performance Indicators cannot be measured directly, they should be derived from Internal Performance Parameters measured for each agent (or obtained from the NMS) by means of suitable Performance Functions.

B. *Local Weighting Process*

As we are considering the case of a single type of flow (see Section III.B), there is no need for a local weighting process: the contributions of the relevant capabilities of each agent are used directly in the global weighting process.

In the most general case, when several types of flows are considered, the contribution of each capability of each agent should be weighted depending on its participation in each flow and the importance or contribution of each flow to the total information flow.

C. *Global Aggregation Process*

For the Global Aggregation Process, simple metrics will be used as far as possible. They are summarized in Table IV.

D. *Valuation and Parameterization Functions*

As mentioned before, the valuation and parameterization functions relate the perceptions for each service to Global Valuation Factors and Global Performance Indicators.

In the next subsections, we provide models for the estimation of perceived quality for the main Basic Service of each Final Service.

1) **IPTV: Digital Video Broadcast**

We have developed our own model for estimating the video quality in IPTV. An early version of the model is described in [2], and a more advanced version in [3].

$$MOS = \begin{cases} 5 - 4 \cdot VQM & VQM \leq 1 \\ \\ 1 & VQM > 1 \end{cases} \tag{1}$$

$$VQM = VQM_C + VQM_L \tag{2}$$

$$VQM_C = VQM_{REF} \cdot (VCR/VCR_{REF})^{-K_C} \tag{3}$$

$$VQM_L = (1 - VQM_C) \cdot (PLR/PLR_1)^{K_L} \tag{4}$$

where

VQM is the Video Quality Metric as specified in [11]
VQM_C is the contribution of coding to VQM
VQM_L is the contribution of packet losses to VQM
VCR_{REF} is a reference VCR (e.g., 1Mbps)
VQM_{REF} is the value of VQM at the reference VCR
PLR_1 is the value of PLR for which VQM = 1

VQM_{REF}, K_C, PLR_1 and K_L depend on the codec, the coding parameters, and the characteristics of the video sequence (type, format, spatial and temporal complexity, information contents, etc.).

TABLE IV. METRICS FOR THE GLOBAL AGGREGATION PROCESS

Indicator	Metric
Delay	Additive
Delay Variance	Additive
Jitter	Rooted Sum of Squares (RSS)
Bandwidth	Concave
Packet Passthrough Ratio	Multiplicative
Packet Loss Ratio	Additive

PLR_1 and K_L also depend on VCR. We have found that their variation with VCR fits very well to a function of the form:

$$F(VCR) = A + B \cdot VCR \cdot (1 + C \cdot e^{-(VCR/D)^{\wedge}K}) \tag{5}$$

In order to estimate the Audiovisual Quality for synchronized audio and video streams, we use the model described in [12][13]:

$$Q_{AV} = K_0 + K_A \cdot Q_A + K_V \cdot Q_V + K_{AV} \cdot Q_A \cdot Q_V \tag{6}$$

where

Q_{AV} is the Audiovisual Quality Factor
Q_A is the Audio Quality Factor
Q_V is the Video Quality Factor

Q_{AV} must be converted to the standard MOS scale using the E-Model conversion function specified in [14]:

$$MOS = \begin{cases} 1 & Q_{AV} < 0 \\ 1 + 0{,}035 \cdot Q_{AV} + Q_{AV} \cdot (Q_{AV} - 60) \cdot (100 - Q_{AV}) \cdot 7 \cdot 10^{-6} & \\ & 0 \leq Q_{AV} \leq 100 \\ 4{,}5 & Q_{AV} > 100 \end{cases} \tag{7}$$

Q_V is derived from the MOS value given by (1) using the E-Model inverse function specified in ITU-T G.107 [14].

There are other factors that contribute to the global quality perception of the IPTV service, such as audio quality, audio-video synchronization (lip sync), channel change time, etc. In order to compute the global quality perception all these factors must be taken into account.

The Perceived Global Quality of the IPTV service will be computed using a nonlinear model:

$$Q_{IPTV} = K_{IPTV} + K_{AV} \cdot Q_{AV} + K_{Tav} \cdot Q_{Tav} + K_{Tcc} \cdot Q_{Tcc} + K_{AVTav} \cdot Q_{AV} \cdot Q_{Tav} + K_{AVTcc} \cdot Q_{AV} \cdot Q_{Tcc} + K_{TavTcc} \cdot Q_{Tav} \cdot Q_{Tcc} \tag{8}$$

where

Q_{IPTV} is the Global Quality of the IPTV service
Q_{AV} is the Audiovisual Quality given by (7)
Q_{Tav} is the Perceived Quality due to audio-video desynchronization (lip sync)
Q_{Tcc} is the Perceived Quality due to Channel Change Time (CCT)
$Q_i \cdot Q_j$ are the interaction terms

The coefficients K_i will be computed using the AHP method.

2) **IP Telephony: Voice Call**

In order to estimate the voice quality perception in IP Telephony, the adaptation of the E-Model [14] for VoIP given in [15] [16] will be used:

$$R = Ro - Is - Id - Ie + A \qquad (9)$$

$$Id = \begin{cases} 0.024 \cdot d & d < 177.3 \\ \\ 0.134 \cdot d - 19.503 & d \geq 177.3 \end{cases} \qquad (10)$$

$$Ie = a \cdot \ln(1 + b \cdot \rho) + c \qquad (11)$$

where

R is the R-Factor of the E-Model
$Ro = 93.2$ is the signal/noise ratio for 0dBr
Is is the degradation of the voice signal
Id is the degradation caused by delay and delay variation (jitter)
Ie is the degradation caused by the equipment (coding and packet loss)
A is the User Expectation Factor
d is the end-to-end delay in miliseconds
ρ is the end-to-end packet loss ratio
a,b,c are coefficients that depend on the codec

The R-Factor will be converted to the standard MOS scale using the E-Model conversion function [14]:

$$MOS = \begin{cases} 1 & R<0 \\ 1+0{,}035 \cdot R + R \cdot (R-60) \cdot (100-R) \cdot 7 \cdot 10^{-6} & 0 \leq R \leq 100 \\ 4{,}5 & R>100 \end{cases} \qquad (12)$$

There are other factors that contribute to the global quality perception of the service, such as dial tone delay, call setup time, etc. In order to compute the global quality perception all these factors must be taken into account.

The Perceived Global Quality of the IP Telephony service will be computed using a nonlinear model similar to that used for the IPTV service (8).

3) **Internet Access: Web Browsing**

In order to estimate the preceived quality for Web Browsing in the Internet Access service, we will use the model proposed in [4], which in turn is based on that proposed in [17]:

$$MOS = \begin{cases} 5 & T < 2 \text{ seg} \\ 5 - \log_2 T/2 & 2 \text{ seg} \leq T \leq 30 \text{ seg} \\ 1 & T > 30 \text{ seg} \end{cases} \qquad (13)$$

$$T = T_{DNS} + 2 \cdot RTD + T_{MAIN} + N \cdot S/B \qquad (14)$$

where

T is the average page download time
T_{DNS} is the time needed for name resolution
RTD is the Round Trip Delay
T_{MAIN} is the main page download time
N is the average number of objects in a page
S is the average object size
B is the effective bandwith

The Perceived Global Quality of the Internet Access service will be computed by combining the Perceived Quality evaluations for each Basic Service using a linear model as proposed in [4]:

$$Q_{IA} = K_{WB} \cdot Q_{WB} + K_{EM} \cdot Q_{EM} + K_{FT} \cdot Q_{FT} + K_{FS} \cdot Q_{FS} \qquad (15)$$

where Q_{IA} is the Perceived Quality of the Internet Access Service
Q_{WB} Web Browsing Service
Q_{EM} Electronic Mail Service
Q_{FT} File Transfer (FTP) Service
Q_{FS} File Sharing (P2P) Service

The coefficients K_i will be computed using the AHP method [10].

E. Global Evaluation Process

In this section, the contributions of the different elements of the model are weighted and combined in order to produce a global evaluation of the perceived quality of the 3P service. The AHP method [10] will be used when the weights cannot be determined in a more specific way.

1) **Evaluation of Perceptions**

Once the different perceptions related to a service have been derived (valuation and parameterization functions), they must be combined in order to obtain the global evaluation of the service. For each service, an AHP matrix [10] should be used to define the relative importance of the different perceptions.

For all services, we have considered that service availability has extreme importance, and service reliability strong importance for the users, relatively to other perceptions. For other perceptions, we have considered the primary perceptions (other than service availability and reliability) as moderately more important than the secondary perceptions. These ratings will be refined once we had more evidence of the relative importance of these perceptions for domestic users.

2) **Evaluation of Services**

In [4], all services are evaluated in a single step. Instead, we have decomposed the Service Evaluation process in two steps: first, the relative importance of the Final Services is rated; then, the relative importance of the Elementary Services of each Final Service is rated. This method scales better to a situation with many Final Services, each in turn composed of many Elementary Services.

The relative weights for the Final Services are shown in Table V. They are derived from service usage data [18].

TABLE V. IMPORTANCE WEIGHTS FOR FINAL SERVICES.

Service	Internet Access	IP Telephony	IPTV
Home percent	63,9	80,6	99,6
Weight	0,2618	0,3302	0,4080

As an example, the relative importance (AHP Matrix) for the Elementary Services of the Internet Access Service is shown in Table VI. The value (rating) in each cell represents the importance of the service in the row relative to the service in the column. The precise meaning of each rating is described in [10], but intuitively a higher rating means that the row is more important relative to the column.

The importance ratings are derived from those in [4] after removing the unused services and including the new ones. We have given File Sharing the same importance as File Transfer, and kept the relative importance of other services.

The Consistency Ratio (CR) of this matrix is 2.25%<10%, so the relative importance factors are acceptably consistent. The corresponding weights are shown in Table VII.

V. Conclusion

A model for the estimation of quality as perceived by the users (i.e., the user Quality of Experience, QoE) in Triple-Play (3P) and Quadruple-Play (4P) services has been presented. The model is based on a matrix framework defined in terms of user types, service components, and user perceptions on the user side, and agents, agent capabilities, and performance indicators on the network side. A Global Quality Evaluation process, based on several layers of evaluation functions, has been described, that allows to estimate the overall quality of a set of convergent services, as perceived by the users, from a set of performance and/or Quality of Service (QoS) parameters of the convergent IP transport network. The model has been refined for the particular case of residential (domestic) users with a specific information flow where the content server is external to the ISP and there is no content caching outside the content provider. The full sets of services, user perceptions, valuation factors, agents and agent capabilities have been provided, as well as the full matrix of matching points between agent capabilities and user perceptions. Performance indicators, as well as valuation and parameterization functions for some representative services (Digital Video Broadcast in IPTV, Voice Call in IP Telephony, and Web Browsing in Internet Access) have been provided. For Global Service Quality evaluation, weights for the Final Services, derived from service usage statistics, have been provided, as well as an example of the use of the AHP method for deriving the weights of the Elementary Services of a Final Service (Internet Access). In summary, the paper shows the applicability of the proposed model to the estimation of perceived quality (Quality of Experience) in convergent 3P/4P services.

Table VI. AHP Matrix for the Internet Access Service (Domestic Users)

	Web Browsing	E-Mail	File Transfer	File Sharing
Web Browsing	1	4	6	6
E-Mail	1/4	1	3	3
File Transfer	1/6	1/3	1	1
File Sharing	1/6	1/3	1	1

Table VII. AHP Weights for the Internet Access Service (Domestic Users)

Web Browsing	E-Mail	File Transfer	File Sharing
0.6121	0.2164	0.0858	0.0858

Acknowledgment

This research was partially supported by the Spanish Ministry of Science and Innovation grant TEC2008-06539 (ARCO Project).

References

[1] ITU-T G.1081, “Performance monitoring points for IPTV,” International Telecommunication Union, October 2008.

[2] P. de la Cruz Ramos, F. González Vidal, and R. Pérez Leal, “Perceived video quality estimation from spatial and temporal information contents and network performance parameters in IPTV,” Proc. of the Fifth IARIA International Conference on Digital Telecommunications (ICDT 2010), pp. 128-131, Athens, Greece, June 2010.

[3] P. de la Cruz, R. Pérez Leal, and F. González Vidal, “A model for perceived video quality estimation from coding and QoS parameters in IPTV,” December 2011. Submitted to IEEE Communications Magazine: Special Issue on QoE Management in Emerging Multimedia Services.

[4] F. Liberal Malaína, “Proposal of a model and a methodology for quality management in telecommunication services,” PhD. Thesis, University of the Basque Country, Spain, September 2005.

[5] F. Liberal Malaína, A. Ferro, and J. O. Fajardo, “PQoS based model for assessing significance of providers statistically,” Proceedings of HETNET’05 Conference, 2005.

[6] F. Liberal Malaína, H. Koumaras, L. Sun, A. Ferro, A. Kourtis, and E. C. Ifeachor, “QoE in multi-service multi-agent networks,” International Journal of Communication Networks and Distributed Systems, 2006.

[7] ITU-R BT.500-12, “Methodology for the subjective assessment of the quality of television pictures,” International Telecommunication Union, September 2009.

[8] ITU-T P.911, “Subjective audiovisual quality assessment methods for multimedia applications,” International Telecommunication Union, December 1998.

[9] ITU-T P.800, “Methods for subjective determination of transmisión quality,” International Telecommunication Union, 1997.

[10] T. Saaty, “The Analytic Hierarchy Process,” McGraw Hill, New York, USA, 1990.

[11] ITU-T J.144, “Objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference,” International Telecommunication Union, March 2004.

[12] M. N. Garcia and A. Raake, “Impairment-factor-based audio-visual quality model for IPTV,” International Workshop on Quality of Multimedia Experience (QoMEx) 2009, San Diego, California, USA, July 29-31, 2009.

[13] M. N. Garcia, R. Schleicher, and A. Raake, “Impairment-factor-based audiovisual quality model for IPTV: Influence of video resolution, degradation type, and content type,” EURASIP Journal on Image and Video Processing, Volume 2011, Article ID 629284, 2011.

[14] ITU-T G.107, “The E-Model, a computational model for use in transmission planning,” International Telecommunication Union, April 2009.

[15] L. Sun, “Speech quality prediction for voice over Internet Protocol networks,” PhD. Thesis, University of Plymouth, U.K., 2004.

[16] L. Sun and E. Ifeachor, “New models for perceived voice quality prediction and their applications in playout buffer optimization for VoIP networks,” Proceedings of IEEE International Conference on Communications (IEEE ICC 2004), Paris, France, pp. 1478-1483, 2004.

[17] R. D. van der Mei, “Performance analysis of communication networks,” Faculty of Science, Vrije Universiteit, 2004.

[18] INE, “Encuesta sobre equipamiento y uso de tecnologías de la información y comunicación en los hogares 2011,” Instituto Nacional de Estadística, Ministerio de Economía y Hacienda, Gobierno de España, October 2011.

Unified Methodology for Broadband Behavior Measurements in the Acreo National Testbed

Christina Lagerstedt, Andreas Aurelius, Hemamali Pathirana, Claus Popp Larsen
Netlab, Acreo AB
Kista, Sweden
christina.lagerstedt@acreo.se
andreas.aurelius@acreo.se
hemamali.pathirana@acreo.se
claus.popp.larsen@acreo.se

Olle Findahl
World Internet Institute
Gävle, Sweden
olle.findahl@wii.se

Abstract— **The continuous evolution of Internet and Internet applications increases the demands on access networks. Understanding user behavior and Internet usage patterns is fundamental in developing future access networks that meet technical as well as end user needs, and from a societal point of view it is equally important to correctly recognize and understand user behavior. In this paper, we present measurements from the Acreo National Testbed where we have access to traffic measurements from real end users. We have developed a unified methodology combining traffic measurements with web questionnaires and diaries to compare the results from different methods as well as investigate user behavior. By comparing the different measurement methods we find that the end users have difficulty in estimating the time they spend on different Internet activities although they are fairly well able to estimate the frequency of usage. We also found that though the diaries are quite accurate, the traffic measurements give us a much more detailed picture of the end user activity. The importance of having a testbed with real end users is invaluable to this kind of study and we emphasize the importance of having access to access network traffic.**

Keywords - Testbed; Traffic measurements; User behavior; FTTH.

I. INTRODUCTION

As the Internet continues to evolve and offer new services, it takes up a larger part of our lives. We find new ways to communicate, interact and entertain ourselves. This puts new demands on access networks [1] and requires new insights into the behavior of those who use them. We believe that understanding user behavior and needs is the key to develop future networks and services that are accessible, reliable and that address the needs of real end users. There are several ways to study user behavior. From the technical side, the data traffic can be measured and analyzed. Other common ways are to use surveys or diaries. Traffic measurements are routinely performed by all larger operators, but results are rarely published because the operators don't want to share this information with competitors. On the other hand, published behavior studies are almost always based on surveys with individuals (e.g., telephone interviews with a large population). Such surveys often attract considerable interest among in the public debate, and far-reaching conclusions may be drawn.

The question is now whether such surveys are reliable. People can forget about their Internet behavior, they may not know what they did, they may not know what the children in the household have done, they may lie about sensible subjects, etc.

The purpose of this study is to develop a unified methodology where different kinds of surveys are combined with actual traffic measurements. We compare three different methods of looking at Internet user behavior; 1) Internet protocol (IP) traffic measurements, 2) web questionnaire and 3) diary. This will lead to a more detailed knowledge of the behavior, and, by evaluating results from different methods, we will obtain a better knowledge of their respective limitations. Furthermore, this way it can be verified whether surveys are correct or whether there perhaps is a systematic bias in survey answers that thus leads to misleading results.

There are few comprehensive traffic measurement studies in the literature and most of these are based on traffic from a campus areas [2], [3] or measure aggregated traffic [4], [5]. Questionnaires and diaries are well known and often used methods [6], [7], [8], but, as far as we know, have not been used together with traffic measurements. In the following sections, we will first describe the measurement setup and population. We will then present and discuss our results and finally presents our conclusions.

Combining technical measurements with surveys require test subjects in order to get statistical data. At Acreo we put a lot of effort into developing the Acreo National Testbed, ANT, which enables us to perform in-depth measurements and test new technology and equipment as well as to interact with end users. Here we can study user behavior in a unique way as we have access to real user traffic measured on a household level.

We have signed agreements with the end users where they agree to give feedback and participate in surveys and investigations. This means that the network conditions such as network topology, link speeds, service setup, etc as well as user metadata such as the number of people in a household, age, etc are known. This gives us a unique opportunity to perform measurements of user behavior, and to compare the results from different measurement techniques in order to evaluate the validity of the results.

II. Test Environment and Population

As mentioned in the introduction, the measurements in this study were performed in the Acreo National Testbed (ANT), which has previously been described in [9] and [10]. Contrary to lab based testbeds, this is a live network with real end users or *test pilots*. In return to being test pilots the end user households are given free access to services like Internet and IPTV. Fiber to the home (FTTH) is the main access technology in the testbed, and a schematic picture of the network is shown in Fig. 1.

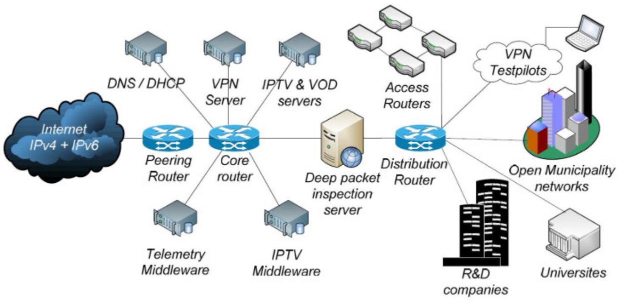


Figure 1. Schematic picture of the Acreo National Testbed, ANT.

The FTTH installation at the test households is active Ethernet providing 100 Mbit/s symmetrical connections to each household.

The number of households in the testbed changes over time according to the current tests that are being performed. At the time of this study, there were approximately 40 active households in the testbed. Approximately 20 were apartments in a building centrally located in the Swedish town Hudiksvall. The rest of the households, approximately 20 single dwelling units, were connected to the testbed via a fixed wireless access network, depicted as HSPA (High Speed Packet Access) in Figure 1.

The measurements in this paper are based on data from 5 testbed households. The reason for using a small population in the study is that we wanted to perform a *qualitative* analysis of their Internet usage, i.e., the main point is not to gather data for statistical analysis, but rather to study the user behavior *in depth*, as we have not found any similar studies in the literature. As this is a novel comparative methodology it was also important to develop methods and analysis tools that can be scaled up to considerably larger populations [13] where the amount of data to be handled will be much more extensive. That is, we do not claim to be able to make statistical conclusions in this study, but interesting trends for further study will be pointed out. Nevertheless, we will develop a unified methodology that, at a later stage, can be used for more substantial conclusions.

A letter of invitation was sent out to all households in the testbed, asking for participation in the study. The participants were selected from those who responded. In the selection process, household details such as number of people in the household, number of computers, ages, etc were taken into account, in order to get a varied test population with different household constellations. The test pilots were assured that their answers would be treated anonymously. This resulted in a test group of 5 households of which 2 were single households and 3 family households. Of the family households, one had small children (<7 years old), one had teenage children and one both teenagers and small children.

III. Measurements

The combined measurement methodology consists of three parts: traffic measurements, web questionnaires and diaries. In this section we will describe these methods.

A. *Traffic Measurements*

The traffic measurements have been performed using PacketLogic (PL) [12], a commercial traffic management device used in many commercial broadband access networks all over the world. Traffic is identified based on packet content (deep packet inspection and deep flow inspection) instead of port definitions. The device can identify more than 1000 Internet application protocols, and the signature database is continuously updated.

Since the PL is a commercial product, the details of its functions are proprietary. However, the identification process is connection-oriented, which means that each established connection between two hosts is matched to a certain application protocol. When a new connection is established, the identification of this connection begins. The identification algorithm searches for specific patterns, signatures, in the connection. The patterns are found in the IP header and application payload. The PL uses the traffic in both directions in the identification process. The measurement point is depicted as “deep packet inspection” in Fig. 1.

The PL can track and identify several hundred thousand simultaneous connections, storing statistics in large databases. The statistics database records the short-time average amount of traffic in inbound and outbound directions as well as the total traffic for all nodes in the network. The data is averaged over 5 minute periods. Data concerning which web sites have been visited is stored in the connection log.

The measurement setup, although giving detailed measurements, has certain constraints. First, the traffic is measured per household and not per person and the analysis in this paper is therefore done on a per household basis.

There is also a 5 minute resolution in the measurements, which may have an impact on measurements of applications that are used in short time periods such as instant messaging. The data cut-off is 1 kbps, which may influence the measurements of certain applications such as gaming where the amount of data is generally very low. The signature database in this study was not up-to-date due to old hardware, which may result in a larger amount of unknown traffic. An upgrade will be performed before follow-up studies are performed.

The Internet traffic of each household was measured both during the days when the household recorded their diaries and for a complete month (May 2009) to get enough statistics to compare with the web questionnaire. Statistics on what web sites were visited by the different households were monitored for two weeks.

B. Web questionnaires

Each household member was asked to answer a web questionnaire regarding their Internet activity and behavior. In the case of small children (<7 years), the parents were asked to answer for them. The web questionnaire contained basic questions concerning family situation, education and occupation as well as questions pertaining to computer knowledge and Internet activity. The Internet activity questions include questions on what kind of medium is used to access Internet applications and frequency of use of different applications.

The questionnaire also surveys how often different types of web sites are visited such as banks or newspapers, etc. Finally the respondents are asked to estimate the amount of time they spend on different Internet activities. The questions are multiple choice with an additional field for comments.

C. Diaries

The members of each household were asked to fill in a diary logging their Internet activity during two consecutive days, 17-18 May 2009. Each day was divided into 15 minute intervals. The diary had four columns that the test pilots were asked to fill in:

1. Daily activity (sleep, work/school, leisure time activities, meal times etc.)
2. Media usage (TV, newspaper , radio, book, etc)
3. Internet activity when at home (web browsing, playing games online, community, downloading material from the Internet, etc)
4. Web address or service used

IV. RESULTS

From the traffic measurements, we find that the Internet activity of the households in the study occurs mainly during afternoons and evenings with shorter bursts of traffic during the morning and lunch hours. This is consistent with the traffic patterns established both in ANT and in municipal networks of similar characteristics, but with much higher populations [10]. The daily traffic pattern in the testbed during May 2009 is shown in Figure 2.

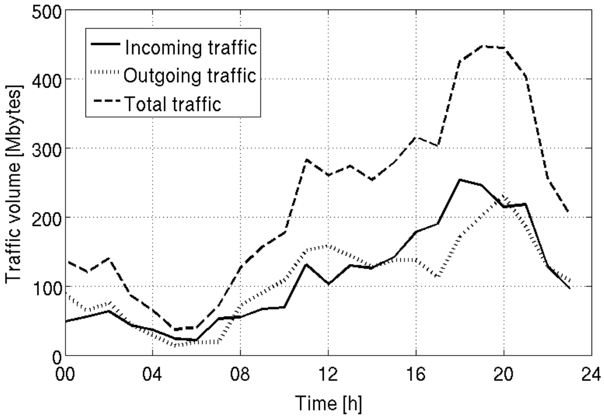


Figure 2. Daily traffic pattern of all active households in ANT during May 2009.

We also note that the average time spent online calculated from the traffic measurements is greater during the weekends than the weekdays for the family households while the opposite is true for the households without children, see Table I.TABLE I. The assumption here is that for the family households, the time of day when household members use Internet applications will be more spread out during the weekends and of course there are more people at home with leisure time. However, this assumption should be confirmed for a larger population.

TABLE I. AVERAGE TIME PER DAY SPENT ON INTERNET APPLICATIONS PER HOUSEHOLD.

Household	All days [min/day]	Weekends [min/day]
1	253	111
2	432(IP1) / 75(IP2)	614 (IP1) / 66(IP2)
3	588	496
4	196	154
5	1047	1162

Concerning application usage, the frequencies of use reported by the household members in the web questionnaires corresponds well with the measured data for the most part. The test pilots were able to give estimations of how often they used specific applications such as Spotify or file sharing applications and how often they visited certain types of web sites such as newspapers or banks. Deviations in the estimation of frequencies were however seen for more general questions about usage. For example, the users were able to estimate how often they use Spotify but had a harder time answering questions about how often they listened to streaming music or watched streaming media. Here it is important how the questions are posed and if the user understands which medium he/she is using.

TABLE II. USER PENETRATION OF SPECIFIC APPLICATIONS SEEN IN THE TRAFFIC MEASUREMENTS.

Application Household	**HTTP**	**HTTP media stream**	**BitTorrent**	**Spotify**	**SSL**	**Flash video**	**MSN messenger**	**Skype**
1	x	x	x		x	x	x	
2	x	x	x	x	x	x	x	x
3	x	x		x	x	x		
4	x	x			x	x		
5	x	x	x		x	x	x	

The user penetration of a number of applications seen in the traffic measurements is found in Table II. This is in good agreement with the answers from the web questionnaires. As is expected, HTTP is used by all households as well as the SSL protocol, which is used by for example Internet shops and banks. All of the households also use HTTP media stream as well as flash video meaning that they look at streaming material on the Internet.

Although the estimation of frequencies agreed fairly well with the traffic measurements, there were discrepancies between the approximations of the time spent on different applications and the measured time. For example, three of the households were able to give a reasonable figure as to how much time they spent using Internet applications though tending to slightly underestimate the actual figure, see Table III. The deviation between measurements and questionnaire was especially large for household number 2. We note that this is a family household and the total time is more difficult to estimate.

TABLE III. TIME SPENT ON INTERNET APPLICATIONS PER HOUSEHOLD AND PER WEEK.

Household	**Questionnaire [h/w]**	**Measured HTTP [h/w]**	**Measured ALL [h/w]**
1	14	16	30
2	97	28 (IP#1)	50
3	8	10	69
4	7	14	23
5	49	48	122

Comparing the diaries with the traffic measurements we find that they are mainly in accordance with each other. However, the traffic measurements add details to the picture given by the diaries, showing the potential that the method has in analyzing user behavior. A limitation of a diary is that users may not record all activities, either because they are done very frequently or because they are a natural part of your daily life that you do without reflecting that you are actually using the Internet. For example, one of the respondents had noted down "watching TV" in the diary. From the traffic measurements, we find that she was watching SVT play (streaming video from the Swedish state television) while at the same time being active on several community sites. This raises, among other things, an interesting question: What is actually meant by watching TV in the future when even more TV material will be available on-demand over the Internet?

The questionnaire contains questions concerning file sharing. The answers correspond well with what is measured and the frequencies of use recorded in the questionnaires agree well with the measurement. Although the use of file sharing applications is a sensitive question in public debate, the test pilots seem comfortable answering questions about this. Three of the households have used file sharing applications during the measurement period but none of the households are heavy users. It should be noted here that the test pilots are used to answer questionnaires, so they may not be representative in the sense that they may be less shy than other users when sharing sensible information.

V. CONCLUSION AND FUTURE WORK

In this paper, we have proposed and applied a unified methodology using three different methods to study Internet user behavior: traffic measurements, web questionnaires and diaries with the purpose of verifying and comparing the different methods as well as gaining more insight into user behavior.

From the measurements, we conclude that the test pilots are well able to describe some of their short term behavior seen in the diaries, although some activities were not noted in the diaries. The long term behavior seen in the web questionnaires are fairly accurate in describing frequencies of use specific applications and visits to specific web sites. The estimation of the amount of time spent on different activities was seen to differ from that of the traffic measurements, with a slight tendency to under-estimate the time spent. An even more powerful conclusion is the complex and rich picture of user behavior, which is obtained via traffic measurements. Here, details and behaviors that are not exposed in diaries or questionnaires are visible. This gives new insights into user behavior as well as valuable feedback for better construction of question based investigations in the future.

Another major result of the study is the importance of the testbed to the study. Here, we have the possibility of making measurements in a controlled environment with real end users. We gain much more insight into the behavior of the end users than can be obtained from only questionnaires or diaries. We also gain understanding of what the end user experiences that complements the traffic measurements. From the network side, this can be used to improve the quality of service both from the technical and the end user perspective.

Our future work will continue with a wider study to follow up on the results presented here as well as a continued development of our testbed, which makes these types of measurements possible.

ACKNOWLEDGMENTS

The work in this project has been partly financed by the VINNOVA project Broadband Behavior and the European Regional Development Fund project Acreo National Testbed. Also, it has been partly funded by the Celtic IPNQSIS project and Swedish National Strategic Research Area (SRA) within the program TNG (The Next Generation) and the project eWIN.

REFERENCES

[1] M. Chesire, A. Wolman, G. M. Voelker, and H. M. Levy, “Measurement and Analysis of a Streaming-Media Workload”, Proceedings of the 3rd conference on USENIX Symposium on Internet Technologies and Systems - Volume 3, 2001.

[2] M. Alvarez-Campana, A. Azcorra, J. Berrocal, D. Larrabeiti, J. I. Moreno, and J. R. Pérez, “CASTBA: Internet Traffic Measurements over the Spanish R&D ATM Network”, 5th HP Openview University Association Workshop, 1998.

[3] Y. Bhole and A. Popescu, “Measurement and Analysis of HTTP Traffic”, Journal of Network and Systems Management, pp. 357-371, Vol. 13, No. 4, December 2005.

[4] C. Fraleigh, S. Moon, B. Lyles, C. Cotton, M. Khan, D. Moll, R. Rockell, T. Seely, and C. Diot, “Packet-level traffic measurements from the sprint ip backbone”, IEEE Network, pp. 6-16, vol. 17, no. 6, 2003.

[5] J. Zhang, J. Yang, C. An, and J. Wang, "Traffic Measurement and Analysis of TUNET," 2005 International Conference on Cyberworlds (CW'05), Nov. 2005.

[6] J. George, “Researching Life in E-Society with Diary Studies”, Paper 67, Proceedings of the 2006 European Conference on Information Systems, Gothenburg, Sweden, June 2006.

[7] K. Vermaas and L. van de Wijngaert, "Measuring Internet Behavior: Total Time Diary and Activity Diary as Research Methods”, *Journal of Information Technology Theory and Application (JITTA)*: Vol. 7: Iss. 1, Article 11, 2005.

[8] O. Findahl, ”Vad säger Internetstatistiken? (What do the Internet Statistics Tell Us?)”, Stiftelsen för internetinfrastruktur .SE, 2008.

[9] C. P. Larsen, C. Lindqvist, H. Pathirana, R. Lindström, E. Modin, and A. Aurelius, “ANT: The Acreo National Testbed – Status and Plans”, in proc. *NOC 2007*, Stockholm, Sweden, June 18-21, 2007.

[10] http://www.acreo.se/en/Technology-Areas/Broadband-Technology/Projects/Current-Projects/Acreo-National-Testbed [Accessed 10th April 2012].

[11] M. Kihl, C. Lagerstedt, A. Aurelius, and P. Ödling, “Traffic analysis and characterization of Internet user behavior”, In Proc. ICUMT 2010, pp. 224-231, Moscow, Oct. 2010.

[12] Procera networks, http://proceranetworks.com [Accessed 10th April 2012].

[13] W. S. Cleveland and D. X. Sun**, “**Internet Traffic Data”**,** Journal of the American Statistical Association, 1995, pp. 979-985. Reprinted in *Statistics in the 21st Century* , edited by A. E. Raftery, M. A. Tanner, and M. T. Wells, Chapman & Hall/CRC, New York, 2002.

A Low-Power Passive Mixer Receiver for Software Defined Radio (SDR) Applications

Pierre Bousseaud, Emil Novakov and Jean-Michel Fournier
Institut de Microélectronique, Electromagnétisme et Photonique (IMEP-LAHC),
MINATEC – 3 parvis Louis Néel, B.P. 257, 38016 Grenoble, France
e-mail: pierre.bousseaud@minatec.grenoble-inp.fr, novakov@minatec.grenoble-inp.fr, fournier@minatec.grenoble-inp.fr

Abstract— **Software Defined Radio (SDR) is a promising concept for the next generation of low-power portable wireless devices. Different communication standards can be processed just by changing the software in the transmitter. This approach allows great flexibility as well as hardware cost reduction. In this paper, we describe discrete-time passive mixer architecture for Radio Frequency (RF) direct sampling receivers. This architecture can achieve wideband quadrature demodulation and is suitable for SDR applications. The performances of the architecture are evaluated by simulation. Currently, we are working on the design of an Integrated Circuit (IC) that will implement a quadrature sampled I/Q receiver front-end. The IC is based on a 130nm CMOS technology. The receiver should be able to work in the 868MHz and 2.4GHz ISM bands. The expected power consumption is less than 2 mW.**

Keywords-Software defined radio; RF sampling; discrete-time mixer; linearity; QAM; I/Q; BER; noise figure; filtering; phase noise; blockers.

I. INTRODUCTION

The pervasive wireless applications and the presence of multiple communication standards based on different modulations, with variable channel bandwidth and carrier frequencies have motivated the development of multi-band and multi-standard radio communication technologies like the Software Defined Radio (SDR). On the other hand the CMOS technology's scaling allows now the fabrication of fully integrated radio transceiver chips able to work practically in any available frequency bands [1]. The final goal of the SDR is to receive a plurality of standards on just one single chip, whose functionalities can be updated by software only. The ideal SDR receiver hardware is composed of an antenna immediately followed by an Analog-to-Digital Converter (ADC) followed by the digital signal processing unit. Nevertheless, sampling a 2.4 GHz Radio Frequency (RF) signal at 4.8 Gsample/sec and 12 bit amplitude resolution would consume too much power which makes such approaches clearly unrealistic for miniaturized battery powered devices. In a more practical approach the ADC can be located at the Intermediate Frequency (IF) output of a frequency mixer (down-converter) in a standard super heterodyne RF receiver. Another possible solution is to use direct RF sampling mixer to down convert the wanted RF frequency band around the DC component and to amplify and process the band-pass signal at low frequency.

By using N paths sampling structures it is possible to sample the RF signal on N passive elements such as switched capacitors with both filtering and down-conversion at base-band [2]. This permits to greatly reduce the constraints for the ADC. Some receivers based on this principle have already been studied, implemented and tested by different research groups [3], [4], [5]. They show very interesting properties concerning the linearity and the power consumption. It is interesting to note that with such structures, very high Q pass-band filters can be realized at high frequency [4]. They easily replace RF SAW filters or LC structures which are bulky, costly and difficult to integrate.

The goal of our project is to design, simulate, implement and evaluate passive four paths mixer architecture for direct RF sampled conversion. In this paper we present the receiver architecture as well as the simulation results of its performances. The sampled receiver is designed to work with Quadrature Amplitude Modulated (QAM) signals but can also work with Amplitude Modulated (AM), Frequency Modulated (FM) and Phase Modulated (FM) signals. The complete receiver front-end will be further fabricated in a 130 nm CMOS RF technology.

Section II presents the RF sampling mixer structure and its basic parameters. Section III presents Matlab simulations of the receiver [6]. Section IV defines the implementation characteristics of the receiver and final conclusions are drawn in Section V. This paper is an intermediate report of a "work in progress".

II. FOUR PATHS MIXER ARCHITECTURE

A. Conventional and 4- paths structure

The conventional I/Q passive mixer structure is shown in Fig. 1. For a QAM modulated RF signal $V_{RF}(t) = A\cos(\omega_0 t+\varphi)$ where w_0 and φ are respectively the carrier pulsation and phase. During the symbol period A and φ are constant. In the passive sampling mixer, I and Q are obtained on two different sample/hold circuits by separating their sampling instant by $T_0/4$, where T_0 is the signal's period ($T_0 = 2\pi/\ \omega_0$). At the sampling time $t_0 = 0$, $I(t_0) = A\cos[\varphi]$. At the sampling time $t_1 = t_0 + T_0/4$, $Q(t_1) = A\cos(\varphi + \pi/2) = -\ A\sin(\varphi)$. In this way it is possible to find the magnitude A and phase φ of the RF signal in an orthogonal basis I/Q created by the voltage values sampled at the moment t_0 and t_1 as follows: $A = (I^2+Q^2)^{0.5}$ and φ = -

arctan(Q/I). Nevertheless, in such architecture due to the low conversion gain the noise figure is high, thus severely limiting the high frequency performances.

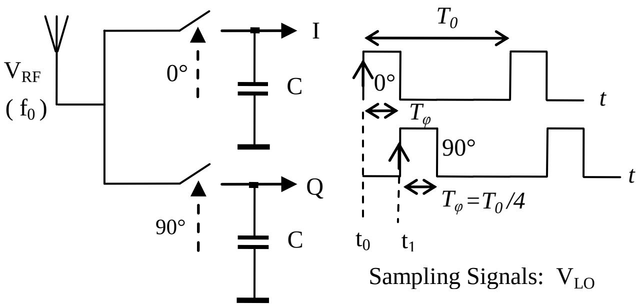


Figure 1. Classical I/Q sampling mixer

It is possible to improve the two phases I/Q sampling mixer by sampling the RF signal at four different points. The four phases mixer operation is based on the relationships: $Acos(\varphi + \pi) = - Acos(\varphi)$ and $Asin(\varphi + \pi) = - Asin(\varphi)$. The structure of the mixer is shown in Fig. 2. Now, *I* and *Q* redundant information are obtained by sampling data *I*+ and *I*- during phases 0° and 180° at and data *Q*+ and *Q*- during phases 90° and 270° respectively. If we subtract both *I*+ and *I*- as well as *Q*+ and *Q*- we'll have $I = I^{+} - (I-) = 2Acos(\varphi)$ and $Q = Q^{+} - (Q-) = 2Asin(\varphi)$. The power gain is improved by 6 dB and in the same time the Noise Factor (NF) is reduced by 3 dB for each *I* and *Q* paths.

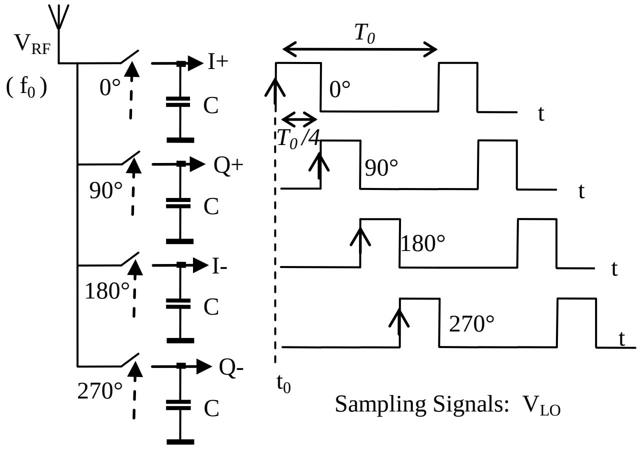


Figure 2. Passive four paths mixer (left) and clock scheme (right)

Further improvement can be obtained by using a fully differential topology. In this case, as shown in Fig. 3, the differential RF signal is sampled on eight different paths and the corresponding values are stored in two capacitors. The differential RF signal can be generated by either a transformer or a single-to-differential buffer amplifier before the sampler. This system is able to work without Low Noise Amplifier (LNA). The gain stage (G) can be moved to the base band signals *I* and *Q* thus increasing the noise figure of the receiver. This sampling scheme has multiple advantages. It is possible to reduce significantly the even order harmonic terms created by the non-linearity in the transistor forming the switch.

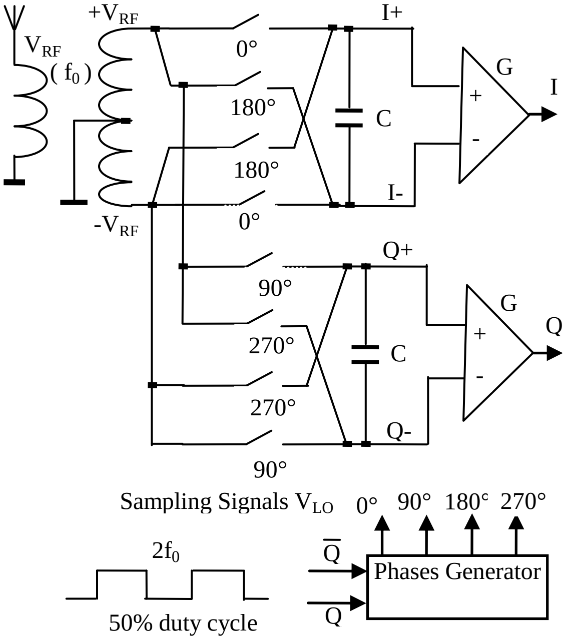


Figure 3. Differential sampling mixer and phase generator

At the sampling moment two opposite switches are activated at the same time. This allows the compensation of the charge injection phenomenon in the switch which causes a distortion in the amplitude of the sampled RF signal. Finally, the error caused by the clock feedthrough is completely rejected [6]. Indeed this error signals appears as a common mode voltage on both top and bottom plate of the capacitor *C*. At least 20 dB of improvement can be done concerning the IIP2.

B. Conversion Gain and Filtering

When the received signal V_{RF} and the sampling signal V_{LO} in Fig. 2 have the same frequency, the difference of frequency seen by each capacitor *C* is null and their equivalent impedance is infinite. Each capacitor behaves as an open circuit. In this case the mixer detects the complex envelope of the modulated signal V_{RF} without attenuation. If *D* is the duty cycle of V_{LO} ($D = T_{\varphi} / T_0$) the conversion gain (*CG*) of the sampler at the carrier frequency f_0 is given by:

$$CG = \sin c(D) \tag{1}$$

When the two signals V_{RF} and V_{LO} are with different frequencies each capacitor observes a periodic sliding of V_{RF}. The equivalent impedance is now non infinite. C starts to conduct and provides some voltage attenuation. If we consider one branch of the mixer, a first order low-pass filter if formed by the antenna resistance R_a, the conducting switch resistance R_{on} and the capacitor *C*. By choosing a time constant $\tau = (R_a + R_{on})C$ higher than the period of the modulated signal we can filter the input signal V_{RF}. If $f_{IF} = f_0 - f_{LO}$ is the difference of the frequency between the radio signal and the sampling signal and $f_C = 1/(2\pi\ \tau)$ is the cut-off frequency of the low-pass filter, the transfer function of the sampler at the base-band is given by [3]:

$$H(f_{IF}) = \frac{\sin c(D)}{1 + j\dfrac{f_{IF}}{Df_C}} \tag{2}$$

The equivalent bandwidth B of the sampler is:

$$B = \frac{D}{2\pi(R_a + R_{on})C} \tag{3}$$

The system is equivalent to a frequency converter and a band-pass filter at the same time. By changing the V_{LO} frequency it is possible to down convert V_{RF} and by choosing the duty-cycle it is possible to control the bandwidth of the converted signal.

C. *Harmonics rejection*

The sampling receiver architecture is based on the N paths filtering concept [2]. At the base band the sampler down converts the signals V_{RF} with harmonic frequencies multiple of f_{LO}. By combining signals from the different paths some harmonics can be suppressed. Using a N paths mixer, only the harmonics of the form $(Ni+1)f_{FLO}$ remain ($i=\pm 1, \pm 2, \pm 3$ …). For the four path mixer in Fig. 3 ($N=4$) the harmonics ($-5, -4, -2, -1, 2, 3, 4, 7$ …)f_{LO} are rejected.

D. *Noise*

In the mixer there are two main noise sources: the thermal noise of the antenna and the thermal noise of the switch resistance R_{on} when the switch it turned *on*. The major contributor to the noise is the resistance of the switch. That's why this resistance must be as low as possible. That imposes the use of larger transistors with higher Wide to Length (W/L) gate ratio. This also increases the parasitic capacitor between the gate and the source, which leads to degradation of the linearity due to more charge injection phenomenon [5]. That's why a trade-off must be found between linearity, noise factor and sizing of transistors realizing the switches. The Noise Figure (*NF*) of the system is given by [4]:

$$NF = (1 + \frac{R_{on}}{R_a})\frac{\pi^2}{8} \tag{4}$$

The best NF is obtained when R_{on} is null. In this case $NF = 0.9$ dB. With the four paths architecture, very low-noise figure can be achieved.

III. Matlab SimulatiOn Results

A. *System level architecture*

Using the Matlab Simulink software package [7], we have simulated the four phase down sampling mixer represented in Fig. 3. The simulation takes into account the phase noise of the sampling signal V_{LO}, the switch thermal noise and the presence of radio interferences and blockers at the mixer input. The radio frequency signal V_{RF} parameters have been chosen as follow: carrier frequency f_0 = 2.4 GHz, 16-QAM modulation and 9MHz RF bandwidth. The cut-off frequency is f_C = 20 MHz and the duty-cycle is D = 0.25. The sampling signals V_{LO} are generated from a 4.8 GHz external signal.

B. *Ideal transmitted signal*

Fig. 4 shows the received signal constellation for an ideally 16-QAM modulated signal without any source of perturbation added. The rotation of the constellation is due to the delay produced by the time constant τ of the receiver. This problem is not harmful and can be compensated by some calibration methods.

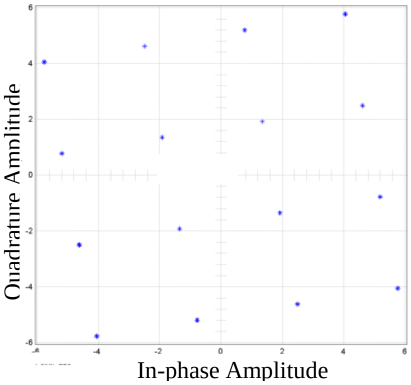


Figure 4. Ideal received constellation at the sampler output

C. *Phase noise*

The presence of phase noise in the sampling signal V_{LO} is a very important problem because it can degrade SNR at the output of the receiver and makes harder to detect the correct symbols. Simulations were performed with a phase noise of -102dBc@1MHz (1ps jitter). This is the maximum noise tolerated by the WiFi system. Fig. 5 shows the received constellation. Phase noise introduces some variations in the phase of received symbols which span an arc around its ideal value. Nevertheless the distortion is acceptable and received symbols still can be detected.

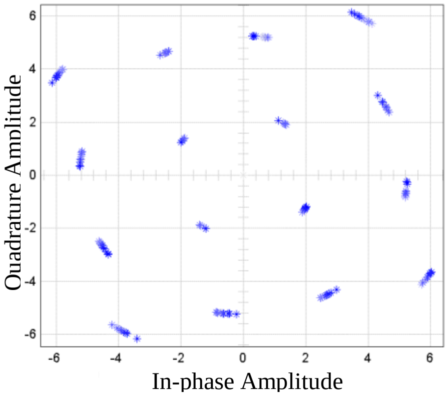


Figure 5. Received constellation with 1ps jitter@2.4GHz

D. Influence of the thermal noise

Fig. 6 shows the effect of the receiver’s thermal noise generated by the antenna and the switch resistances. By taking R_a = 50 Ω and R_{on} = 100 Ω, NF is about 5.68 dB for 20 MHz bandwidth. The noise power at the switch output is -95.3 dBm. Here we assume 10 dB of SNR at the output, which is 10 dB lower than the expected one for a correct 16-QAM demodulation. In Fig. 6 we can observe clouds of symbol randomly distributed around the ideal one.

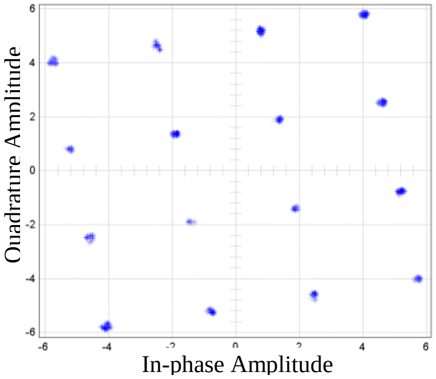


Figure 6. Received constellation with input AWG noise

E. Influence of interferers

Selectivity is one of fundamental criteria for a receiver and determines how much strong power blockers it is able to receive without deteriorating the Bit Error Rate (BER). Here we placed a sinusoidal interferer at a distance of 50 MHz from the modulated signal. The interference signal power is 10 dB higher than the power of the modulated signal. In Fig. 7 we observe symbols distributed on a circle around the ideal symbol position. This phenomenon is the consequence of AM to PM modulation in QAM systems. It shows the limits of the mixer for very close interferers in terms of filtering.

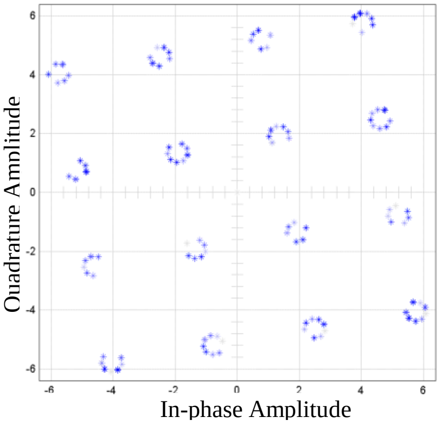


Figure 7. Received constellation with a 10dBm interferer@50MHz

That’s why some passive filtering needs to be made at the antenna level. Some digital filtering can also be done on the down converted signal.

IV. AIM OF THE WORK

The main objective of this work is to design, realize, test and evaluate the performances of a four paths sampling mixer implemented in a 130 nm CMOS technology. The Integrated Circuit (IC), based on the structure represented on Fig. 3, includes the mixer (switches and capacitors) the four phase generator, and the low frequency amplifiers (G). The mixer is designed to work in the ISM frequency band of 868 MHz and 2.4 GHz with special care to the low power consumption. The design target IC power consumption of less than 2 mW. The finality is to demonstrate the possibility to receive different communication standards with a simple structure with low power consumption, high linearity and low noise figure.

V. CONCLUSION

The present article describes low complexity four paths differential sampling mixer architecture. The passive mixer has the property to act also as a high Q pass-band filter. The mixer is flexible in frequency and achieves good linearity. It consumes low dynamic power, has low noise figure and is easy to integrate on silicon. It is a serious candidate for future SDR receivers and cognitive radio applications. Some Matlab simulation results are presented in the paper. A 130 nm CMOS integrated circuit implementing the mixer and the phase generator is under development. It is dedicated to be used in the 868 MHz and 2.4 GHz ISM frequency bands.

REFERENCES

[1] A. Abidi, “RF CMOS Comes of Age”, IEEE Journal of Solid-State Circuits, Vol. 39, No. 4, April 2004.

[2] L.E. Franks and J.W. Sandberg, “An alternative approach to the realization of network transfer functions : the N paths filter”, The Bell System Technical Journal, vol. 39, pp. 1321-1350, September 1960.

[3] M. Soer, E. Klumperink, Z. Ru, F. van Vliet and B. Nauta, “A 0.2-to-2.0GHz 65nm CMOS Receiver without LNA achieving >11dBm IIP3 and <6.5 dB NF”, In: IEEE International Solid-State Circuits Conference, ISSCC, San Francisco, U.S.A., 8-12 February 2009.

[4] C. Andrews and A. Molnar, “Implications of Passive Mixer Transparency for Impedance Matching and Noise Figure in Passive Mixer-First Receivers”, IEEE Transactions on Circuits and Systems-I: Regular papers, Vol. 57, No. 12, December 2010.

[5] "Discrete-Time Mixing Receiver Architecture for RF-Sampling Software-Defined Radio", IEEE Journal of Solid-State Circuits, Vol. 45, No. 9, September 2010.

[6] G. Wegmann et al. , “Charge injection in analog MOS switches”, IEEE Journal of Solid-State Circuits, Vol. SC-22, No. 6, December 1987

[7] Matlab, “The Wath Works”, version 7.1.11, August 2010

A Hybrid VOX System Using Emulated Hardware Behaviors

Eduardo Gonzalez
Ingram School of Engineering
Texas State University
San Marcos, TX 78666, USA
email: eg1196@txstate.edu

Stan McClellan
Ingram School of Engineering
Texas State University
San Marcos, TX 78666, USA
email: stan.mcclellan@txstate.edu

***Abstract*—This paper analyzes two well-known but complementary speech detection algorithms, and combines them to create a robust, low complexity method of speech detection. Software emulation of behaviors important in venerable hardware-based voice-operated switches is key to hybrid system performance. We test the hybrid system in the context of amateur radio, where speech and in-band data is accurately detected in real-time, even in the presence of significant noise.**

***Keywords*-silence detection, voice activity detection, VAD, VOX, voice-activated squelch, voice-activated switch**

I. Introduction

Speech detection plays an important role in applications where communication may be intermittent, or hands-free operation is desirable. Examples of this class of applications include emergency radio services, amateur radio, and communications for infrastructure maintenance and development. These environments require monitoring of communications channels for the presence of speech, which places a psychological strain on operators who must listen to constant noise and interference. Often, voice-operated switch systems are used to detect the presence of speech on a channel, and automatically "gate" the signal to an audio amplifier. Automated speech detection can effectively relieve operator strain and mute the speaker/receiver until active speech is present in the incoming transmission.

This paper analyzes two complementary approaches to speech detection, compares their operating characteristics, and presents a combination of elements to produce a hybrid, easily implemented and robust speech detection system. We focus on use of this approach in amateur radio systems and explore the performance and requirements of an automated squelch for convenient, hands-free operation. Conventional terminology among amateur radio operators uses the term "VOX" for a voice-activated switch, or voice-operated squelching unit. Thus, we refer to this system as a "VOX" in the remainder of this paper.

In Section II, we describe a venerable but popular hardware driven approach with some operational features which are very attractive for the user community. In Section III, we examine a software-driven approach which is similar to well-known pitch detection schemes, but optimized for low computational complexity. In Section IV, we describe the characteristics of a hybrid system which derives operational features from both of the preceding architectures. In Section V we evaluate the three complementary approaches and present performance comparisons, and Section VI concludes with observations about the examined systems and their application in real-time systems.

II. Hardware Driven Approach

In the 1970's, Motorola engineers developed a transistor circuit for hardware-based voice detection [1]. This circuit, which we refer to as the "MICOM" implementation, had very good characteristics for speech detection in noisy analog transmissions, and variants of this system were popular in the amateur radio community. Such variants include the Smart Squelch, popularized in *73 Magazine* [2] and an implementation by the Jet Propulsion Laboratories Amateur Radio Club [3] for retransmission of NASA Select Audio over the JPL voice/packet repeater network in Southern California.

The MICOM circuit was popular with amateur radio enthusiasts since it provided a simple and easily implemented speech detection subsystem. The MICOM VOX continuously monitors a specified channel, suppressing non-speech noise in the idle channel while allowing detected speech signals to activate the speaker.

MICOM-like circuits exploit the syllabic rate of human speech (3 syllables per second) and include a detector for short-term frequency modulation which is characteristic of voiced speech. The main components of MICOM implementations include a high gain amplifier, a trigger circuit to produce constant width pulses, a 3.25 Hz lowpass filter, comparators and timing circuitry to create hysteresis on the output "voicing" signal.

Motivated by the popularity and continued use of the MICOM VOX architecture [3] we performed an in-depth analysis of of this circuit to understand its behavior and model its features in a software simulation. First, we analyzed the MICOM circuit by hand and modeled it using a SPICE variant (MultiSim [4]) to accurately decompose its functional components. Then, we duplicated these functional

ISBN: 978-1-61208-193-9

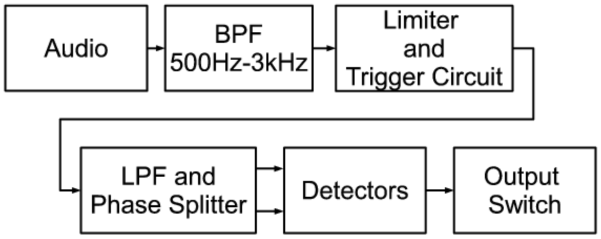


Figure 1. High level block diagram of MICOM algorithm.

components using a simulation package (Simulink [5]) to model the subsystems using signal processing algorithms.

In effect, we modeled the hardware implementation to extract performance measurements. This enabled a common reference to compare subsequent speech detection algorithms. Following subsections describe this process and illustrate the performance of the MICOM system.

A. MICOM Subsystems

To baseline MICOM performance, both the MultiSim and Simulink simulations used an 8kHz audio file which was manipulated through the stages shown in Figure 1. Each of the stages play an important role as described below:

- Band-Pass Filter (0.5 - 3 kHz): Removes non-voice-band energy.
- Limiter (85dB amplifier): Amplifies the signal so that non zero samples are saturated at the extrema. The effect of this function is a zero crossing detector for positive going excursions.
- Trigger Circuit (0.33ms pulses): Triggered by the amplified and limited voice band signal to create a steady stream of pulses that have uniform width, one per zero-crossing.
- Low-Pass Filter (3.25 Hz): Extracts the syllabic envelope from the pulse stream, estimating energy < 3 Hz.
- Phase Splitter: The first output of the phase splitter removes the DC component from the LPF output, and the second output inverts the resulting signal. This separates the original output of the LPF into a “top phase” and “bottom phase” for the detector.
- Detector: Creates a detection event if either of the phase voltages is above a manually-set threshold. This threshold must be set by the user each time a different channel is selected (ex. carrier frequency in amateur radio) or if the noise floor of the channel changes. In the analog implementation, a potentiometer provided decent control. However, in software, the tuning of this threshold becomes difficult.
- Output Switch: Incorporates a timing capacitor that creates a one second holdover from a single detection event. This is done in order to remove “dropouts” in the middle of active speech. The output switch also incorporates hysteresis by lowering the threshold whenever a detection event occurs. This, like the holdover capacitor, is intended to reduce false negatives.

B. MICOM Problems

Although the MICOM circuit was robust and simple to implement in an analog system, some subtleties of modeling analog phenomena make it less stable and more difficult to implement directly in a discrete time system. Certain components such as DC removal, which are simply capacitors in an analog circuit, become complicated in a discrete environment. Further, slight usability issues revolve around the threshold setting, which is sensitive and has small tolerance. Issues also arise whenever modulated data is transmitted on the channel, or when noise changes slowly producing localized energy < 3 Hz in the detection circuit.

Although much of the MICOM VOX functionality may have been supplanted by modern signal processing techniques, many of the MICOM operational characteristics are powerful and attractive to the user community. Thus, we attempt to model and emulate selected features in a discrete fashion.

III. Software Driven Approach

Robust speech detection systems often incorporate separate detection or classification of voiced and unvoiced speech. Many approaches to detection of voiced, unvoiced, and silence segments have been described in the literature, including for example: pitch detection [6], spectral characterizations [7], [8], and distance measures or statistical tests applied to harmonic and/or nonparametric models [9], [10].

However, in some classes of systems, detection of voiced segments is performed by subtracting estimated noise power from the output of a comb filter at the dominant frequency of the voiced speech. This result is compared to a threshold that determines whether speech is present. This type of “discriminate and threshold” system is functional, but presents a heavy computing load.

An approach to reducing the compute burden, which we refer to as the “Harris Algorithm,” provides an approximation of the voiced detector through a single lag autocorrelation process [11]. This method has been used by Harris Corp. to provide dynamic channel routing and activation for ADPCM (Adaptive Differential Pulse-Code Modulation) channel encoding.

The Harris Algorithm has several useful features for robust speech detection. However, in a complete implementation it may be lacking key features which are provided very effectively by aspects of the MICOM system.

A. Harris Subsystems

The Harris algorithm was designed in the 1990’s to meet the demand for a functional and simple voice detector [11]. For the purpose of this paper we summarize the general

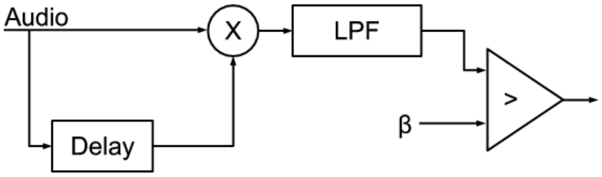


Figure 2. High level diagram showing the Harris algorithm components.

operation of the Harris algorithm and refer the reader to the literature for a complete discussion.

A block diagram of the Harris system is shown in Figure 2. The system incorporates a delay and multiply operation which essentially computes a running autocorrelation at a single pre-determined lag, according to Equation 1. In the equation, l is the fixed lag and $\bar{X}$ is the complex conjugate of X:

$$ACF(l) = \sum_{n} X_n \bar{X}_{n-l} \tag{1}$$

The output from this delay and multiply operation is fed into a simple lowpass filter implemented as an accumulator. The resulting low frequency component of the running autocorrelation is then compared to a threshold to determine the presence of speech. The effect of the Harris approach is to detect strong, stable correlations around the pre-determined lag value, which is related to pitch frequency.

B. Harris Problems

The Harris Algorithm performs well in detecting the onset of speech, but is inconsistent during active speech segments. The detect output has many false negatives within active speech, and resulting audio is choppy and incomprehensible. When the threshold is lowered to prevent these dropouts, the same results occur during silence intervals since the noise creates a high enough output to repeatedly trigger a detect event. Furthermore, since the Harris Algorithm relies on the low frequency components of the ACF, the slow spectral rolloff caused by an accumulator (a poor lowpass filter) allows low-frequency components to interfere with the approximation.

The core idea within the Harris approach is valuable, but by itself it does not provide a reliable system. The hybrid implementation described here uses aspects of the MICOM system to address these problems.

IV. Hybrid Approach

In order to achieve a robust hybrid speech detection algorithm, fundamental features of the MICOM circuit and the Harris Algorithm were taken into consideration and then extended. The components that are used from each system are outlined below, as well as the additional modifications made to increase detection speed, reduce false positives, and reduce the need for manual operation of the threshold.

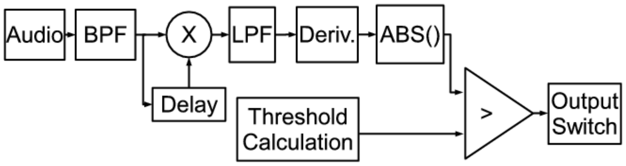


Figure 3. High level diagram showing the hybrid algorithm incorporating MICOM, Harris and new components.

A. Hybrid Inner Workings

Figure 3 presents a high level block diagram of the hybrid system. Each of the hybrid blocks is explained below:

- Band-Pass Filter (300-700 Hz): The BPF provides the same function as the BPF in the MICOM circuit but the voice band is decreased so that processing is done on more selective data.
- Delay and Multiply: Extracts short term periodicities in filtered audio. The delay chosen of 50 samples with a sampling frequency of 8000Hz provides smooth operation and good sensitivity.
- MICOM Low-Pass Filter: Instead of using a simple accumulator, the 3.25Hz lowpass filter from the MICOM circuit is used to extract syllabic rate information from the delay and multiply. This filter also provides a much sharper cut-off, eliminating unwanted frequency components that interfered with the estimation in the Harris algorithm.
- Derivative and Absolute Value: The derivative converts the slowly changing output of the LPF into a more defined and faster changing waveform which increases the tolerance and sensitivity of the threshold. Since the output of the LPF contains information about the changes in syllabic rate, like the phase splitter subsection of the MICOM circuit, both positive and negative deviations are important. The absolute value allows a single threshold to considers both deviations.
- Threshold Calculation: Removes the need for manual setting of the threshold value. To accomplish this, whenever speech is not detected, the energy of the noise is continuously calculated and the baseline threshold is established according to this changing energy level. This allows detection in varying noise floors.
- Modified MICOM Output Switch: Forces a holdover in detection via a counter that resets every time there is a detect event. The output is turned off only when the counter saturates to a holdover value. Instead of using a 1 second holdover (as in the MICOM circuit) the hybrid algorithm uses a 0.25s holdover which results in few dropouts and does not overly extend a detect event.

V. Performance and Comparisons

The hybrid algorithm accurately performs the VOX function in low-noise as well as high-noise conditions. Figure

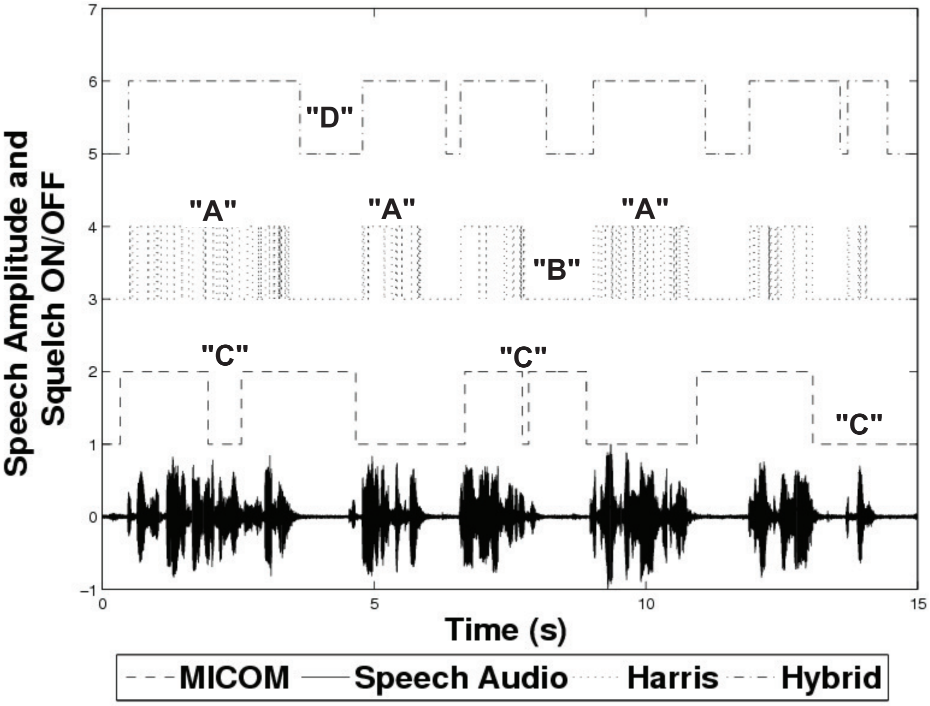


Figure 4. Performance of all three voice detection algorithms in a low noise, natural environment. The utterance was captured from an amateur radio transmission, and contains some non-speech noise. Annotations “A” through “D” indicate detection errors in each algorithm.

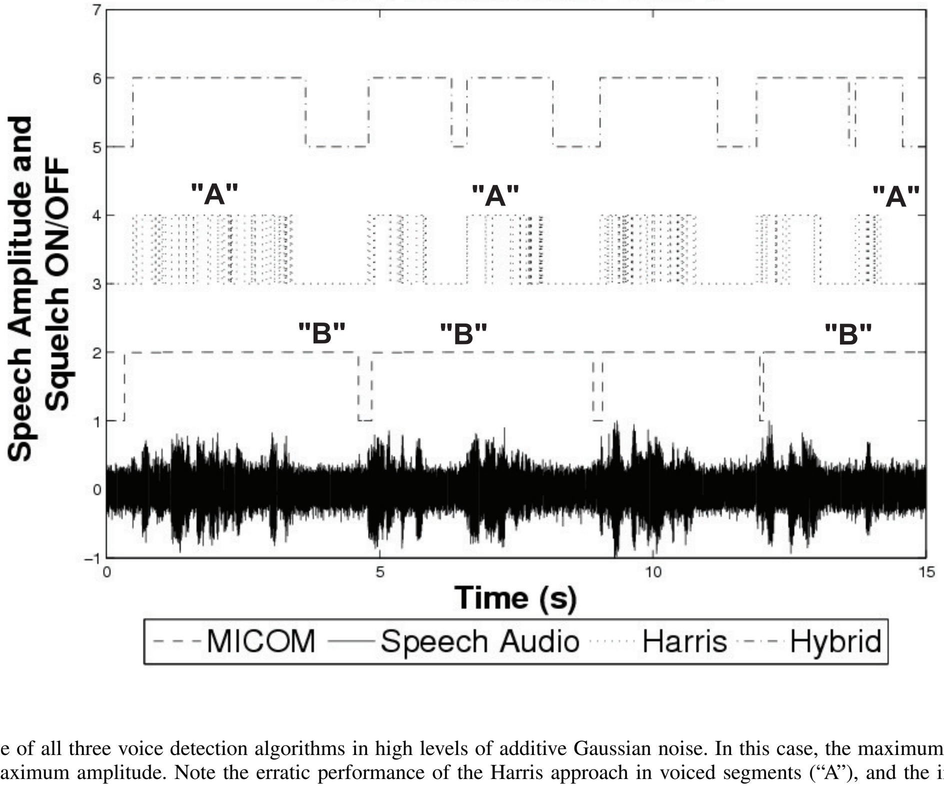


Figure 5. Performance of all three voice detection algorithms in high levels of additive Gaussian noise. In this case, the maximum noise amplitude is half the audio waveform maximum amplitude. Note the erratic performance of the Harris approach in voiced segments (“A”), and the inability of the MICOM approach to discriminate between noise and silence (“B”).

4 shows performance of the Harris, MICOM, and hybrid VOX implementations in low-noise conditions. Although Figure 4 seems to display a fairly "clean" or lab quality original signal, the signal is actually a speech utterance captured from an amateur radio transmission, and contains some objectionable, non-speech noise.

In the figure, several error conditions are labeled. Note the highly erratic performance of the Harris approach in voiced segments ("A"), but the ability of the Harris approach to reliably (albeit aggressively) determine non-speech segments ("B"). Also note the inaccurate voiced/non-voiced decisions of the MICOM approach ("C"). The hybrid approach typically produces accurate voicing indicators with acceptable overhang, and without aggressive penetration into non-voiced segments. There are a few exceptions (e.g. a missed onset at "D"). However, this style of performance is quite acceptable for real-time implementation, which avoids clipping, slow-attack, and other behaviors which are objectionable to amateur radio operators.

The performance of the Harris, MICOM, and Hybrid approaches in noisy environments is shown in Figure 5, where Gaussian noise was added to a speech signal to simulate poor quality amateur radio channels. The additive noise amplitude is adjusted to be half of the waveform's maximum, or 6dBV down from the signal's peak amplitude. In the figure, several error conditions are labeled.

The top trace of Figure 5 shows voicing indicators generated by the hybrid implementation, which accurately track the voicing segments of the original speech, even in the presence of significant additive noise.

The second trace shows voicing indicators generated by the Harris algorithm, which switches erratically between ON/OFF states during voiced segments ("A"), generating numerous false positives and false negatives for voiced and unvoiced speech, as well as inter-word gaps.

The third trace from the top of Figure 5 shows voicing indicators generated by our software emulation of the the MICOM VOX system. In noisy environments, the MICOM system remains in the ON or "voicing" state for the majority of the utterance, and has difficulty discriminating between noise and silence ("B").

Neither the MICOM VOX nor the Harris algorithm are sufficiently robust to generate stable voicing indicators in the presence of mild to moderate additive noise. Furthermore, and not discussed here in detail, the MICOM and Harris approaches are highly susceptible to colored noise, tone bursts, and in-band data.

The hybrid implementation works significantly better than the other two approaches even though the thresholds of the other systems were carefully set to extract maximum performance for the tests typified by Figure 5. In contrast with the other approaches, the hybrid system meshes the MICOM and Harris extremes together and tracks the speech in real-time, with minimal computational burden, and only a small, configurable detection delay.

To complete our analysis, the hybrid algorithm was also tested using several "in-band" data transmissions which are popular in amateur radio [12]. In-band tests included modulation schemes such as WSJT, CW1, PSK31, FSK, Pactor 1&2, and RTTY. Figure 6 provides the combined results of this testing. As shown in the figure, none of these modulation schemes triggered a speech detection event in the hybrid VOX, which would have been indicated by a low-to-high excursion of the voicing indicator. In the figure, the voicing indicator is shown as a dotted line just above each data sequence.

This testing demonstrates the robustness and stability of the hybrid approach in realistic applications and environments. These results are important in amateur radio and infrastructure applications where operators rely on hands-free VOX operation and robust voicing detection in noisy channels.

VI. Conclusion

The results of our comparison of VOX systems has shown that a combination of features from hardware-driven and software-driven approaches provides a robust and low complexity system capable of meeting important application requirements in a variety of environments.

In particular, amateur radio channels with in-band data transmissions and significant noise and non-speech interference are well-served by the hybrid VOX system. The approach described here combining venerable techniques with newer signal processing approaches and emulated hardware behaviors results in a stable, sensitive speech detection algorithm.

Further development and testing will improve the performance of the hybrid implementation in other environments and in different applications. Specifically, work is ongoing to compare the hybrid VOX system to well-known VAD schemes via standardized test frameworks, such as [13].

References

[1] *Service Manual for Motorola Micom HF SSB Transceiver*, Motorola, Inc., 1975, Part No. 68-81025E95A, The "Constant SINAD" Squelch was used in the Motorola Micom HF SSB Transceiver. The MICOM squelch board part number is TRN6175.

[2] F. Reid and L. David, "Smart Squelch for SSB," *73 Magazine*, pp. 44–49, Aug. 1982.

[3] J. Tarsala and R. Hammock, "The Jet Propulsion Laboratories 'Smart VOX'," Available: http://www.repeater-builder.com/projects/jpl-vox-sq/ssb-squelch.html, 2005.

[4] *Multisim 11.0*, National Instruments, Jul. 2011.

[5] *Simulink 2011a*, MathWorks Inc., Jan. 2011.

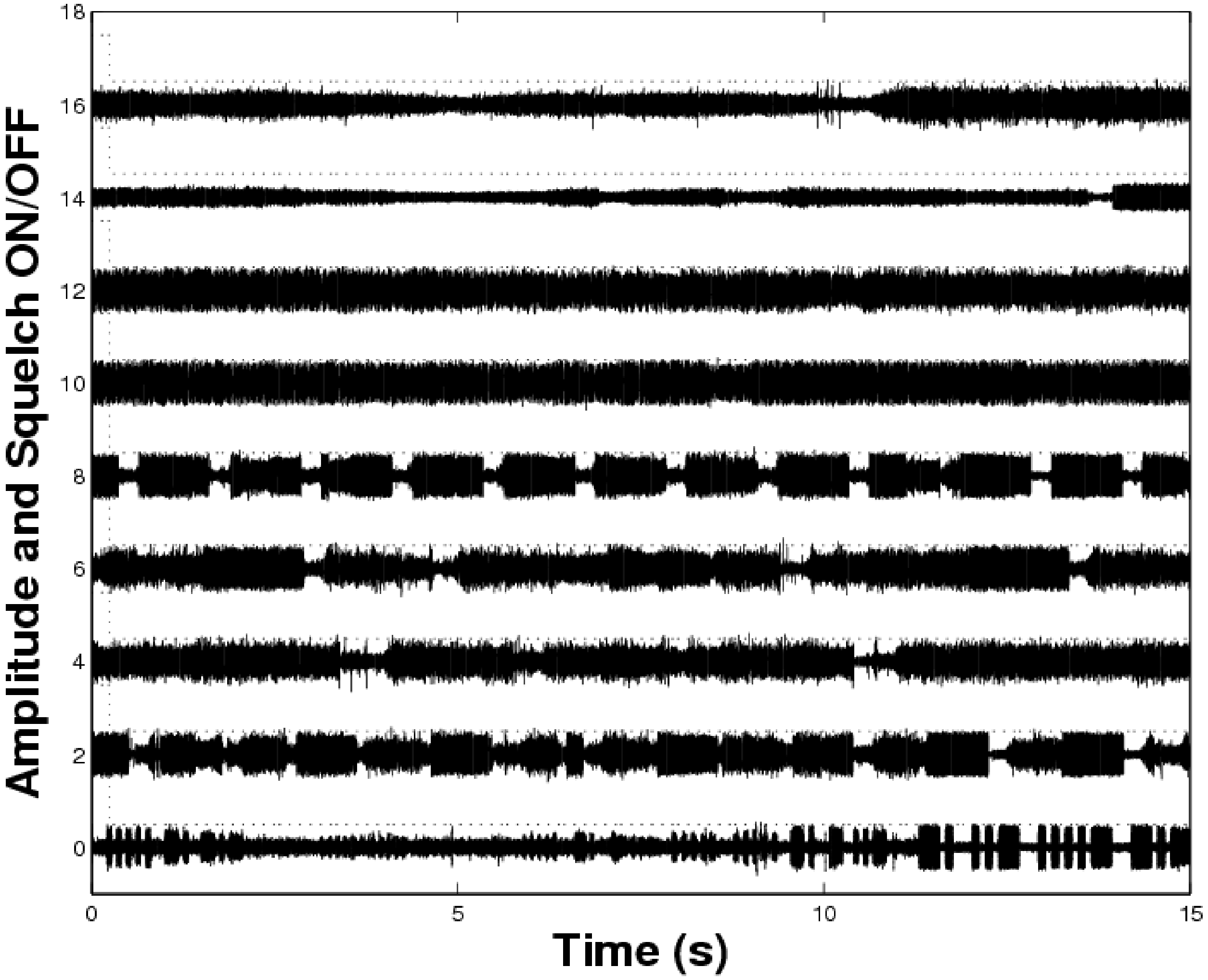


Figure 6. Performance of the hybrid system for common amateur radio data transmissions [12]. Top to bottom: WSJT, RTTY, Strong PSK31, PSK31, PACTOR2, PACTOR1, Noise, FSK, and CW1. The hybrid voicing indicator is shown as a dotted line just above each data sequence.

[6] L. Rabiner, M. Cheng, A. Rosenberg, and C. McGonegal, “A comparative performance study of several pitch detection algorithms,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 5, pp. 399–418, Oct. 1976.

[7] L. Rabiner and M. Sambur, “Application of an LPC distance measure to the voiced-unvoiced-silence detection problem,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 25, no. 4, pp. 338–343, Aug. 1977.

[8] S. McClellan and J. Gibson, “Variable-rate CELP based on subband flatness,” *IEEE Trans. Speech Audio Process.*, vol. 5, no. 2, pp. 120–130, Mar. 1997.

[9] E. Fisher, J. Tabrikian, and S. Dubnov, “Generalized likelihood ratio test for voiced-unvoiced decision in noisy speech using the harmonic model,” *IEEE Trans. Audio, Speech, and Language Proc.*, vol. 14, no. 2, pp. 502–510, Mar. 2006.

[10] B. Cox and L. Timothy, “Nonparametric rank-order statistics applied to robust voiced-unvoiced-silence classification,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 5, pp. 550–561, Oct. 1980.

[11] M. Webster, G. Sinclair, and T. Wright, “An efficient, digitally-based, single-lag autocorrelation-derived, voice-operated transmit (VOX) algorithm,” in *Military Communications Conference (MILCOM’91)*, vol. 3, Nov. 1991, pp. 1192–1196.

[12] *The ARRL Handbook for Radio Communications*, 89th ed., The American Radio Relay League, Newington, CT, Oct. 2011.

[13] *IEEE Std 269-2010 (Revision of IEEE Std 269-2002): IEEE Standard Methods for Measuring Transmission Performance of Analog and Digital Telephone Sets, Handsets, and Headsets*, IEEE, 2010.

About M2M standards

M2M and Open API

Manfred Sneps-Sneppe
Ventspils University College
Ventspils International Radioastronomy Centre
Ventspils, Latvia
manfreds.sneps@gmail.com

Dmitry Namiot
Lomonosov Moscow State University
Faculty of Computational Mathematics and Cybernetics
Moscow, Russia
dnamiot@gmail.com

***Abstract*—In this paper, we will discuss the current state of open APIs for M2M applications, as well as propose several possible changes and extensions. Our article based on open standards provided by ETSI. An open specification, presented as an Application Programming Interface (OpenAPI), provides applications with a rich framework of core network capabilities upon which to build services while encapsulating the underlying communication protocols. OpenAPI is a portable platform for services that may be replicated and ported between different execution environments and hardware platforms. We are proposing possible extensions for ETSI documents that let keep telecom development in sync with the modern approaches in the web development.**

Keywords-m2m; REST; open API; XML; web intents.

I. INTRODUCTION

As per classical definition from Numerex, Machine-to-Machine (M2M) refers to technologies that allow both wireless and wired systems to communicate with other devices of the same ability. M2M uses a device (such as a sensor or meter) to capture an event (such as temperature, inventory level, etc.), which is relayed through a network (wireless, wired or hybrid) to an application (software program), translates the captured event into meaningful information [1].

The next related acronym is Internet of things (IoT), referring to the networked interconnection of everyday objects [2]; it can be regarded as an extension of the existing interaction between humans and applications through the new dimension of "things" communication and integration. In IoT, devices are clustered together to create a stub M2M network, and are then connected to its infrastructure, i.e., the traditional "Internet of people" [3].

Considering M2M communications as a central point of Future Internet, European commission creates standardization mandate M/441 [4]. The Standardization mandate M/441, issued on 12th March 2009 by the European Commission to CEN, CENELEC and ETSI, in the field of measuring instruments for the development of an open architecture for utility meters involving communication protocols enabling interoperability, is a major development in shaping the future European standards for smart metering and Advanced Metering Infrastructures. The general objective of the mandate is to ensure European standards that will enable interoperability of utility meters (water, gas, electricity, heat), which can then improve the means by which customers' awareness of actual consumption can be raised in order to allow timely adaptation to their demands.

Besides the describing the current state of standards, our goal main here is the proposal for some new additions in M2M APIs architecture. We are going to propose web intents as add-on for the more traditional REST approach in order to simplify the development phases for M2M applications. The key moments in our proposals are: JSON versus XML, asynchronous communications and integrated calls.

The rest of the paper is organized as follows. Section II contains an analysis of M2M API standardization activities. In Section III, we consider Open API for M2M, submitted to ETSI. Sections IV and V are devoted to our offerings. In Section IV, we offer the never web tool – Web Intents for enhancement of M2M middleware. Sections V and VI are devoted to discussions.

II. THE CURRENT STATE OF M2M STANDARDS

Let us start from the basic moments. Right now, market players are offering own standards for M2M architecture. We refer to the recent ETSI TC M2M Workshop held on October 26-28, 2011. Figure 1 illustrates the basics of M2M infrastructure (as per ETSI) [5].

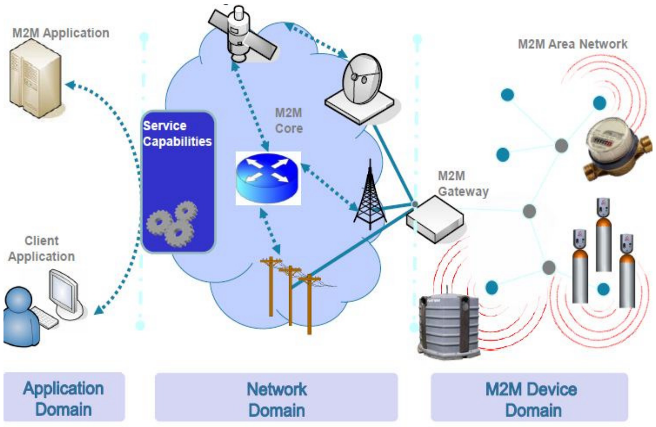


Figure 1. M2M infrastructure (as per ETSI)

The goals for M2M middleware are obvious. M2M middleware helps us with heterogeneity of M2M applications. Heterogeneity of service protocols inhibits the interoperation among smart objects using different service protocols and/or APIs. We assume that service protocols and API's are known in advance. This assumption prevents existing works from being applied to situations where a user wants to spontaneously configure her smart objects to interoperate with smart objects found nearby [6]. M2M API provides the abstraction layer necessary to implement interactions between devices uniformly. The M2M API provides the means for the device to expose its capabilities and the services it may offer, so that remote machines may utilize them. Consequently, such an API is necessary to enable proactive and transparent communication of devices, in order to invoke actions in M2M devices and receive the relating responses as well as the simplified management of resources.

ETSI is not the only source for the standardization in M2M area. Actually, ETSI has created a dedicated Technical Comittee for developing standards on M2M communications [7]. This structure aims at developing and maintaining an end-to-end architecture for M2M systems, as well as addressing various M2M communication considerations, such as naming, addressing, location, QoS, security, charging, management, application interfaces and hardware interfaces. Additionally, a major concern of the committee is to integrate sensor networks. The above-mentioned M/411 is just one example [12]. Other examples cover eHealth [8], Connected Consumer [9], City Automation [10] and Automotive Applications [11].

The 3rd Generation Partnership Project maintains and develops technical specifications and reports for mobile communication systems. Mobile networks are also concerned with the integration and support of M2M communications, as the nature of M2M systems is substantially differentiated than that of Human-to- Human services, i.e. plain telephone calls, which mobile networks originally addressed. Therefore, the 3GPP Technical Specifications Group dealing with Service and System Aspects [13], has issued a number of specifications dealing with requirements that M2M services and M2M communication imposes on the mobile network.

The Telecommunications Industry Association is the United States developing industry standards for a wide variety of telecommunication products. The standardization activities are assigned to separate Engineering Committees. The TR-50 Engineering Committee Smart Device Communications [14], has been assigned the task to develop and maintain physical-medium-agnostic interface standards, that will enable the monitoring and bi-directional communication of events and information between smart devices and other devices, applications or networks. It will develop a Smart Device Communications framework that can operate over different types of underlying transport networks (wireless, wired, etc.) and can be adapted to a given transport network by means of an adaptation/convergence layer.

The International Telecommunication Union as a specialized agency of the United Nations is responsible for IT and communication technologies. The Telecommunications Standardization Sector (ITU-T), covers the issue of M2M communication via the special Ubiquitous Sensor Networks-related groups [15]. ITU address the area of networked intelligent sensors.

Open Mobile Alliance (OMA) [16] develops mobile service enabler specifications. OMA drives service enabler architectures and open enabler interfaces that are independent of the underlying wireless networks and platforms. An OMA Enabler is a management object designated for a particular purpose. It is defined in a specification and is published by the Open Mobile Alliance as a set of requirements documents, architecture documents, technical specifications and test specifications. Examples of enablers would be: a download enabler, a browsing enabler, a messaging enabler, a location enabler, etc. Data service enablers from OMA should work across devices, service providers, operators, networks, and geographies.

As there are several OMA standards that map into the ETSI M2M framework, a link has been established between the two standardization bodies in order to provide associations between ETSI M2M Service Capabilities and OMA Supporting Enablers [17]. Specifically, the expertise of OMA in abstract, protocol-independent APIs creation, as well as the creation of APIs protocol bindings (i.e. REST, SOAP) and especially the expertise of OMA in RESTful APIs is expected to complement the standardization activities of ETSI in the field of M2M communications. Additionally, OMA has identified areas where further standardization will enhance support for generic M2M implementations, i.e. device management, network APIs addressing M2M service capabilities, location services for mobile M2M applications [18].

Actually, there should be a mapping of OMA service enablers to the ETSI M2M framework.

III. OPEN API FROM ETSI

This section describes an Open API for M2M, submitted to ETSI. By our opinion it is probably the most valuable achievement at this moment.

The Open API for M2M applications developed jointly in Eurescom study P1957 [19] and the EU FP7 SENSEI project makes. The OpenAPI has been submitted as a contribution to ETSI TC M2M [20] for standardization.

Actually, in this Open API we can see the big influence of Parlay specification. Parlay Group leads the standard, so called Parlay/OSA API, to open up the networks by defining, establishing, and supporting a common industry-standard APIs. Parlay Group also specifies the Parlay Web services API, also known as Parlay X API, which is much simpler than Parlay/OSA API to enable IT developers to use it without network expertise [21].

The goals are obvious, and they are probably the same as for any unified API. One of the main challenges in order to support easy development of M2M services and applications will be to make M2M network protocols "transparent" to applications. Providing standard interfaces to service and application providers in a network independent way will allow service portability [22].

At the same time, an application could provide services via different M2M networks using different technologies as long as the same API is supported and used. This way an API shields applications from the underlying technologies, and reduces efforts involved in service development. Services may be replicated and ported between different execution environments and hardware platforms [23]

This approach also lets services and technology platforms to evolve independently. A standard open M2M API with network support will ensure service interoperability and allow ubiquitous end-to-end service provisioning.

The OpenAPI provide service capabilities that are to be shared by different applications. Service Capabilities may be M2M specific or generic, i.e., providing support to more than one M2M application.

Key points for Open API:

- It supports interoperability across heterogeneous transports
- ETSI describes high-level flow and does not dictate implementation technology
- It is message-based solution
- It combines P2P with client-server model
- It supports routing via intermediaries

At this moment, all points are probably well developed except the message-based decision. Nowadays, publish-subscribe method is definitely not among the favorites approaches in the web development, especially for heavy-loading projects.

Main API sections are:

- Subscription and Notification (e.g., Publish/Subscribe).
- Grouping.
- Transactions.
- Application Interaction: Read, Do, Observe.
- Compensation (micro-payment).
- Sessions.

Let us provide more details for Open API categories and make some remarks:

Grouping

A group here is defined as a common set of attributes (data elements) shared between member elements. On practice, it is about the definition of addressable and exchangeable data sets. Just note, as it is important for our future suggestions, there are no persistence mechanisms for groups

Transactions

Service capability features and their service primitives optionally include a transaction ID in order to allow relevant service capabilities to be part of a transaction. Just for the deploying transactions and presenting some sequences of operations as atomic.

In the terms of transactions management, Open API presents the classical 2-phase commit model. By the way, we should note here that this model practically does not work in the large-scale web applications. We think it is very important because without scalability we cannot think about “billions of connected devices”.

Application Interaction

The application interaction part is added in order to support development of simple M2M applications with only minor application specific data definitions: readings, observations and commands.

Application interactions build on the generic messaging and transaction functionality and offer capabilities considered sufficient for most simple application domains.

Messaging

The Message service capability feature offers message delivery with no message duplication. Messages may be unconfirmed, confirmed or transaction controlled. The message modes supported are single Object messaging, Object group messaging, and any object messaging; (it can also be Selective object messaging); thinking about this as Message Broker.

Event notification and presence

The notification service capability feature is more generic than handling only presence. It could give notifications on an object entering or leaving a specific group, reaching a certain location area, sensor readings outside a predefined band, an alarm, etc.

It is a generic form. So, for example, geo fencing [32] should fall into this category too.

The subscriber subscribes for events happening at the Target at a Registrar. The Registrar and the Target might be the same object. This configuration offers a publish/subscribe mechanism with no central point of failure.

Compensation

Fair and flexible compensation schemes between cooperating and competing parties are required to correlate resource consumption and cost, e.g., in order to avoid anomalous resource consumption and blocking of incentives for investments. The defined capability feature for micro-payment additionally allows charging for consumed network resources.

It is very similar to Parlay’s offering [33] for Charging API. Again, it is a big question from the modern large-scale applications point of view: shall we develop a special API for the compensations or create a rich logging functionality where the external log processing should be responsible for the things as charging.

Sessions

In the context of OpenAPI, a session shall be understood to represent the state of active communication between Connected Objects

OpenAPI is REST based, so, the endpoints should be presented as some URI's capable to accept (in this implementation) the basic commands GET, POST, PUT, DELETE.

Actually, ETSI uses the Smart Meter profile as 'proof of concept' for the M2M service platform in Release 1.

For example: requests execution of some function.

URI: http://{nodeId}/a/do
Method: POST

Request

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<appint-do-request xmlns="http://eurescom.eu/p1957/openm2m">
<requestor>9378f697-773e-4c8b-8c89-27d45ecc70c7</requestor>
<commands>
<command>command1</command>
<command>command2</command>
</commands>
<responders>9870f7b6-bc47-47df-b670-2227ac5aaa2d</responders>
<transaction-id>AEDF7D2C67BB4C7DB7615856868057C3</transaction-id>
</appint-do-request>
```

Response

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<appint-do-response xmlns="http://eurescom.eu/p1957/openm2m">
<requestor>9378f697-773e-4c8b-8c89-27d45ecc70c7</requestor>
<timestamp>2010-04-30T14:12:34.796+02:00</timestamp>
<responders>9870f7b6-bc47-47df-b670-2227ac5aaa2d</responders>
<result>200</result>
</appint-do-response>
```

Note that because we are talking about server-side solution, there is no problem with so called sandbox restrictions. But, it means of course, that such kind of request could not be provided right from the client side as many modern web applications do.

IV. The modern web vs. Open API from ETSI

Let us describe the proposed standards from the modern web development points of view. As seems to us it is a correct approach, because Open API declares REST support right for the web development. In other words, support for web developers as the first class citizens is one of the obvious goals for ETSI.

It is almost impossible for developers to anticipate every new service and to integrate with every existing external service that their users prefer and thus they must choose to integrate with a few select APIs at great expense to the developer.

As per telecom experience we can mention here the various attempts for unified API that started, probably, with Parlay. Despite a lot of efforts, Parlay API's actually increase the time for development. It is, by our opinion, the main reason for the Parlay's failure.

Web Intents solves this problem. Web Intents is a framework for client-side service discovery and inter-application communication. Services register their intention to be able to handle an action on the user's behalf. Applications request to start an action of a certain verb (for example share, edit, view, pick etc.) and the system will find the appropriate services for the user to use based on the user's preference. It is the basic [24].

Intents play the very important role in Android Architecture. Three of the four basic OS component types - activities, services, and broadcast receivers - are activated by an asynchronous message called as intent.

Intents bind individual components to each other at runtime (you can think of them as the messengers that request an action from other components), whether the component belongs to your application or another.

Created intent defines a message to activate either a specific component or a specific type of component - an intent can be either explicit or implicit, respectively.

For activities and services, an Intent defines the action to perform (for example, to "view" or "send" something) and may specify the URI of the data to act on (among other things that the component being started might need to know). For example, our intent might convey a request for an activity to show an image or to open a web page. In some cases, you can start an activity to receive a result, in which case, the activity also returns the result in an Intent (for example, you can issue an intent to let the user pick a personal contact and have it returned to you - the return intent includes a URI pointing to the chosen contact) [25].

Going to M2M applications, it means that our potential devices will be able to present more integrated data for the measurement visualization for example. The final goal of any M2M based application is to get (collect) measurements and perform some calculations (make some decisions) on the collected dataset. We can go either via low level APIs or use (at least for the majority of use cases) some integrated solutions. The advantages are obvious. We can seriously decrease the time for development.

Web Intents puts the user in control of service integrations and makes the developers life simple.

Here, is the modified example for web intents integration for the hypothetical web intents example:

1. Register some intent upon loading our HTML document

```
    document.addEventListener("DOMContentLoaded",
function() {
        var regBtn = document.getElementById("register");
        regBtn.addEventListener("click", function() {
      window.navigator.register("http://webintents.org/m2m",
undefined);    }, false);
```

2. Start intent's activity

```
        var startButton =
document.getElementById("startActivity");
        startButton.addEventListener("click", function() {
          var intent = new Intent();
          intent.action = "http://webintents.org/m2m";
            window.navigator.startActivity(intent);  }, false);
```

3. Get measurements (note – in JSON rather than XML) and display them in our application

```
        window.navigator.onActivity = function(data) {
          var output = document.getElementById("output");
          output.textContent = JSON.stringify(data);
        }; }, false);
```

Obviously, that it is much shorter than the long sequence of individual calls as per M2M Open API.

The key point here is *onActivity* callback, which returns JSON (not XML!) formatted data. In contrast, as per suggested M2M API, we should perform several individual requests, parse XML responses for the each of them and only after that make some visualization. Additionally, Web Intents based approach is asynchronous by its nature, so, we don need to organize asynchronous calls by our own.

Also, Web Intents approach let us bypass sandbox restrictions. In other words developers can raise requests right from the end-user devices, rather than always call the server. The server-side only solution becomes bottleneck very fast, and vice-versa, client side based request let developers deploy new services very quickly. Why do not use the powerful browsers in the modern smart-phones? At the end of the day Parlay spec were born in the time of WAP and weak phones. Why do we ignore HTML5 browsers and JavaScript support in the modern phones?

Generally speaking, we expect the initiatives from software companies that will opposite to telecom approach. For example, Paho project [26] (IBM et al.) directly declares the need to provide open source implementations of open and standard messaging protocols that support current and emerging requirements of M2M integration with Web and Enterprise middleware and applications. It will include client implementations for use on embedded platforms along with corresponding server support as determined by the community. This will enable a paradigm shift from legacy point-to-point protocols and the limitations of protocols like SOAP or HTTP into more loosely coupled yet determinable models. It will bridge the SOA, REST, Pub/Sub and other middleware architectures already well understood by Web 2.0 and Enterprise IT shops today, with the embedded and wireless device architectures inherent to M2M.

We think that XML days are over, JSON (and especially JSONP) is a key.

But, here goes the next big question: persistence.

V. DATA PERSISTENCE

The next question we would like to discuss relating to the M2M API's is probably more discussion able. Shall we add some persistence API (at least in the form of generic interface)?

The reasons are obvious – save the development time. Again, we should keep in mind that we are talking about the particular domain – M2M. In the most cases our business applications will deal with some metering data. As soon as we admit, that we are dealing with the measurements in the various forms we should make, as seems to us a natural conclusion – we need to save the data somewhere. It is very simple – we need to save data for the future processing.

So, the question is very easy – can we talk about M2M applications without talking about data persistence? Again, the key question is M2M. It is not abstract web API. We are talking about the well-defined domain.

As seems to us, even right now, before the putting some unified API in place, the term M2M almost always coexists with the term "cloud". And as we can see, almost always has been accompanied by the terms like automatic database logging, backup capabilities etc.

So, maybe this question is more for the discussions or it even could be provocative in the some forms, but it is: why there is no reference API for persistence layer in the unified M2M API? It is possible in general to create data gathering API without even mentioning data persistence? Shall we define cloud database API as a part of M2M standard or not?

The use of cloud computing means that data collection, processing, interface, and control can be separated and distributed to the most appropriate resource and device. In contrast, currently M2M implementations tend to combine data collection, processing, interface, and control.

Once transmitted to the cloud, data can be stored, retrieved and processed without having to address many of the underlying computing resources and processes traditionally associated with databases. For M2M applications, this type of virtualized data storage service is ideal [27]

As soon as ETSI standards define the interfaces, the developers we will be able to introduce various implementations. For example, it looks like NoSQL solutions are perfect fit for M2M applications.

These data stores operate by using key-value associations, which allows for a flatter non-relational form of association. NoSQL databases can work without fixed table schemes. It makes easy to store different data formats as well as change and expand formats over time. It is very important for M2M applications (as well as for any type of applications tied with hardware). There are no "unified" devices in the real word. We simply cannot create an efficient schema that will serve all the devices (including new entrants). So M2M stores should be schema-less.

NoSQL databases could be easily scaled horizontally. Data is distributed across many servers and disks. Indexing is performed by keys route the queries to the datastore for the range that serves that key. This means different clusters respond to requests independently from other clusters, what increases throughput and response times. Quick adding new servers, database instances and disks and changing the ranges of keys can accommodate growth.

There are more then enough NoSQL systems on the market, they all have own APIs, so the question for M2M standardization body becomes even more important: shall we include the "unified" interface to data store into standard?

Suppose we do not as it is now. Does it mean that for OMA interfaces for example we will define own persistence approach each time we need data saving?

The topic that is tight linked with data persistence is a cloud. Obviously, for big data we should be able to integrate the information gathered via M2M into a large virtual information platform in a cloud [28]. This moment is completely missed in Open API. Shall we live with it, shall we pass problem to OMA enablers or what? As seems to us, this question should be addressed and answered.

We think, that in addition to developing open interfaces and standard system architectures, M2M ecosystems also need to establish a set of common software and hardware platforms to substantially reduce development costs and improve time to market.

VI. New signaling demand

Eventually, billions of devices — such as sensors, consumer electronic devices, smart phones, PDAs and computers — will generate billions of M2M transactions. For example, price information will be pushed to smart meters in a demand-response system. Push notifications will be sent to connected devices, letting a client application know about new information available in the network. The scale of these transactions will go beyond anything today's largest network operators have experienced. Signaling traffic will be the primary bottleneck as M2M communications increase. Alcatel-Lucent Bell Labs traffic modeling studies support this by comparing network capacity against projected traffic demand across multiple dimensions (such as signaling processing load on the radio network controller, air-interface access channel capacity, data volume and memory requirement for maintaining session contexts). The limiting factor is likely to be the number of session set-ups and tear-downs. For the specific traffic model and network deployment considered in the study, it is seen that up to 67 percent of computing resources in the radio network controller is consumed by M2M applications. [29].

How much of the traffic sent is network overhead? As an analysis carried on by A. Sorrevad [30] shows for ZigBee solution, a node is sending at least 40 Mbytes per year with the purpose of maintaining the network and polling for new data. The trigger data traffic for a year is much less - around1-10 Mbytes. Thus we see that the relationship between network and trigger traffic can range between 40:1 to 4:1 in a ZigBee solution that is following the home automation specification.

The traffic sent when maintaining a 6LoWPAN network is application specific. The relationship between network and trigger traffic can then be in the range 2:1 to 5:1.

As per [31], we can describe the several traffic-related issues for M2M. Time-controlled traffic is transmitted and received at periods of time that are defined well in advance. Time-tolerant traffic can support significant delays in data transmission and reception. This implies that the system can give lower access priority to or defer data transmission of time-tolerant traffic. When data traffic is "one way," it is only control signaling that is transmitted in the opposite direction. Digital signage and consumer devices are use cases where data may be device-terminated. One-way traffic may require changes to the network entry and addressing protocols. Extremely low latency requires that both network access latency and data transmission latency be reduced. This feature is required in many emergency situations (e.g., healthcare). Changes to the bandwidth request and network entry/re-entry protocols may be required to support extremely low latency. Infrequent traffic is common in many M2M use cases. This feature may enable sleep/idle mode improvements that save power and channel resources.

Due to the salient features of M2M traffic it may not be supported efficiently by present standards [31].

Why do we think also it is a time for traffic-related talks? By our opinion the reason is very simple. It is not obvious exactly how can we support transactional APIs (as per ETSI draft), without the dealing with the increased traffic. Simply – in our transactions we need the confirmation that device is alive, that operation has been performed, etc. All this is signaling traffic. Actually, this may lead to next provocative questions: do we really need transactional calls for all use cases? For example, the modern large-scale web applications (e.g. social networks) are not transactional internally.

VII. Conclusion

In this article, we briefly described the current state for the open unified M2M API from ETSI. We proposed some new additions – Web Intents as add-on for the more traditional REST approach. The main goal for our suggestions is the simplifying the development phases for M2M applications. The key advantages are JSON versus XML, asynchronous communications and integrated calls. Also we would like to point attention of readers to the couple of important questions that are not covered yet by our opinion: data persistence, clouds and signaling traffic.

VIII. Acknowledgement

The paper is financed from ERDF's project SATTEH (No. 2010/0189/2DP/2.1.1.2.0/10/APIA/VIAA/019) being implemented in Engineering Research Institute «Ventspils

International Radio Astronomy Centre» of Ventspils University College (VIRAC).

REFERENCES

[1] Bob Emmerson, "M2M: The Internet of 50 Billion Devices", WinWin Magazine, Jan. 2010, pp.19-22.

[2] Commission of the European communities, Internet of Things in 2020, EPoSS, Brussels, 2008.

[3] Antoine de Saint-Exupery, Internet of Things – Strategic Research Roadmap, Sep. 15, 2009. http://www.internet-of-things-research.eu Retrieved: Jan, 2012

[4] Standartisation mandate to CEN, CENELEC and ETSI in the field of measuring instruments for the developing of an open architecture for utility meters involving communication protocols enabling interoperability, European Commission, M/441, 2009.

[5] ETSI Machine-to-Machine Communications http://www.etsi.org/website/technologies/m2m.aspx Revised: Feb, 2012

[6] Hyunho Park, Byoungoh Kim, Yangwoo Ko, and Dongman Lee, "InterX: A service interoperability gateway for heterogeneous smart objects" in in: Pervasive Computing and Communications Workshops (PERCOM Workshops), 2011 IEEE International Conference 21-25 March 2011 pp. 233 - 238.

[7] http://www.etsi.org/Website/Technologies/M2M.aspx Retrieved: Jan, 2012

[8] ETSI TR 102 732 V0.3.1, Machine to Machine Communications (M2M); Use cases of M2M applications for eHealth.

[9] ETSI TR 102 857 V0.3.0, Machine to Machine Communications (M2M); Use cases of M2M applications for Connected Consumer

[10] ETSI TR 102 897 V0.1.1, Machine to Machine Communications (M2M); Use cases of M2M applications for City Automation

[11] ETSI TR 102 898 V0.4.0, Machine to Machine Communications (M2M); Use cases of Automotive Applications in M2M capable networks.

[12] ETSI TR 102 691 V1.1.1, Machine-to-Machine communications (M2M); Smart Metering Use Cases

[13] 3GPP TS 22.368 V11.0.1, 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Service requirements for Machine-Type Communications (MTC); Stage 1, (Release 11)

[14] TR-50 standards http://www.tiaonline.org/standards/ committees/committee.cfm?comm=tr-50 Retrieved: Jan, 2012

[15] Jea-Il Han, Anh-Duy Vu, Jin-Won Kim, Jun-Soo Jeon, Seung-Min Lee, and Young-Man Kim The fundamental functions and interfaces for the ITU-T USN middleware components Information and Communication Technology Convergence (ICTC), 2010 International Conference on 17-19 Nov. 2010 pp.: 226 – 231 Print ISBN: 978-1-4244-9806-2 DOI=10.1109/ICTC.2010.5674664

[16] OMA http://www.openmobilealliance.org/ Retrieved: Jan, 2012

[17] Niklas Blum, Irina Boldea, Thomas Magedanz, and Tiziana Margaria Service-oriented access to next generation networks: from service creation to execution Journal Mobile Networks and Applications archive Volume 15 Issue 3, June 2010 Kluwer Academic Publishers Hingham, MA, USA DOI=10.1007/s11036-010-0222-1 Retrieved Feb, 2012

[18] IoT project: http://www.iot-a.eu/public Retrieved Feb, 2012

[19] EURESCOM project P1957, Open API for M2M applications, http://www.eurescom.de/public/projects/P1900-series/P1957/. Retrieved Feb, 2012

[20] Draft ETSI TS 102 690 V0.13.3 (2011-07) Technical Specification

[21] Jong-choul Yim, Young-il Choi, and Byung-sun Lee Third Party Call Control in IMS using Parlay Web Service Gateway Advanced Communication Technology, 2006. ICACT 2006. The 8th International Conference Issue Date: 20-22 Feb. 2006 pp. 221 – 224

[22] Grønbæk I., Architecture for the Internet of Things (IoT): API and interconnect, The Second International Conference on Sensor Technologies and Applications, IEEE August 2008, DOI 10.1109/SENSORCOMM.2008.20, 809.

[23] Inge Grønbæk and Karl Ostendorf Open API for M2M applications In: ETSI M2M Workshop Oct. 2010

[24] Web Intents http://webintents.org/ Retrieved: Feb, 2012

[25] Android Developers http://developer.android.com/guide/topics/fundamentals.html Retrieved: Jan, 2012

[26] Paho project: http://eclipse.org/proposals/technology.paho/ Retrieved: Jan, 2012

[27] Cloud + Machine-to-Machine: http://www.readwriteweb.com/cloud/ 2011/03/cloud-machine-to-machine-disruptive-innovation-part-1p2.php Retrieved: Jan, 2012

[28] T. Osawa Practice of M2M Connecting Real World Things with Cloud Computing FUJITSU Sci. Tech. J. vol. 47 No. 4 pp. 401-407

[29] Harish Viswanathan, "Getting Ready for M2M Traffic Growth" http://www2.alcatel-lucent.com/blogs/techzine/2011/getting-ready-for-m2m-traffic-growth/ Retrieved: Jan, 2012

[30] A. Sorrevad M2M Traffic Characteristics, KTH Information and Communication Technology Master of Science Thesis Stockholm, Sweden 2009 TRITA-ICT-EX-2009:212 http://web.it.kth.se/~maguire/DEGREE-PROJECT-REPORTS/091201-Anders_Orrevad-with-cover.pdf Retrieved: Feb, 2012

[31] Geng Wu, Shilpa Talwar, Kerstin Johnsson, Nageen Himayat, and Kevin D. Johnson, M2M: From Mobile to Embedded Internet IEEE Communications Magazine, April 2011 pp. 35-43

[32] A. Greenwald, G. Hampel, C. Phadke, and V. Poosala An economically viable solution to geofencing for mass-market applications Bell Labs Technical Journal Special Issue: Application Enablement Volume 16, Issue 2, September 2011 pp. 21–38,

[33] SunHwan Lim, JaeYong Lee, and ByungChul Kim Open API and System of Short Messaging, Payment, Account Management Based on RESTful Web Services Advanced Communication and Networking Communications in Computer and Information Science, 2011, Volume 199, pp. 66-75, DOI: 10.1007/978-3-642-23312-8_9

Adaptive Resource Allocation Mechanism for Broadband Mobile Network

Moo Wan Kim
Tokyo University of Information Sciences
Chiba, Japan
mwkim@rsch.tuis.ac.jp

Shintaro Uno
Aichi University of Technology
Gamagoori, Japan
Uno-shin@aut.ac.jp

***Abstract*—Mobile networks have been continuously developed from 3.5G to 3.9G/4G with high speed wireless technologies (i.e., broadband mobile network). So, mobile networks need to provide more sufficient QoS mechanism to provide enhanced user's satisfaction. In this paper, we propose a new adaptive resource allocation mechanism based on utility function borrowed from the field of microeconomics. Through the simulation and calculation evaluation, we show that adaptive resource allocation based on user preferences is effective.**

Keywords- QoS; utility function; mobile network;

I. INTRODUCTION

Currently mobile networks are being evolved from 3.5G to 3.9G/4G, which is broadband mobile network, integrated into IP core network based on IMS (IP Multimedia Subsystem). In this environment, it is essential to provide sufficient QoS (Quality of Service) mechanism which enables different services to provide enhanced user satisfaction. QoS for mobile networks is being studied mainly in 3GPP [1], but still deal with only class oriented mechanism. For example, QoS class is defined according to the service types, and mapping between QCI (QoS Class Identifier) and DSCP (DiffServ Code Point) is studied [2] [3]. So, it is required that QoS mechanism has to provide more flexible resource allocation mechanism based on each user's preferences.

In this paper, we propose a new adaptive resource allocation mechanism based on each user's preferences by using the utility function borrowed from the field of microeconomics. The utility function qualifies the value that a user perceives for all possible amount of resources allocated. The field of microeconomics addresses the issue of resource allocation when many users compete for a limited amount of resource. Microeconomics has been used to address problems in networks by several authors. Some authors use it to address pricing issues [4] [5] [6] [7], and others to the problem of resource allocation [8] [9].Utility functions have been used by other authors for resource allocation in networks [7] [10]. These papers develop distributed mechanisms for resource allocation assuming that the network does not have knowledge of the utility functions of users. Our work is different in that we assume that the network has knowledge of the utility functions of all users.

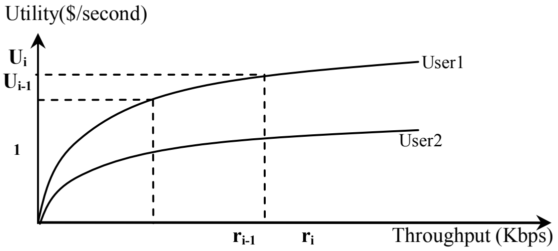


Figure 1. Example of Utility Function

II. BASIC TECHNOLOGIES

A. Utility Function

Utility is defined in the field of microeconomics that the level of satisfaction acquired from the consumption of properties such as services or commodities [11]. The user's total utility obtained from a network service will depend on several QoS metrics, such as throughput, delay, and jitter. In the throughput perspective, the user's utility depends on the bandwidth availability in the network to satisfy the resource requirement of service. Fig. 1 shows the example of utility function regarding the throughput allocated to user. User should define or select multiple utility functions for each service because the utility functions will be different from the kinds of services (e.g., streaming service).

Let us consider the network situation in which we have M users in the system. We let $U_{i}(r_{i})$ denote the utility derived by user's flow i for a bandwidth allocation r_i and C is the total link capacity. User i is allocated $\underline{r}_i$ units of resource that is the i-th component of the solution r=[r_1, r_2, …, r_M] of the following optimization problem: [12]

$$\operatorname*{Max}_{[r1,r2,\ \ldots\ r_M]} \sum_{i=1}^{M} U_{i}(r_{i}) \qquad (1)$$

$$\text{subject to } \sum_{i=1}^{M} r_i \leq C$$

$$r_i \geq 0 \text{ for } i = 1,2,..,M$$

In this network situation, in order to adopt utility function to the network QoS management, the following requirements should be satisfied.

- Continuous utility functions should be represented in discrete functions in order to allocate the resource unit where the discrete segmental value of utility function is compared each other.
- Resource allocation should be done based on discrete utility function satisfying above (1)

Fig. 2 shows the example of discrete utility functions, which is the case that the two lines are differentiated by users. This value can be thought as the price that a user would be willing to pay to obtain a specific amount of resource. For example, in the upper line to obtain the throughput $r_{i-1,}$ a user will pay U_{i-1} ($/second) and to obtain the throughput $r_{i,}$ a user will pay U_{i}($/second). In Fig. 2, the slope of utility (ΔU) is more important since it means a unit price for a unit bandwidth. If a flow is a high priority flow, it has a higher ΔU than other flows.

Following, there are the properties of discrete utility functions:

- Non-negativity: $U(r) \geq 0$ for all $r \geq 0$. Obviously the users cannot associate a negative utility with a positive resource allocation.
- Non-decreasing nature: $U(r)$ has to be a non-decreasing function. Clearly also is the fact that users cannot associate a higher utility with a smaller allocation that with a higher allocation.

B. MSS Resouce Allocation Algorithm

To satisfy (1), the MSS (Maximum Segmental Slope) resource allocation algorithm [12] has been developed by our colleagues based on the standard optimization problem solving method [15]. The main characteristic of this algorithm is that it allows resource allocation maximizing the user's satisfaction by allocating a unit of resource firstly to the flow that has the highest segmental slope. But, this algorithm is rather straightforward and it should be updated to enhance the performance. For example, some heuristic sorting algorithm should be considered because the major portion of the algorithm is a sorting process.

.

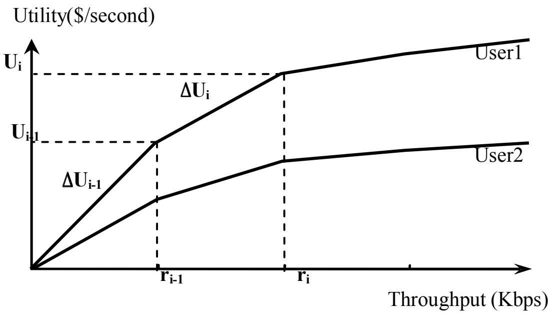


Figure 2. Example of Discrete Utility Function

III. ADAPTIVE RESOURCE ALLOCATION MECHANISM

A. Network Architecture Model

We assume the conceptual network architectural model as depicted in Fig. 3, based on 3.9G network (i.e., LTE) which comprises RAN(Radio Access Network) with MNs (Mobile Nodes), and CN (Core Network) with two main players: the QS (QoS Server) and the MM(Mobility Manager) [7]. The ARs (Access Router) in CN are key control points in the network. They are IP routers that are one IP hop distant from the mobile node via BS (Base Station, i.e., e-Node B in case of LTE). All data packets to and from the mobile node, and signaling messages between the mobile node and various servers in the network pass through the ARs. The CRs (Core Router) are high speed routers that lie in the core network. The main function of QS is an admission control based on the adaptive resource allocation mechanism proposed in the next section. The MM manages MN's location information where the MN is located in the mobile network, and performs handover based on the handover policies provided by the network operator. The MM interacts with QS during the call setup and termination.

B. Adaptive Resource Allocation Mechanism

There are already some studies in which the utility function is adopted to QoS [13] [14]. But, these studies do not deal with the detailed adaptation algorithm considered the flow (i.e., consider caller and callee side together) proposed in this paper. Under the assumption that network has knowledge of utility function for all users, we adopt MSS algorithm from caller to callee side together. Fig. 4 shows the proposed a new adaptive resource allocation mechanism and details of the mechanism are as follows.

- AR1: caller AR, AR2: callee AR
- fi: new flow's entry
- fi-AR1-re-flow-list: a list of existing flows in the AR1 to be re-allocated due to fi's joining

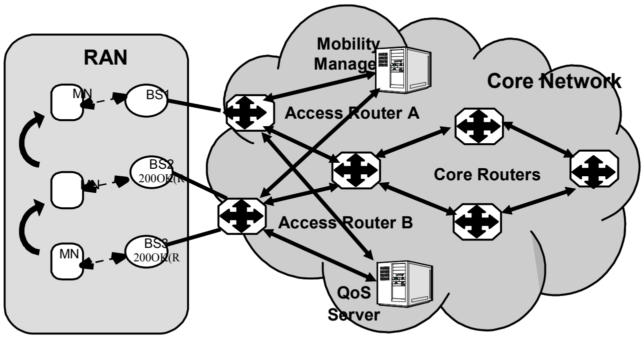


Figure 3. Architectual Model for Mobile Network

- fi-AR2-re-flow-list: a list of existing flows in the AR2 to be re-allocated due to fi's joining
- fi-AR1-alloc: An obtained temporary allocation to fi in AR1 by using the utility function based on MSS algorism
- fi-AR1-re-flow-list-temp-alloc: a list of obtained flows in the AR1 by using the utility function based on MSS algorism
- fi-AR2-alloc: an obtained temporary allocation to fi in AR2 by using the utility function based on MSS algorism.
- fi-AR1-re-flow-list-temp-alloc: a list of obtained flows in the AR2 by using the utility function based on MSS algorism
- fi-mid-alloc: minimum of [fi-AR1-alloc, fi-AR2-alloc]
- bw: bandwidth
- fi-AR1-re-flow-list-last-alloc: the last allocation for the existing flows in AR1, resulting in last-alloc-a, last-alloc-b, last-alloc-c,.. by re-calculating the bandwidth of the existing flows in fi-AR1-re-flow-list if there is remaining un-utilized bandwidth in AR1.
- fi-AR2-re-flow-list-last-alloc: the last allocation for the existing flows in AR2, resulting in last-alloc-x, last-alloc-y, last-alloc-z,.. by re-calculating the bandwidth of the existing flows in fi-AR2-re-flow-list if there is remaining un-utilized bandwidth in AR2.
- fi-AR1-remain-bw: the remained bandwidth in AR1
- fi-AR1-remain-bw: the remained bandwidth in AR2

a) Initial phase(1-4),

QS provides AR's flow list for the caller and callee side.

1). QS receives resource allocation request from caller or caller's AR called AR1.

2). QS generates this new flow's entry called f_i which has the information of its caller's access router (f_i –AR1), and callee's access router (f_i –AR2).

3)&4). From previous step 2, QS generates a list of the existing flows that need to be re-allocated due to f_i's joining. It is called f_i-re-flow-list. This list is made of flows that are passing through the caller and callee ARs and that are affected by f_i joining. For example, f_i-AR1-re-flow-list and f_i-AR1-re-flow-list

b) The second phase(5-7),

Whenever new flow is added, for AR's existed flows for both of caller and callee side, temporary bandwidth for each existing flows and new flow are calculated according to the MSS algorithm. For the new flow, select the minimum value between caller's temporary bandwidth and callee's temporary bandwidth and set it as middle allocation value.

5). By using the utility function based on MSS algorithm, a temporary allocation is made to f_i and the flows in the caller AR list. So, f_i gets f_i-er-alloc (bandwidth/sec), and the flows in f_i-AR1-re-flow-list gets Σf_i*-AR1-re-flow-list-temp-alloc* = [temp-alloc-a + temp-alloc-b + temp-alloc-c + ...]).

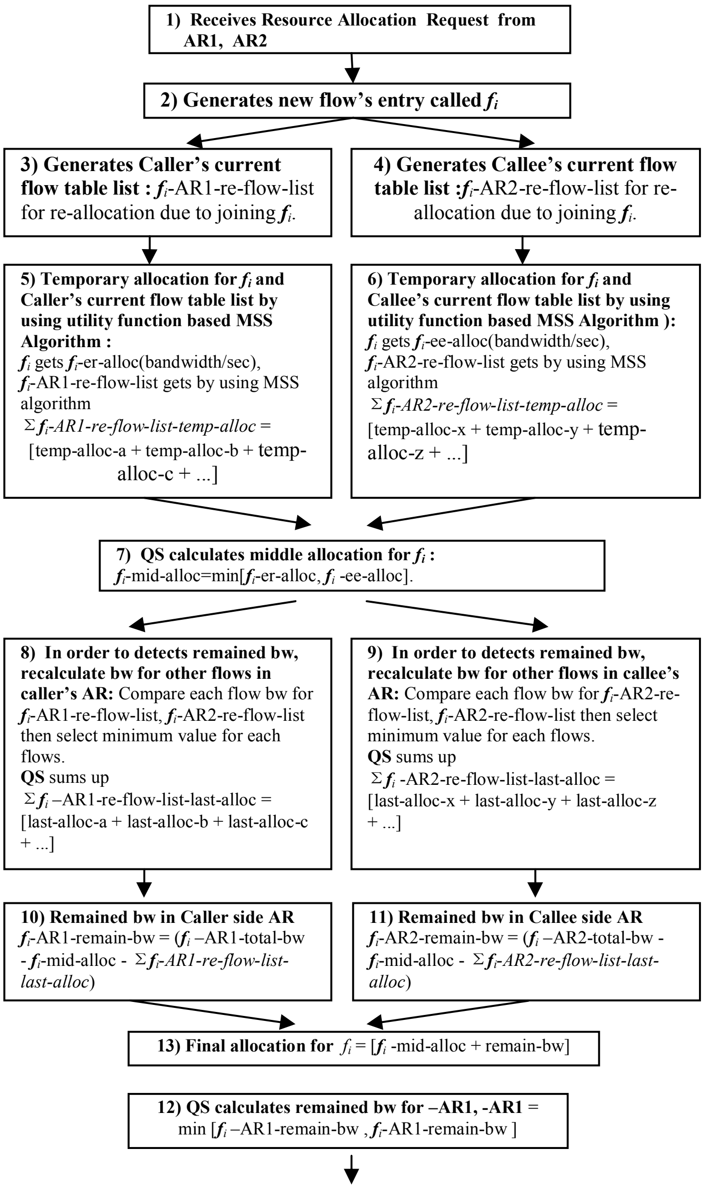

1) Receives Resource Allocation Request from AR1, AR2
2) Generates new flow's entry called f_i
3) Generates Caller's current flow table list : f_i-AR1-re-flow-list for re-allocation due to joining f_i.
4) Generates Callee's current flow table list :f_i-AR2-re-flow-list for re-allocation due to joining f_i.
5) Temporary allocation for f_i and Caller's current flow table list by using utility function based MSS Algorithm : f_i gets f_i-er-alloc(bandwidth/sec), f_i-AR1-re-flow-list gets by using MSS algorithm Σf_i-AR1-re-flow-list-temp-alloc = [temp-alloc-a + temp-alloc-b + temp-alloc-c + ...]
6) Temporary allocation for f_i and Callee's current flow table list by using utility function based MSS Algorithm): f_i gets f_i-ee-alloc(bandwidth/sec), f_i-AR2-re-flow-list gets by using MSS algorithm Σf_i-AR2-re-flow-list-temp-alloc = [temp-alloc-x + temp-alloc-y + temp-alloc-z + ...]
7) QS calculates middle allocation for f_i : f_i-mid-alloc=min[f_i-er-alloc, f_i -ee-alloc].
8) In order to detects remained bw, recalculate bw for other flows in caller's AR: Compare each flow bw for f_i-AR1-re-flow-list, f_i-AR2-re-flow-list then select minimum value for each flows. QS sums up Σf_i –AR1-re-flow-list-last-alloc = [last-alloc-a + last-alloc-b + last-alloc-c + ...]
9) In order to detects remained bw, recalculate bw for other flows in callee's AR: Compare each flow bw for f_i-AR2-re-flow-list, f_i-AR2-re-flow-list then select minimum value for each flows. QS sums up Σf_i -AR2-re-flow-list-last-alloc = [last-alloc-x + last-alloc-y + last-alloc-z + ...]
10) Remained bw in Caller side AR f_i-AR1-remain-bw = (f_i –AR1-total-bw - f_i-mid-alloc - Σf_i-AR1-re-flow-list-last-alloc)
11) Remained bw in Callee side AR f_i-AR2-remain-bw = (f_i –AR2-total-bw - f_i-mid-alloc - Σf_i-AR2-re-flow-list-last-alloc)
13) Final allocation for f_i = [f_i -mid-alloc + remain-bw]
12) QS calculates remained bw for –AR1, -AR1 = min [f_i –AR1-remain-bw , f_i-AR1-remain-bw]

Figure 4. Adaptive Resource Allocation Mechanism

6). Same procedures are done for the flow f_i at the callee AR. Flow f_i gets f_i-ee-alloc (bandwidth/sec), and the flows in the list for callee AR(f_i-AR2-re-flow-list) get Σf_i*-AR2-re-flow-list-temp-alloc* = [temp-alloc-x + temp- alloc-y + temp-alloc-z + ...] .

7). QS calculates middle allocation for f_i, this value is called f_i-mid-alloc which is the minimum of [f_i-er-alloc, f_i -ee-alloc].

c) The third phase(8-11),

In order to detect remained bandwidth for both caller side AR and callee side AR, compare with the temporary value for each existed flow's in caller side AR and one of callee side, select minimum value and set as a last allocation for each existed flows, then sum up all as the last allocation for each caller side AR and callee side AR.

8). QS will detect if there is remaining un-utilized bandwidth in the ARs. This is done by re-calculating the bandwidth of the other flows in f_i-AR1-re-flow-list similarly to the process above. The calculation is done for the corresponding callee-AR of every other flow in the list. And again a minimum is selected for each flow. This is the last allocation for the other flows in the caller AR, resulting in last-alloc-a, last-alloc-b, last-alloc-c. Finally, QS sums up these values [last-alloc-a + last-alloc-b + last-alloc-c + ...], which can be expressed in *Σf_i-AR1-re-flow-list-last-alloc.*

9). In the similar manner, QS re-allocates each flow in f_i-AR2-re-flow-list and results in *Σf_i–AR2-re-flow-list-last-alloc* which is the sum of [last-alloc-x + last-alloc-y + last-alloc-z + ...].

10). Then the remaining bandwidth in caller side AR is f_i-AR1-remain-bw = (f_i –AR1-total-bw - f_i-mid-alloc – *Σf_i-AR1-re-flow-list-last-alloc*).

11). Similarly, the remaining bandwidth in the callee side AR is calculated in the way as above and results in f_i –AR2-remain-bw = (f_i-AR2-total-bw – f_i-mid-alloc – *Σf_i –AR2-re-flow-list-last-alloc*).

d) Final phase(12-13),

Compares the remained bandwidth for caller side AR with the one for callee side AR, then select minimum value and set as a network remained value. Then, allocate the middle allocation bandwidth and remained bandwidth to the new flow.

12). The remaining bandwidth for AR1 and AR2 for the flow f_i is remain-bw = min [f_i –AR1-remain-bw , f_i-AR2-remain-bw].

13). At last, as the result, the final allocation for f_i = [f_i -mid-alloc + remain-bw].

IV. EVALUATION

A. Network Topokogy and Conditions

In order to evaluate the feasibility of our proposal, we have developed the model system by using NS2 simulator with the topology shown in Fig. 5. We have implemented the proposed adaptive resource allocation mechanism. All the entities in the model system are implemented on the NS2. It is assumed that location registration is conducted before call setup. When a MN enters into an area covered by BS, MN sends Registration Request message to its AR. Then the MN's location is recorded in the Location server (i.e.MM). Each AR has 4 units of bandwidth to administrate (1 unit = 40 kbps) and the data flows use RTP packets. In this topology, MN1, 2, 3 and 4 under AR1 have session with MN5, 9, 13, and 17 respectively. Each node request 4 units of bandwidth (1unit = 40 kbps) and each AR has 4 units of bandwidth to administrate (1 unit = 40 kbps). In detail, the bandwidth between AR and CR, AR and MN, are 160kbps (4units). The data flows are using RTP packet and are called Flow1, 2, 3 and 4. Packet size is 250Byte. In the simulation video phone service is assumed as the application.

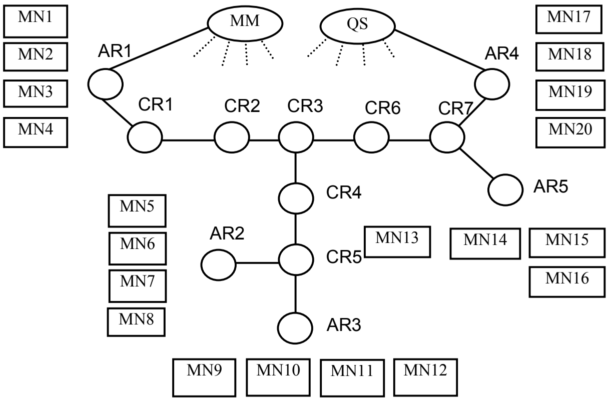


Figure 5. Simulated Network Model System

B. Simulation Results

Fig. 6 shows the bandwidth allocation of each user's flow when a handover takes place. Utility value of Flow B, between MN2 and MN9, Flow S between MN6 and MN10, and Flow H (handover), between MN8 and MN14 are listed in Table I. Flow H starts its call at 30 sec and MN8 starts to move toward to new AR (AR3) at 40 sec. Before Flow H enters to AR3, there are two flows in AR3, our algorithm will therefore allocate 2 units of bandwidth to each flow (Flow H and Flow Sl). Once a new MN having an additional utility function joins to an AR, there is a need to re-allocate the existing resources among the users in the AR according to the utility function that the users have contracted. Hence, when Flow H enters the area of AR3, the allocation becomes Flow B : Flow H : Flow S = 2 : 1 : 1 units. The reason for this allocation is that Flow H has higher utility than Flow H, and therefore it is not affected by Flow H joining the AR. On the other side, Flow S has lower utility than Flow H and therefore one unit is re-allocated to Flow H.

TABLE 1. UTILITY VALUES FOR FLOWS

BW Flow	U1	U2	U3	U4
Flow B	0.7	0.5	0.14	0.1
Flow S	0.45	0.3	0.14	0.1
Flow H	0.5	0.4	0.14	0.1

BW: allocated bandwidth

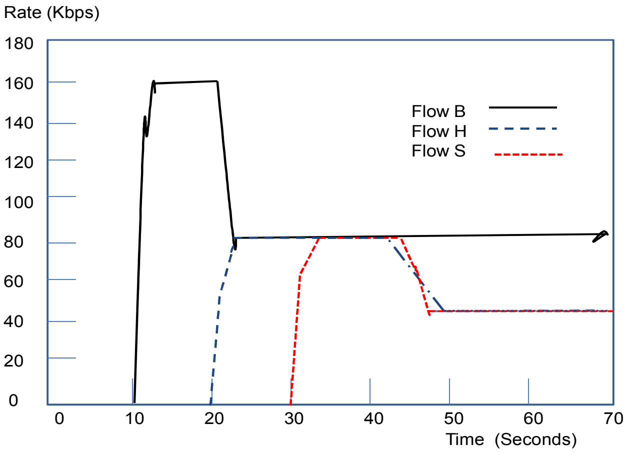


Figure 6. 3 Flows Evaluation Result with Handover

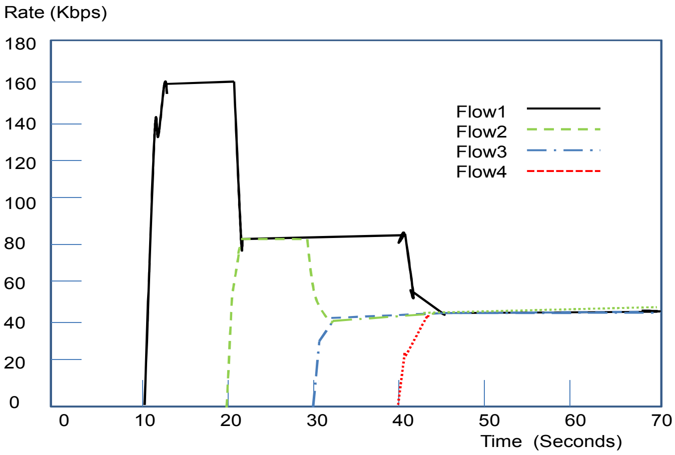


Figure 7. 4 Flows Evaluation Result

Fig. 7 shows another bandwidth allocation result for 4 flows. The utility value of Flow1 from MN1 to MN5, Flow2 from MN2 to MN9, Flow3 from MN3 to MN13, Flow4 from MN4 to MN17 are listed in Table2. For example, for Flow1, the utility values are 0.9 in U1 (0-40kbps), 0.3 in U2 (40-80bkps),0.15 in U3 (80-120kbps), and 0.1 in U4 (120-160kbps) respectively. Flow1 enters the network at time 10s. As it is the first flow in the system, it is allocated the entire AR1's bandwidth. At time 20s, Flow2 enters the network. Then at time 30s, with Flow3's join, the allocation becomes Flow1:Flow2:Flow3 = 2:1:1. Finally, after Flow4 enters, each flow has same bandwidth and this is equal to the allocation that expected. We should consider both the caller side and callee side. However, in this case the caller side is congested and the callee side is not congested at all in each flow so that it is enough that only the caller side is considered.

TABLE 2. UTILITY VALUES FOR FLOWS

BW Flow	U1	U2	U3	U4
Flow1	0.9	0.3	0.15	0.1
Flow2	0.8	0.25	0.14	0.1
Flow3	0.7	0.2	0.2	0.1
Flow4	0.4	0.4	0.1	0.1

Based on Fig. 7, the bandwidth allocation can be explained as follows.

1) At time 10s there is only Flow 1 so that 4 units (160kbps) are allocated to Flow1.
2) At the time 20s, Flow2 enters the network. According to calculation based on utility values in Table 1, 2 units (80kbps) are allocated to Flow1 and Flow2 respectively.
3) At the time 30s, Flow3 enters the network. Then 2 units (80kbps) are allocated to Flow 1, and 1 unit (40kbps) is allocated to Flow 2 and Flow 3 respectively.
4) At the time 40s, Fkow4 enters the network. Then 1 unit(40kbps) are allocated to Flow1, Flow2, Flow3 and Flow4 respectively.

Fig. 8 shows the total bandwidth allocation result of the current method (i.e. static resource allocation method). The situation in Fig.7 is the same as Fig.6. That is, at time 10s, 20s, 30s and 40s, Flow1, Flow2, Flow3 and Flow4 enters the network via AR1 respectively. The allocated bandwidth is fixed (40Kbps) and always the same for all users. So the total bandwidth at AR1 is increased from 40kbps to 160kbps from 10s to 40s. On the other hand, the total bandwidth of the adaptive allocation method is constantly 160kbps.

We have calculated and compared the total user satisfaction in the case of Fig. 7 and Fig. 8. The calculation result is as follows,

1) Proposed adaptive allocation method (Fig. 7)
Flow1:0.9*10+1.2*10+1.2*10+1.45*10=47.5,
Flow2:1.05*10+0.8*20=26.5
Flow3:0.7*20=14, Flow4:0.4*10=4

- Total user satisfaction: 47.5+26.5+14+4=92

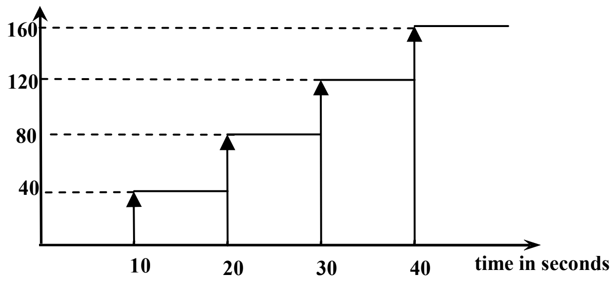


Figure 8. Total Bandwidth Allocation for Current Method

ISBN: 978-1-61208-193-9

2) Current static resource allocation method (Fig. 8)
Flow1: 0.9*40=36, Flow2: 0.8*30=24,
Flow3: 0.7*20=14, Flow4: 0.4*10=4

- Total user satisfaction: 36+24+14+4=78

C. *Considerations*

Through the simulation and calculation we have confirmed the following facts.

1) First, we have confirmed that the allocation in the simulation is equal to the allocation calculated theoretically. Actually, the proposed algorithm has been simulated and it is confirmed that allocation in Fig. 6 and in Fig. 7 are equal to the allocation calculated theoretically.

2) Second, from the user's viewpoint, users can obtain the more satisfied service by this mechanism. Normally when the bandwidth is fully utilized (i.e., congested), the new service request is rejected. But by this mechanism, the new service can be prioritized even in the congested situation case.

3) Third, from the operator's viewpoint, the revenue of operators will be increased by using this mechanism. Actually by this mechanism, it is confirmed that the bandwidth can be utilized at maximum. Moreover, if service price is linked to the value of the utility function, the user will pay more payment so that the operators can obtain more revenue and profit.

V. CONCLUSIONS

In this paper, we have proposed a new adaptive resource allocation mechanism based on the utility function. As the next step, we can expand this mechanism on the following items.

1) In case that some of existing user's utility values are lower, and new users having higher utility value are joining, some of existing user's flow may be suddenly suspended based on our proposed mechanism. This is issue to be solved from the service quality viewpoint. However, even in this case, the existing flow can be continued with some minimum bandwidth if we modify MSS algorithm and the procedure described in Fig, 4.

2) In this paper, it is not clearly mentioned who will assign the utility function for each flow. About this we assume two cases; first case is that the mobile operators can define utility functions for flows and save the information in QS, second case is that mobile users can select a utility function for a flow and send this information to QS by using signaling. resources.

3) We have focused on only bandwidth allocation, but other QoS metrics such as the delay or jitter can be studied in the next step. Also, we have to consider the performance aspect of the proposal, especially how to decrease the call setup delay by applying more efficient and heuristic sorting algorithm because the major portion of MSS algorithm is a sorting process.

REFERENCES

[1] 3GPP TS 29.212. : 3rd Generation Partnership Project; Technical Specification Group Core Network and Terminals ; Policy and Charging Control over Gx reference point(Release 8), 2008-09

[2] Ludwig, R., Eksrom, H., Willars, P., Lundin, N., : AN Evolved 3GPP QoS Concept, VTC Spring 2006, May 2006

[3] Y. Kaneda. : "Policy-based End-to-End QoS Guarantee Using On-Path Signaling for Both QoS Requests and Feedback," ICOIN 2008, I-1, Jan.2008

[4] Jerey K. Mackie-Mason and Hal R. Varian, "Pricing the internet," in Public access to the Internet, Brian Kahin and James Keller, Eds. Prentice Hall, Englewood Cliffs, New Jersey, 1995.

[5] Jerey K. Mackie-Mason and Hal R. Varian, "Pricing congestible network resources," IEEE Journal on Selected Areas in Communications, vol. 13, no. 7, pp. 1141-1149, September 1995.

[6] R. J. Gibbens and F. P. Kelly, "Resource pricing and the evolution of congestion control," Automatica, vol. 35, no. 12, pp. 1969-1985, December 1999.

[7] F. P. Kelly, A. K. Maulloo, and D. K. H. Tan, "Rate control for communication networks: shadow prices, proportional fairness and stability," Journal of Operations Research Society, vol. 49, no. 3, pp. 237-252, March 1998,

[8] James F. Kurose and Rahul Simha, "A microeconomic approach to optimal resource allocation in distributed computer systems," IEEE Transactions on Computers, vol. 38, no. 5, pp. 705-717, May 1989.

[9] N. Liu, J. Bigham, "Utility-maximization bandwidth adaptation for multi-class traffic QoS provisioning in wireless networks",the 1st ACM international workshop on Quality of service and security in wireless and mobile networks, 2005-10.

[10] S. H. Low and D. E. Lapsley, "Optimization flow control, I: Basic algorithm and convergence," IEEE ACM Transactions on Networking, vol. 7, no. 6, pp. 861-875, December 1999.

[11] H. R. Varian. : Microeconomic Analysis, W. W. Norton and Co., 1992.

[12] O. Gonzalez and Michael Needham, "QoS Provisioning Architecture for Next Generation Mobile Networks," IEICE Trans. Commun., Vol.E87-B, No5, May. 2004.

[13] T. Nomura, K. Yamori, E. Takahashi, T. Miyoshi, Y. Tanaka, : "Waiting time versus utility to download images," APSITT2001, session 5, pp.128-132, Nov.2001

[14] K. Yamori, Y. Tanaka. : "Relation between willingness to pay and guaranteed minimum bandwidth in multiple-priority services," APCC2004, no.MA06-1, pp.113-117, Aug.2004

[15] Cao, Z., Zegura, E. "Utility max-min: An application oriented bandwidth allocation scheme," IEEE INFOCOM`99. (1999) 793-801

Performance Modeling of Database Servers in a Telecommunication Service Management System

Maria Kihl[1], Payam Amani[1], Anders Robertsson[2], Gabriela Radu[1], Manfred Dellkrantz[1], and Bertil Aspernäs[3]
[1]Dept. of Electrical and Information Technology, Lund University, Sweden
[2]Dept. of Automatic Control, Lund University, Sweden
[3]Ericsson AB, Karlskrona, Sweden
{maria.kihl, payam.amani, manfred.dellkrantz}@eit.lth.se
luminita.radu@gmail.com, andersro@control.lth.se, bertil.aspernas@ericsson.com

***Abstract*— Resource optimization mechanisms, as admission control and traffic management, require accurate performance models that capture the dynamics of the system during high loads. The main objective of this paper is to develop an accurate performance model for database servers in a telecommunication service management system. We investigate the use of a server model with load dependency. Concurrent requests add load to the system and decrease the server capacity. We derive explicit equations for the state probabilities, the average number of jobs in the system and the average response times. Further, we present some heuristics on how to tune the parameters for given measurement data. Also, using testbed experiments, we validate that the model accurately captures the dynamics of a database server with write-heavy workload.**

***Index Terms*— Performance management; telecommunication systems; queuing theory; database servers.**

I. INTRODUCTION

Resource management of server systems has gained much attention in the last years, since poorly managed resources can severely degrade the performance of a computer system. The experience is that enterprise servers are often the bottlenecks, whereas the network backbone is often underutilized. Therefore, the server systems must provide performance guarantees which satisfy the service-level agreements (on delay, QoS, etc). Also, the system must provide graceful performance degradation during overload.

However, all optimization techniques require accurate performance models of the involved computing systems. The operation region is mainly high traffic load scenarios, which means that the computing systems show non-linear dynamics that needs to be characterized accurately. A software system is basically a network of queues, as examples, the CPU ready queue, semaphore queues, socket queues, and I/O device queues, which store requests in waiting of service in the processors. Therefore, queuing models can be used when describing the dynamic behavior of server systems [1][2][3][4].

In a previous work [1], we have shown that web servers with dynamic content can be modeled as single server queuing systems with processor sharing, where the high load dynamics can be captured with an M/M/1 system. However, this result is only valid for systems with CPU intensive workload. Some recent experiments on databases have shown that the high load dynamics of database servers are completely different for queries involving write operations [5]. Since database servers are important components in Telecommunication service management systems, it is, therefore, important to develop new models for database servers with write-heavy operations.

The concept of load dependent server (LDS) models in which the response time of the jobs in the system is a function of the service time of the jobs and current number of jobs waiting to be serviced has, to the best of our knowledge, firstly been introduced in [6]. Rak *et al.* [7], Curiel *et al.* [8] and Perros *et al.* [6] used standard benchmarks for workload generation and also regression models to capture the system dynamics. A multi-step model parameter calibration strategy was used for fine tuning of the parameters in the model. The resulting models were classified as data driven models. Mathur and Apte [9] presented a queuing network model which represents the load dependent behavior of the LDS. Their model was not analyzed theoretically and was only validated with simulations. A theoretical analysis of the D/G/1 and M/G/1 models with load dependency assumptions was presented in [10] by Leung. These models were developed to be used in congestion control in broadband networks.

In this paper, we add the load dependency behavior to an M/M/m model. The steady state probabilities, average number of jobs in the system and average response times are determined using queuing theory. Also, we perform the same analysis for the case where the queue is limited. The model has a simple structure and can be tuned for various load and database configurations.

Further, in order to tune the parameters of the model to represent the current database and load setup, effects of variations of each parameter on the mean response time of the queries sent to the database as a function of mean effective arrival rate has been studied. Furthermore, some heuristics for tuning the model parameters are introduced. By means of these heuristics, the model parameters can be tuned to match the measurements from the database in a few steps. Finally,

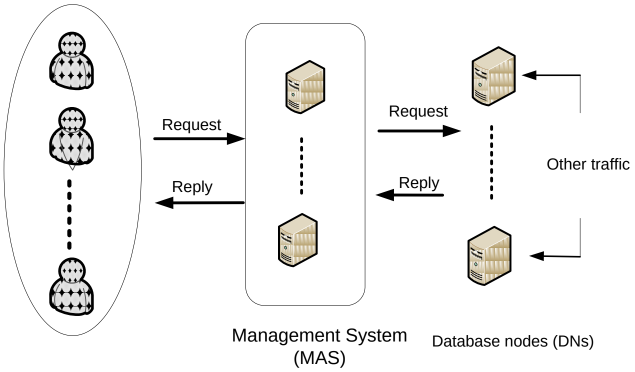


Figure 1. Telecommunication service management systems.

experimental results show that this model is able to capture the high load dynamics of the database server.

This paper is organized as follows. The description of the system is introduced in Section II. Section III is dedicated to introduction of the load-dependent server model. Experiment setups and results are shown in Section IV. Section V concludes the paper.

II. System and Problem description

In telecommunication service oriented architectures, as mobile networks, all services, either user services as telephony, or administrative services as location updates or billing, are handled by a service management system with its own networks and protocols, as illustrated in Figure 1. The service management systems have a complex architecture, usually implemented as large distributed server systems, with Management application servers (MAS) processing service requests from the telecom networks, and databases (DNs) storing subscriber and service data. The DNs are loaded with service traffic from other networks. All signaling is performed across IP networks, with standardized application layer protocols, as Lightweight Directory Access Protocol (LDAP) or Simple Object Access Protocol (SOAP). The system is required to have high reliability for varying traffic loads, where the DNs may be overloaded by the traffic coming from other networks. The nodes can be owned by different network operators, limiting the available information of traffic loads and service progress. All signaling is performed across IP networks.

In this paper, we focus on the modeling aspects of database servers in telecommunication service management systems. The objective is to develop a performance model for the database server that captures the dynamics during high loads. The performance model can be used in resource optimization schemes, as admission control systems, in order to maximize the throughput of the database server, while keeping some latency constraints. One of the challenges for these database servers is that they have a write-heavy workload, which means that the CPU is not the bottleneck during high loads. This means that previous work on performance modeling of server systems is not applicable since they assume CPU-intensive workload.

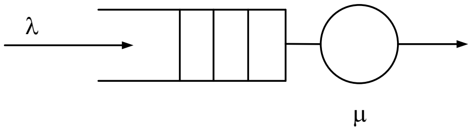


Figure 2. M/M/1 model.

III. Load-dependent server model

Performance models are aimed to be used in the process of designing management entities for server systems. Therefore, the performance model should capture the dominant dynamics of the server system. Most service performance metrics such as response time, service rate and processing delay depend on queue state dynamics.

A. Single server queues

For the objective of performance control, simple models, based on the assumption of a single server queue, are often preferred. The model should only capture the dominating load dynamics of the system, since a well-designed control system can handle many model uncertainties [11]. The classical M/M/1 model, where a single-server queue processes requests that arrive according to a Poisson process with exponential distributed service times, see Figure 2, has been shown to accurately capture the response time dynamics of a web server system [1].

B. M/M/m model with load dependency (M/M/m-LDS)

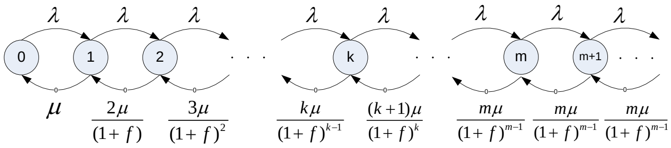
In this paper, we propose to add load dependency to an M/M/m system. In all load-dependent server models, the service time for a request will be dependent on the number of concurrent requests in the system. This load dependency will model effects of the operating system, memory use, etc., which may cause service degradation when there are many concurrent jobs in a computing system [8]. In the experiment section, we will show that the M/M/m-LDS model accurately captures the behavior of write-heavy workload.

The properties of the load dependent M/M/m model (*M/M/m-LDS*) are set by an exponential distributed base processing time $x_{base} = 1/\mu$ and a dependency factor (*f*). When a request enters the system, it gets the base processing time x_{base} assigned to it. A single request in the system will always have a processing time of x_{base}. Each additional request inside the system increases the residual work for all requests inside the system (including itself) by a percentage equal to the dependency factor *f*. When a request leaves the system all other requests have their residual work decreased by *f* percent again. This means that if *n* concurrent requests enter the system at the same point, they will all have a processing time of

$$x_s(n) = x_{base} \cdot (1+f)^{n-1} \qquad (2)$$

A special case is when *f* = 0. It means that there is no load dependency, and all requests will have processing time x_{base}.

The system can process a maximum of *m* concurrent requests at each time instance. Any additional request will have to wait in the queue. New requests arrive according to a Poisson process with average rate λ.



λ
0
1
2
k
m
m+1
. . .
μ
$\frac{2\mu}{(1+f)}$
$\frac{3\mu}{(1+f)^2}$
$\frac{k\mu}{(1+f)^{k-1}}$
$\frac{(k+1)\mu}{(1+f)^k}$
$\frac{m\mu}{(1+f)^{m-1}}$

Figure 3. Illustration of M/M/m-LDS model as a Markov chain.

Therefore, the system can be modeled as a Markov chain as illustrated in Figure 3.

The average service rate of the system that depends on the number of concurrent requests in the system, is derived as followed:

$$\mu_k = \begin{cases} \dfrac{k\mu}{(1+f)^{k-1}} & if \quad 0<k<m \\ \dfrac{m\mu}{(1+f)^{m-1}} & if\ k \ge m \end{cases} \tag{3}$$

By solving the balance equations, stationary probability distribution of existence of k concurrent requests in the system is calculated as below:

$$\pi_k = \begin{cases} \dfrac{\left(\dfrac{\lambda}{\mu}\right)^k}{k!}(1+f)^{\frac{k(k-1)}{2}}\pi_0 & if \quad 0<k<m \\ \dfrac{\left(\dfrac{\lambda}{\mu}\right)^k}{m^{k-m}\cdot m!}(1+f)^{(m-1)(k-\frac{m}{2})}\pi_0 & if\ k \ge m \end{cases} \tag{4}$$

As the sum of the probabilities of all possible states equals to one, π_0 can be derived as follows:

$$\sum_{k=0}^{\infty}\pi_k = 1 \rightarrow$$
$$\pi_0 = \frac{1}{1+\sum_{k=1}^{m-1}\dfrac{\left(\dfrac{\lambda}{\mu}\right)^k}{k!}(1+f)^{\frac{k(k-1)}{2}} + \dfrac{\mu\left(\dfrac{\lambda}{\mu}\right)^m (1+f)^{\frac{m(m-1)}{2}}}{(m-1)!(\mu m-\lambda(1+f)^{m-1})}} \tag{5}$$

The stability condition in this case is

$$\frac{\lambda}{\mu m}(1+f)^{m-1} < 1 \tag{6}$$

The average number of requests in the system, N, can be calculated as below:

$$N = \sum_{k=1}^{\infty} k\cdot\pi_k = N_1 + N_2$$
$$N_1 = \sum_{k=0}^{m-1}\frac{\left(\dfrac{\lambda}{\mu}\right)^k (1+f)^{\frac{k(k-1)}{2}}}{(k-1)!}\pi_0 \tag{7}$$
$$N_2 = \frac{\left(\dfrac{\lambda}{\mu}\right)^m (1+f)^{\frac{m(m-1)}{2}}(\mu m^2-\lambda(m-1)(1+f)^{m-1})\mu}{(m-1)!(m\mu-\lambda(1+f)^{m-1})^2}\pi_0$$

Finally by means of Little's theorem, the average time each request spends in the system, T, can be derived as follows.

$$T = \frac{N}{\lambda} \tag{8}$$

C. M/M/m/n model with load dependency (M/M/m/n-LDS)

In case that the queue is limited to n positions, the probability for an empty system, π_0, can be determined as follows. This queuing system is named as *M/M/m/n-LDS*.

$$\pi_0 = \frac{1}{I+II+III}$$
$$I = 1+\sum_{k=1}^{m-1}\frac{\left(\dfrac{\lambda}{\mu}\right)^k (1+f)^{\frac{1}{2}k(k-1)}}{k!} \tag{9}$$
$$II = \frac{(1+f)^{\frac{1}{2}m^2+\frac{1}{2}m+mn-n-1}\lambda^{n+m+1}}{m^n\mu^{n+m}m!(\lambda(1+f)^{m-1}-\mu m)}$$
$$III = -\frac{(1+f)^{\frac{1}{2}m(m-1)}\lambda^m}{\mu^{m-1}(m-1)!(\lambda(1+f)^{m-1}-\mu m)}$$

Further, the average number of requests in the system is as follows:

$$N = N_1 - N_2$$
$$N_1 = \sum_{k=0}^{m-1}\frac{k\left(\dfrac{\lambda}{\mu}\right)^k (1+f)^{\frac{1}{2}k(k-1)}\cdot\pi_0}{k!}$$
$$N_2 = \frac{\mu(1+f)^{\frac{1}{2}m^2-\frac{1}{2}m-1}}{m^{m-1}\left(-\lambda(1+f)^{m-1}+\mu m\right)}\cdot\frac{N_{2_{n1}}+N_{2_{n2}}-N_{2_{n3}}}{N_{2_{D1}}+N_{2_{D2}}+N_{2_{D3}}}$$
$$N_{2_{n1}} = -\lambda(n+m)(1+f)^{\left(\frac{1}{2}`m^2+\frac{3}{2}m+mn-n-1\right)}\left(\frac{\lambda}{\mu}\right)^{n+m+1}\left(\frac{1}{m}\right)^{n+1}$$
$$N_{2_{n2}} = \left(m(1+f)^m\mu(n+m+1)(1+f)^{\left(\frac{1}{2}`m^2+\frac{1}{2}m+mn-n\right)}\right)\left(\frac{\lambda}{\mu}\right)^{n+m+1}\left(\frac{1}{m}\right)^{n+1}$$
$$N_{2_{n3}} = \left(-\lambda(1+f)^m\mu(m-1)+(1+f)\mu m^2\right)\left(\frac{\lambda}{\mu}\right)^m(1+f)^{\frac{1}{2}m(m-1)}$$
$$N_{2_{D1}} = \left(\frac{1}{m}\right)^m(1+f)^{\frac{1}{2}m(m-1)}m!\left(-\lambda(1+f)^{m-1}+\mu m\right)\left(\sum_{k=1}^{m-1}\frac{\left(\dfrac{\lambda}{\mu}\right)^k(1+f)^{\frac{1}{2}k(k-1)}}{k!}\right) \tag{10}$$
$$N_{2_{D2}} = \left(\frac{-\lambda(1+f)^{m^2+mn-n-1}}{(\mu m)^{n+m}}\right)+\left(\frac{1}{m}\right)^m(1+f)^{\frac{1}{2}m(m-1)}m!\left(-\lambda(1+f)^{m-1}+\mu m\right)$$
$$N_{2_{D3}} = \mu m\left(\frac{\lambda(1+f)^{m-1}}{\mu m}\right)^m$$

Finally, the average response time for a request can be derived using Little's theorem.

D. Parameter tuning

In a telecom system with latency constraints, the dominant dynamic of the system is often characterized by the average response time, T, when varying the average arrival rate, λ. Tuning of the parameters of the load dependent server model in a way that it fits the measured data from the actual server system is a necessary step in modeling of such systems. Assuming that λ and *T* are measureable, and therefore, known, there are three main parameters for the M/M/m-LDS model, *m*, *f* and μ, to tune in order to fit the model on the measured data. Further, for the M/M/m/n-LDS there is an extra parameter, *n*, to tune. Therefore, in Figures 4-8, we have illustrated the effect of changing model parameters. In the rest of the paper, this graph will be called the *λ/T* graph. In each figure, we have assumed that two (three) of the parameters are fixed and the one that is mentioned is the variable. As the equations for calculating the mean response time, is rather complex and the parameters are interdependent, more than one set of parameters can be fit on the measured data. Thus using these figures, we can achieve a heuristic rule for tuning the parameters of the load dependent server model.

In the cases where the M/M/m-LDS model is used, the first parameter to be tuned is the number of servers, *m*. As it can be seen in Figure 4, by increasing the maximum number of concurrent requests that can be processed in the system, the linear part of the λ/T graph will be shorter and the exponential rising rate of the graph is increased. In this case it is assumed that (*f*, μ) = (0.7, 22).

The second parameter to be tuned is the dependency factor, *f*. As shown in Figure 5, by decreasing the dependency factor, the linear part of the λ/T graph is increased, however, the change is slower than in the case where *m* is decreased. On the other hand the exponential rising rate of the graph is increased in comparison with the case where *m* is decreased. Here, it is assumed that (*m*, μ) = (3, 22).

The effects of changing μ on the λ/T graph while fixing the two other parameters is illustrated in Figure 6. As shown in the figure, by increasing μ in equal steps, the λ/T graph will be shifted to the right in equal steps. In this case, the rate of rising of the graph is decreased. In this case, (*m*, *f*) = (3, 0.7).

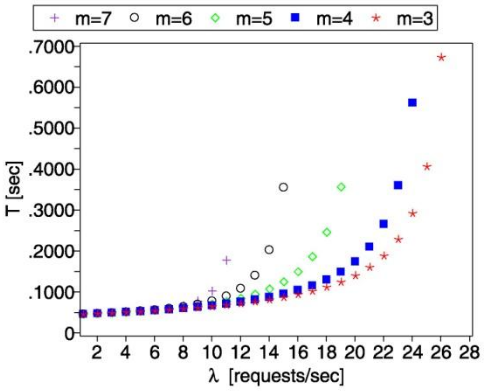


Figure 4. Variations of the λ/T graph for a special scenario with *m* as variable when (*f*,μ) = (0.7, 22).

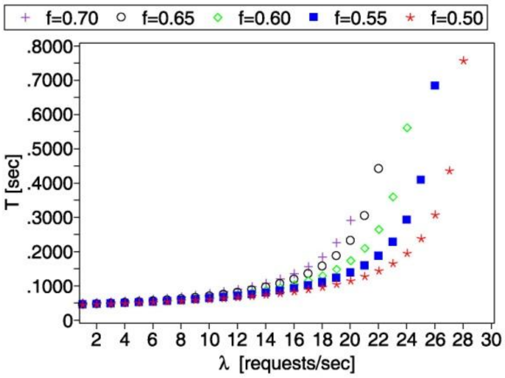


Figure 5. Variations of λ/T graph for a special scenario with *f* as variable when (*m*,μ) = (3, 22).

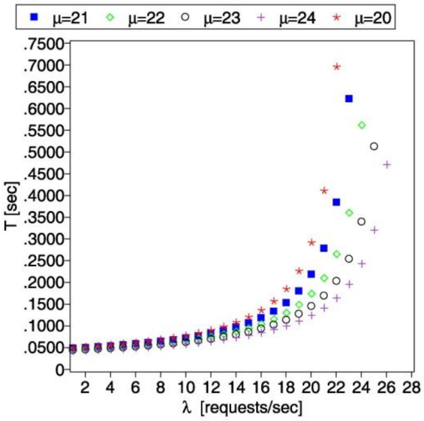


Figure 6. Variations of λ/T graph for a special scenario with μ as variable when (*m*, *f*) = (3, 0.7).

In cases where the M/M/m/n-LDS model is used, there will be a saturation of the response times when the load is high enough to overload the queue. Here it is assumed that the default values are (*m*, *n*, *f*, μ) = (4, 15, 0.6, 22). Figure 7 and Figure 8 show the effects when varying *m* and *f* respectively. In each case the values of the other three parameters are constant. The general effect of changing the parameters is similar as for the case with the infinite queue, with the difference that the response times saturate when the load is high.

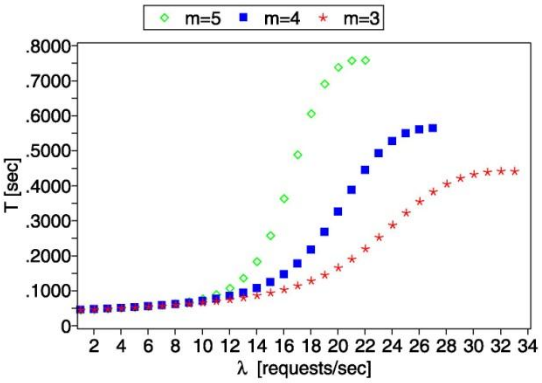


Figure 7. Variations of λ/T graph for a special scenario with *m* as variable when (*n*, *f*, μ) = (15, 0.6, 22).

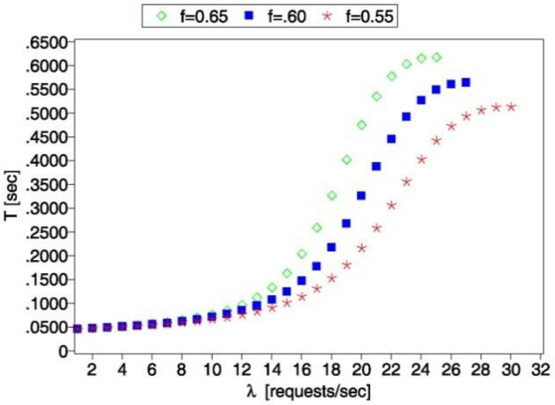


Figure 8. Variations of λ/T graph for a special scenario with *f* as variable when (*m*, *n*, μ) = (4, 15, 22).

IV. Experiments and results

To validate the proposed model, we have performed a series of experiments in our server lab. We developed a database server testbed with a traffic generator and a MySQL 5.1.41 database server, see Figure 9. The computers were connected to a local Fast Ethernet 100 Mbit/s network.

A. Testbed

The traffic generator was implemented in Java, using the JDBC MySQL connector, and it was executed on a computer with an AMD Phenom II X6 1055T Processor at 2.8 GHz and 4 GB main memory. The operating system is Ubuntu 10.04.2 LTS. The traffic generator used 200 working threads and generates MySQL queries according to a Poisson process with average rate λ queries per second. The behavior of the traffic generator was validated in order to guarantee that it was not a bottleneck in the experiments.

The database server has several relations with the same structure but with different number of tuples. The maximum number of allowed concurrent connections is set to 100. The structure of the relations comes from the Scalable Wisconsin Benchmark [12] with 10 million tuples. Two basic types of queries are used, `SELECT` (read) and `UPDATE` (write).

The queries look like this

```
SELECT * FROM <relation> WHERE unique1=?;
UPDATE <relation> SET unique2=? WHERE unique1=?;
```

The question marks are replaced with uniformly distributed random numbers from zero to ten million.

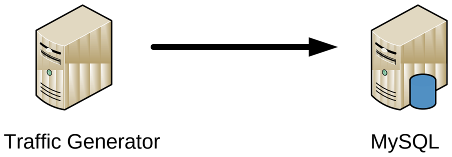


Figure 9. Database server testbed.

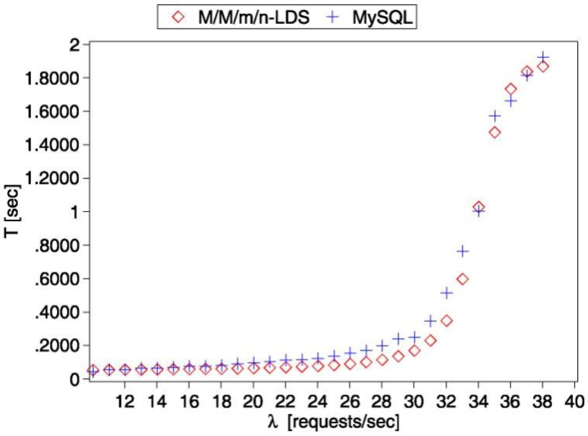


Figure 10. Performance of the M/M/m/n-LDS queuing model in modeling steady state dynamics of a MySQL database server using only update queries.

B. Results

The dynamics of a database server highly depends on the mix of requests, since Select and Update queries require different amount of server capacity. Therefore, we have performed experiments with varying workload mix. Figures 10 and 11 show the results from experiments where the arrival rate is varied from low load to high load. The graphs show the average response times of update queries as a function of the arrival rate. We have fitted M/M/m/n-LDS models for the data using the tuning steps in Section III. In both scenarios, the CPU utilization was very low, also for high loads. The maximum CPU load was about 5%.

In Figure 10, the workload is based on 100% Update queries. The fitted model in this case has the following parameters (*m*, *n*, *f*, μ) = (3, 81, 0.75, 37.1). In order to model the network delays, we have added a bias of 0.023 seconds in the average response times of the proposed model.

Figure 11 depicts the same experiment setup when using a mix of 25% Select and 75% Update queries. The fitted M/M/m/n-LDS model in this case has the following parameters (*m*, *n*, *f*, μ) = (6, 73, 0.44, 35.2). In order to model the network delays, we have added a bias of 0.023 seconds in the average response times of the proposed model.

Figures 10 and 11 verify that the proposed model can represent the average dynamics of a database server with write-heavy workload very well.

V. Conclusions and future works

Resource management schemes require accurate performance models that capture the dominant dynamics of the system in high loads. For server systems with write-heavy workload, load dependent server (LDS) models can be used to model the dynamic overhead effects of concurrent requests. In this paper, queuing theoretic metrics like average number of the requests in the system, average time in the system for each request and the steady state probabilities for M/M/m-LDS models with both unlimited and limited queues have been derived. Further, it has been shown via experiments that the M/M/m/n-LDS model represents the average dynamics of the database server very well. The results are aimed at single database servers, and not aimed at data centers, which have different dynamics. Furthermore, we will use this model in order to design controllers and state estimators for resource management and admission control of database servers.

VI. Acknowledgments

The authors at Lund University are members of the Lund Center for Control of Complex Engineering Systems. The work is partly funded by the Swedish Research Council, grant VR 2010-5864.

References

[1] J. Cao, M. Andersson, C. Nyberg and M. Kihl, “Web Server Performance Modeling using an M/G/1/K*PS Queue”, Proc. of the International Conference on Telecommunication, 2003.

[2] J. Dilley, R. Friedrich, T. Jin, and J. Rolia, “Web server performance measurement and modeling techniques”, *Performance Evaluation*, Vol. 33, No. 1, 1998.

[3] D. A. Menascé and V. A. F. Almeida. *Capacity Planning for Web Services*, Prentice Hall, 2002.

[4] R. D. van der Mei, R. Hariharan, and P. K. Reeser, “Web server performance modeling”, *Telecommunication Systems*, Vol. 16, No. 3, 2001.

[5] M. Kihl, G. Cedersjö, A. Robertsson, B. Aspernäs, “Performance measurements and modeling of database servers”, Sixth International Workshop on Feedback Control Implementation and Design in Computing Systems and Networks, 2011.

[6] H. Perros, Y. Dallery, and G. Pujolle, “Analysis of a queuing network model with class dependent window flow control,” INFOCOM ’92. Eleventh Annual Joint Conference of the IEEE Computer and Communications Societies, IEEE, pp. 968–977 vol.2, May 1992.

[7] A. Rak, A. Sgueglia, "Instantaneous Load Dependent Servers (iLDS) Model for Web Services," In Proceedings of the International Conference on Complex, Intelligent and Software Intensive Systems, 2010.

[8] Curiel, M. and Puigjaner, R., “Using load dependent servers to reduce the complexity of large client-server simulation models”, *Performance Engineering*, Springer-Verlag Berlin Heidelberg, pp. 131-147, 2001.

[9] V. Mathur and V. Apte, “A computational complexity aware model for performance analysis of software servers,” in Proc. of Modeling, Analysis and Simulation of Computer and Telecommunication Systems (MASCOTS), IEEE Computer Society, pp. 537–544, 2004.

[10] Kin K. Leung, “Load-dependent service queues with application to congestion control in broadband networks”, *Performance Evaluation*, Vol. 50, Issue 1, pp. 27-40, October 2002.

[11] K J. Åström and B. Wittenmark, “*Computer–Controlled Systems*”, Upper Saddle River, NJ: Prentice Hall, 1997.

[12] D.J. DeWitt, “The Wisconsin benchmark: Past, present, and future”, The Benchmark Handbook for Database and Transaction Processing Systems, 1, 1991.

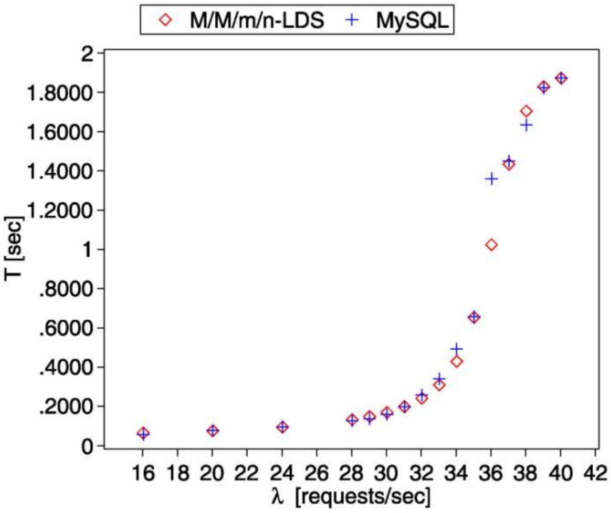


Figure 11. Performance of the M/M/m/n-LDS queuing model in modeling steady state dynamics of a MySQL database server using mixed queries.

Real-Time Telecom Revenue Assurance

Seamless real-time & batch processing of telecom transaction records

Debnarayan Kar, Prateep Misra, Prasun Bhattacharjee, Aniruddha Mukherjee
TCS Innovation Labs - Kolkata
Tata Consultancy Services Ltd.
Kolkata, India
Emails: {debnarayan.kar, prateep.misra, prasun.bhattacharjee, aniruddha.mukherjee}@tcs.com

Abstract— **Telecommunication industry has seen phenomenal growth and intense competition in recent times. Basic connectivity has been commoditized and service providers need to provide personalized and differentiated services to stay competitive and form lasting customer relationship. One of the ways they can do this is by analyzing the data that are generated or pass through their network in real-time, and use the insight to offer personalized services promptly and dynamically. In this scenario it will be immensely beneficial to them to have solutions which allow them a smooth transition path from offline batch processing of stored data to real-time analysis of data as it is available to them. The current paper demonstrates an approach to address the challenge of solution development using Stream Processing, where the same solution can be used to process stored data in offline mode and analyze data stream in real-time for actionable intelligence. IBM stream processing platform Infosphere Streams has been used for implementation of the solution proposed in this work. As a further benefit, if some problem is discovered during real-time analysis of data streams, the same program can be used to analyze the stored data after necessary update is made to the program. This work focuses on seamless dual mode processing of both offline and real-time analysis. The approach has been carried out with reconciliation component of telecommunication revenue assurance and has used CDR (Call Detail Record) for analysis. However, the approach is very generic and it can be applicable in other areas of telecom like churn management, campaign management and in other sectors, like Utilities, Banking etc., which can benefit from real-time and seamless dual mode processing of account transaction records.**

Keywords-revenue assurance; real-time; CDR processing, stream processing

I. INTRODUCTION

There has been phenomenal growth in mobile telecommunication in different parts of the world and the growth momentum is continuing still, more specifically in the emerging markets. As per International Telecommunications Union (ITU) Telecom World 2011 report, globally, there are 5.9 billion mobile subscriptions [1]. This number will rise even further in the future, as mobile penetration in different parts of the world including India, China, and Africa has significant room for increase [2].

Apart from the large number of subscribers, the operators are also facing huge competitive pressure as the high growth prospect in mobile and telecommunication has attracted multiple players to the field.

Communication Service Providers (CSP) are increasingly adding digital content, applications, and other value added service in their offerings to attract and retain customers. If they do not do this, they loose on two counts:

- They do not have differentiators, and face service-churn and subscriber-churn.
- They incur the cost of expensive network equipments in their network, but cannot get the optimum value out of it. They simply become fat-pipe provider and the ‘Over The Top’ (OTT) suppliers profit at their cost by selling content, application to consumers using their network.

The service providers have been offering digital content like movies, music, games, and applications. The present day offerings are limited in supply from most service providers and are expected to grow phenomenally. Even at the current level of service offerings, some service providers have to process multiple millions and for some large operators billions of Call Detail Records (CDR) or Usage Detail Records (UDR) per day.

To remain competitive in the increasingly competitive business environment, the service providers need the capability to source content from multiple partners and dynamically create differentiating offerings using the sourced content and applications. They need the ability to charge for the consumed services in real-time and prevent misuse or unauthorized use. This will increase the number of accounting records required to be processed by multiple order of magnitude. In this scenario, there will be multiple billions of usage accounting records which needs to be processed in real-time to get maximum value from installed network and offered services.

These activities require processing and analyzing high volume of data originating or passing through their network in real-time, near-real-time or in very low latency mode.

Phenomenal subscriber growth and emergence of machine-to-machine (M2M) communication have been leading to very high volume of data stream. To derive the maximum benefit, it is necessary to have computing solution to process high volume live data stream in real-time or near-real-time.

Section II provides an overview of the telecommunication reconciliation process and also the problem, which is addressed by the current work. Section III describes the challenges encountered while addressing the problem. Section IV outlines some of the other methods and mechanisms to analyze high volume of data promptly. Section V provides an overview on stream processing techniques in general and also certain key features of IBM Infosphere Streams. Section VI outlines the approach adopted for the solution and Section VII describes and analyses the results from the current implementation. Section VIII lists some of the ways to enhance the current work and also enumerates some other fields, where the techniques of this work can be utilized. Finally, the concluding section reiterates the usefulness of the current work.

II. Overview And Problem Statement

The reconciliation component of Telecommunication Revenue Assurance system was taken as an example for exploring different approaches to speed up high volume CDR processing. The component reconciles CDR data from Mobile Switching Center (MSC) and Telecommunication Billing System. In a mobile communication network, the MSC coordinates communication channels and necessary processesfor end-to-end connection handling, mobility management, routing charging and real-time account monitoring. The initial work has been limited to the case of reconciling pre-paid outgoing voice call because this use case will benefit most through real-time reconciliation. Service providers may have hybrid customers who use pre-paid and post-paid services on the same account. Hybrid customers have been excluded from this work, because their combined service usage can be reconciled in suitable batch modes.

After the calls are setup between the calling party and called party and as the calls progress, the CDRs are generated in the MSC. The CDR contains several fields including identity of calling party and called party, call start date, call start time, call duration etc. The CDRs contain meta-data about the call or communication service usage which are useful for charging and other management purpose, however they do not contain the actual content of the call or communication.In this current work, since the CDRs are used for their intended purpose of rightful charging, the usage does not violatesubscribers' privacy concern.The current work does not, in any way, prioritize or block subscribers' data content in the delivery process. Thus, it does not raise any “Network Neutrality” (Net Neutrality) issues. The format of the CDRs usually differs among MSC equipment vendors. The formats of theCDRs are also different between MSC and other downstream applications like Billing System and Customer Care System which process them. A Mediation Device receives CDRs from MSC, performs aggregation and correlation among relatedCDRs and reformats them to the format of the downstream application like Billing System. The MSCs, Mediation Device, and Billing Systems are from different vendors, and at times due to error in interface between them, discrepancies crop up in some CDRs as it passes through them. To address the problem of discrepancy, the CDRs need to be reconciled between MSC and the Billing System. Reconciliation is the process of comparing records from multiple sources to verify their accuracy.

When the charge duration between MSC and Billing side for corresponding CDRs differ by more than a threshold, an alert needs to be generated and saved. Corresponding CDRs are matched by their respective values of calling party, called party, call date, call direction, and allowable tolerance threshold in call start time.

As per current deployments in various service provider organizations, the reconciliation is an end-of-day processing where reconciliation needs to be completed in a small window of time during the off-peak hours, usually in the night. This work attempts to suggest an approach which will be able to support both the current deployment scenarios of batch processing as well as the real-time reconciliation scenarios for telecom service providers.

The scenario of the problem is described with the help of Fig. 1. The CDRs from the MSC need pass to a few rounds of filter to:

- Filter in only outgoing call CDRs
- Filter out CDRs which comes from hybrid customer types based on information available from a look-up file.
- Filter out CDRs from post paid customer types

The Billing CDRs need to pass through one simple filter to allow only outgoing call CDRs. The MSC and Billed CDRs needs to be joined where calling party, called party, call date, call direction are same in MSC and Billed CDR and Call Start Time in MSC and Billed CDR are within a configurable tolerance threshold of a few seconds. Finally, alerts needs to be generated from the joined CDRs where the Charge Duration in matched MSC and Billing CDR differ by more than a user configurable threshold.

The purpose of current reconciliation process is to capture and report in real-time, the CDRs between MSC and Billing System, which are under-charged or over-charged beyond a user-configurable tolerance threshold value.

III. Challenges

In case of current work, we are required to support an end-of-day processing with huge volume of data accumulated over the day and also we need to process these CDRs very fast with high throughput as well as real-time reconciliation.

In case of real-time processing, the data are processed as they arrive with the use of suitably designed windows to optimally manage the flow of data streams. In case of end-of-day batch processing, we have two accumulated data sets whose sizes are different.

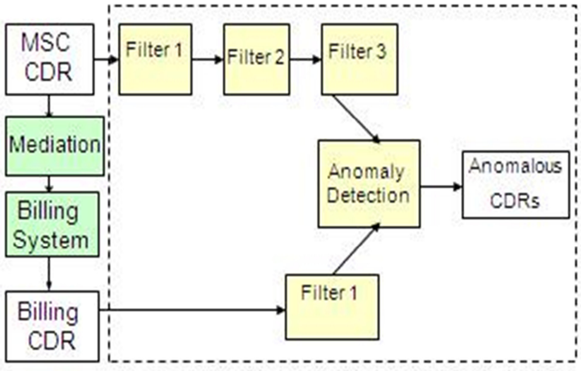


Figure 1. Reconciliation of MSC and Billed CDRs

The data from the MSC have all billable CDRs whereas the CDRs coming from the Billing System have already gone through a few rounds of filtering. Thus, the number of MSC CDRs is more than those of Billed CDRs in any given period of time. MSC CDRs need to pass through three levels of filtering whereas the Billing CDRs need to go through a single simple filter thereby taking a longer time to filter the MSC CDRs. If the application starts the ingestion of MSC and Billing CDR at the same time, and the windows for MSC and Billing CDRs are not properly selected, then it may happen that by the time a particular MSC CDR arrives in its window after

ISBN: 978-1-61208-193-9

passing through multiple cycles of filters, its corresponding Billing CDR has already been removed from the Billing window. When this happen, then the related MSC and Billing CDRs cannot be joined. If there is a deviation in Chargeable Duration between those MSC and Billing CDRs, then the corresponding alert will not be generated.

This necessitates synchronization of the ingestion of the two data sets with careful selection of the windows and achieve high throughput processing without any data loss due to windowing problems.

IV. State of the Art

Very large amount of data can be reduced in space using wave-based approximation [3]. Using this approach, queries can be run on stored data promptly and with reasonable accuracy. Signature based methods have been applied for very fast processing of telecommunication data streams in fraud detection applications [4]. This paper utilizes the fact that fraudsters will have usage patterns which are different from regular users, captures the fraudsters' pattern in the form of statistical signature, and use those for prompt comparison with the CDR stream. This approach is very specifically geared towards fraud management applications in telecom and other industries.

Arroyo solution from Telcordia can efficiently extract, transform and load relevant fields from CDRs stored in data ware house [5]. XML-based transformation rules are used to extract the required CDR fields for faster processing.

There are applications outside the telecom industry which require real-time processing of huge volume of transaction records. In capital market surveillance space, stream processing mechanism has been used for very low latency processing of high throughput transactional data [6].

Graph-theory-based innovations are in exploration to promptly answer queries on huge datasets. Graph-based, disk resident indexing has been devised and utilized for fast queries on high volume RDF data [7]. Complex queries have been answered using sub-graph query matching techniques on a cluster of computers in social network domain which can be applicable generically in other areas also [8].

The works explored in the start of art outlines various methods and mechanisms to analyze high volume of data promptly. Majority of them can provide very high throughput processing on stored data in offline mode. There are also mechanisms for real-time processing of data streams on the fly [9]. Current work addresses the space where the same solution can be used to process high volume of data in real-time as well as in offline batch processing mode.

V. Overview of Stream Processing

Stream processing is a relatively new computing paradigm which is useful for processing and analyzing high volume of data very fast [10]. It can be appreciated by contrasting with database management system (DBMS). In DBMS, the data is stored and is static in nature. Queries are periodically executed over the data to gain insight. In contrast, stream processing has queries residing in the system which are known as continuous queries, and which execute continuously over data streams that are passed through the system. Input data stream can be potentially infinite and stream processing employs the concept of windowing to limit the amount of data records which need to be processed at any particular time. Size of a window can be decided by a number of criteria as specified below [11]:

- **By row count (Count-based window)**: For a count-based window with count value of N, it can accumulate a maximum of N tuples or records. Newly arriving tuples cause older tuples to be evicted if the window is already full.
- **By elapsed time (Time based window)**: For a time-based window with aduration of N seconds, the tuples/records arriving in the last N seconds are accumulated. Tuples which arrived earlier than N second are evicted.
- **By delta of an attribute (Attribute Delta based window)**: In Streams, like in many other system, the data records are called Tuples and the columns or fields in the data records are called Attributes. When the difference in an attribute's value between the earliest tuple in the window and latest tuple is more than the delta threshold, the earliest tuple is removed and the latest tuple is added to the window. If an attribute-delta based window is defined on an attribute named CallStartTime and delta value of 300, then earlier tuples are removed from the window, if their CallStartTime value differs by more than 300 seconds. The attribute value needs to be continuously increasing.

Window can be Tumbling Window or Sliding window. In tumbling window, after the content of the window are processed, its entire content is evicted. In sliding window oldest N records are evicted as newer N records arrive for processing. One tuple can be present in multiple processing steps in sliding window, but in tumbling window one tuple will be present in one processing step only.

The IBM product, Infosphere Streams, has been used for its capability of big data processing and high performance computing [12].It can handle petabytes of data per day and can support traditional and non-traditional data (audio, video etc.). It delivers insights with microsecond latencies and supports the user-defined functions (custom analytics) written in languages like C++ or Java, which makes implementation of complex analytics very easy for developers. Moreover, a single instance of it can support multiple applications.

InfoSphere Streams has a set of built-in stream relational operators which can take care of wide range of requirements.

InfoSphere Streams provides

- A programming model and a language (SPL) for defining data flow graphs consisting of datasources (inputs), operators, and sinks (outputs)
- Controls for fusing operators into processing elements (PEs)
- Infrastructure to support the composition of scalable stream processing applications from these components
- Deployment and operation of these applications across distributed x86 processing nodes, when scaled-up processing is required
- A visualization tool can monitor the running Streams applications distributed across hundreds of servers.

The set of built-in stream relational operators of the SPL can take care of wide range of requirements; some of these are as stated below [11]:

- **Source**: a Source operator is used for creating a stream from data flowing from an external source. This operator is an edge adapter, capable of performing parsing and tuple creation as well as of interacting with *external* devices.

- **Sink**: a Sink operator is used for converting a stream into a flow of tuples that can be used by components that are not part of an InfoSphere Streams instance. This operator is also an edge adapter and its main task consists of converting tuples into corresponding objects, accessible externally through the file system, network, or some other external device.
- **Functor**: a Functor operator is used for performing tuple-level manipulations such as filtering, projection, mapping, attribute creation and transformation.
- **Aggregate**: an Aggregate operator is used for grouping and summarization of incoming tuples. This operator supports a large number of grouping mechanism and summarization functions.
- **Join**: a Join operation is used for correlating two streams. Streams can be paired up in several ways and the join predicate, for example, the expression determining when tuples from the two streams are *joined* can be arbitrarily complex.
- **Sort**: a Sort operator is used for imposing an order on incoming tuples in a stream. The ordering algorithm can be tweaked in several ways.
- **Punctor**: a Punctor operator is used for performing tuple-level manipulations, with the exception of filtering. Unlike a Functor, a Punctor can insert punctuations into the output stream based on a user-supplied punctuation condition.
- **Split**: a Split operator is used for splitting a stream into multiple output streams, based on a split condition that is used to determine which of the output streams a tuple is to be forwarded to.

This platform can be augmented by its simple integrating ability with other visualization products, analytics tools or report-generation tools; through its broad range of stream adapters to consume and publish data from external sources such as network sockets and relational and XML database.

This product is gaining popularity because of its high scalability and robustness in the run-time environment. Additionally it has the ability to add or remove the resources to or from the cluster without impacting the running applications.

VI. Approach To Solution

The application keeps N hours' filtered records in the MSC window, where N can be configured by user and the filtered billing records (or tuples) are compared with the records in the MSC window as they arrive. To ensure that the MSC records are already filtered and are inside the window, a carefully selected small delay is introduced before the ingestion of the billing CDRs.

The window size should be sufficiently large, so that it can hold MSC records of N hours. Count-based window cannot be used as it would require a-priori knowledge of the number of records in N hours' CDR data, which is not available when records are processed in real-time as they arrive. While a time based window can accommodate the requirement real-time processing, it cannot be used in the offline batch processing requirement, because in batch processing mode, the record will arrive at the speed of disk read. So, a one hour window will have all CDRs, which are read in one hour's time rather than all calls those were made in a particular one hour interval.

An attribute-delta based window on the call start time can satisfy both real-time and batch processing requirement. Say for example, for a three hour processing window, records are accumulated in the window as long as the call start time of the arriving records are within N hours (3600xN seconds) of the oldest record's call start time. As this method does not require a-priori knowledge of record number, it can be used in both real-time and batch processing mode.

This means that the filtered billing CDR, as it arrives, is compared with all the available records in the MSC window and once the joining is completed, the billing record is evicted.

As can be seen in Fig. 2, the MSC window is an attribute-delta based sliding window which contains N hours' CDR data. Each Billed CDR arrives in the Billing window and is compared against all the CDRs which are present in the MSC window at that instance of time for the purpose of joining. If the join predicate (joining criteria) is satisfied, the MSC and Billed CDRs are joined. The billing window is a count based window of size 1, and after the comparison is performed, the CDR in the billed window is evicted.

One advantage of this approach is that all the MSC records, which match against a single billing record, are located sequentially in one place in the joined records. This situation is particularly useful to remove anomaly in the joined records which happens if more than one call is made between the same parties within the tolerance threshold of K seconds during joining. The scenario is elaborated through an example bellow.

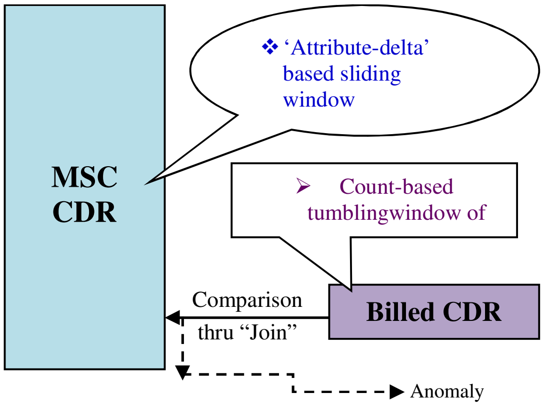

❖ 'Attribute-delta' based sliding window
MSC CDR
➢ Count-based tumblingwindow of
Comparison thru "Join"
Billed CDR
Anomaly

Figure 2. CDR joining with different sized windows

During the joining process, in this example scenario, a tolerance threshold of 60 seconds is allowed on the field Call Start Time. This allowance is made because in many instance the call start time in the billing CDRs is shifted by certain amount from that in the corresponding MSC CDRs.

If same calling and called parties are involved in more than one calls of different duration within the tolerance threshold of 60 seconds, there are multiple matches for same CDRs leading to some false alarms. Table I shows 2 separate calls between the same calling party (A) and called party (B) within the tolerance threshold interval and call start time in the MSC and the corresponding Billed CDRs are shifted by certain amount. The MSC CDRs have among other things, the fields in column

1, 2, 3 and 4 in Table I and the Billed CDRs have the fields in column 1,2, 5 and 6 in Table I

TABLE I. TEST RESULT

Calling Party	Called Party	Call Start Time in MSC	Charge Duration in MSC	Call Start Time in Billing	Charge Duration in Billing
A	B	37030	2	37015	2
A	B	37040	18	37025	18

Table II shows the result after the MSC and Billed CDRs are joined based on calling party, called party and a tolerance threshold of 60 seconds on the “call start time” field.

TABLE II. TEST RESULT (AFTER JOIN)

Calling Party	Called Party	Call Start Time in MSC	Charge Duration in MSC	Call Start Time in Billing	Charge Duration in Billing	Deviation
A	B	37030	2	37015	2	0
A	B	37040	18	37015	2	16
A	B	37030	2	37025	18	16
A	B	37040	18	37025	18	0

In this example, 1st and 4th matches are valid match. The 2nd and 3rd matches are not the intended matches even though they satisfy the joining criteria and subsequently generate alarms which are false alarms. In the subsequent post processing step, the joined matches which cause false alarms are removed from the live data stream.

VII. TEST RESULT & ANALYSIS

The solution was tested on an Intel based X86 HP Server with 12 GB RAM and 2 Quad core CPUs. The CPU in the test environment was equipped with Intel(R) Xeon(R) CPU with 2.00GHz speed and 4MB Cache size. Red Hat Enterprise Linux version 5.4 has been used for this work. IBM Infosphere Streams has been used as stream processing platform. This approach has been tried on 1 hour data set and 3 hour data set from one MSC respectively and tests were repeated 5 times for each case. The time taken for each case is show in Table III.

TABLE III. TEST RESULT (AFTER 5 TIMES REPETITION)

Date Set	Times Taken in Seconds
10,61,421 CDRs in MSC side	Median : 33 , Max : 34, Min : 33
3,53,336 CDRs in MSC side	Median : 13, Max : 13, Min : 13

The tests were repeated 5 times on two different datasets of different durations. As can be seen, the solution can process more than 30,000 CDRs per second. Anomaly or false alerts due to multiple calls by same parties in the tolerance threshold interval are taken care of through some post processing as detailed bellow

The post processing after join cannot be performed by mere use of the Infosphere Streams built-in operator (operator called Functor) for following reasons:

- The Functor operator allows history access or past record access only through numeric literals. Variables cannot be used for history access.
- State variables cannot be used in all part of Functor operator.

To overcome the problem of performing the required post processing using built-in operators, a user-defined operator (UDOP) was written using C++ interface of Stream Processing Language.

For the scenario described above, all the MSC CDRs that are joined against the same billing CDR are present sequentially. For all the joined records in the same sequence the calling party number, called party number, and the Billing CDR call start time are the same. One joined record in this sequence is a valid join and the other invalid joined record(s) in the same sequence needs to be discarded.

For multiple joined records with same calling party number, called party number, call start time, the post-processing component selects the one where the difference in call start timebetween MSC and Billing CDR is the minimum. However if the MSC CDRs has already been joined with another Billing CDR, then that that MSC CDR is not considered by post-processing module. As can be seen in Table III, it takes 13 seconds to process 1 hour data and 33 seconds to process 3 hours’ data. This happens because sufficient number of MSC CDRs are filtered and made available before ingestion of Billing CDRs in the batch processing mode. As the CDR volume grows, the impact of this delayed injection of billed CDRs is significantly minimized in the case of offline processing mode.

In case of real-time processing, the MSC window does not need to hold N hours’ CDR data. Assuming there is a processing delay of maximum of t seconds between MSC CDRs and Billed CDRs, then MSC window of $\mathbf{t + t_2 + \delta}$ seconds will be sufficient for real-time processing of CDRs where δ is the maximum time required for processing t seconds’ CDRs in the stream processing platform and t_2 is if any optional tolerance threshold is required between MSC CDR and corresponding Billed CDR. The Billed CDR window will stay as it is i.e. a count-based window of size 1.

VIII. FURTHER WORK

The anomaly detection logic can be broken up and program flow can be updated to fix different parts of the logic to run on different cores of the host server. The approach is called **core-pinning**. Subsequently the performance needs to be compared with and without core-pinning.

For even more prompt processing, multiple host nodes can be placed in a cluster and thus the processing can be distributed across multiple hosts. Infosphere Streams allows seamless addition and removal of host in the computing cluster. The processing can be distributed across the host in different ways [13]:

- Split the stream of CDR data into different sub-streams and process different sub-streams on different host in parallel.
- Divide the application logic into components, which can be processed sequentially in a pipe-line fashion, and perform different component on different host.

The scenario can be extended to include other type of services in addition to pre-paid voice call. Other type of pre-paid service, services used by hybrid subscribers and even the post-paid services can be brought under scope of near-real-time reconciliation. With multiple data streams for different services running on InfosphereStreams computing cluster, it will be useful to segment and position the data streams judiciously on the hosts of the cluster to achieve optimum load balancing.

The current work has been tried on a limited set of CDR data. This work and its proposed extension can be run on various CDR data streams of different durations and also on archived and real-time CDR streams. The process of running the application on different data sets and scenarios will provide more insight into the solution domain.

In the deployment scenario where a huge volumes of CDRs needs to be reconciled in a relatively small time window, Hadoop Map-Reduce based approach can be considered [14]. Apache Hadoop is open-source and it can be implemented on large clusters of multiple commodity servers. It scales linearly to handle huge data by adding more nodes to the cluster transparently. This approach is useful for scenario which require very high throughput, but does not need very low latency and is suitable for batch processing operation, whereas the Infosphere Streams base clustering is useful for real-time on the fly stream processing.

The discussion so far in this section has focused on how the current work can be enhanced or extended. However the techniques of this work, i.e., stream-based real-time analytic processing, can be applied in other domains also. Low latency processing of CDRs or other kind of records can be re-used in various other kinds of applications like capacity management, traffic analysis, user trending, Quality of Experience (QoE) metric collection etc some of which are outline below:

- Churn Management - Monitor and correlate key attributes from different sources like Customer Relationship Management (CRM), Tariff Plan, Provisioning, Mediation, Billing, Network Switch to predict probability of churn before it occurs
- Campaign Management - Real-time analytics to deliver right message to right customer segment/prospects at right time and in right place
- Network Traffic Monitoring - Real-time correlation of traffic data with other sources like CRM, historical Usage Summary to set differentiating priority for traffic and optimize network usage for greater profitability.

IX. CONCLUSION

Stream processing platforms are ideally suited for real-time or near real-time processing. They can also be used for complex event processing where one or more real-time data streams are correlated with historical archived data and look-up information to arrive at actionable intelligence on the fly. However with careful consideration, applications can be developed using stream processing which can be used in both real-time and offline batch processing mode, just by providing different input parameter and without requiring any program logic modification. Applications like revenue assurance are currently done by many operators in offline batch processing mode, however they can greatly benefit by being done in real-time. This dual mode application can provide a smooth transition path from batch processing to real-time processing mode. It can validate the processing logic with off line data before deploying it for online real-time analysis.

During real-time analysis, if any problem is found in the processing logic which is found out due to some kind of new or unforeseen data in the field, then the application can be modified to correct the processing logic, and the same application can run in offline mode on the archived data.

ACKNOWLEDGMENT

The authors thank Mr Rajat Garg and his team from Tata Consultancy Services Ltd (TCS) telecommunication business unit for their active support and valuable input during this work.

REFERENCES

[1] The World in 2011 ICT Facts and Figures, ITU Telecom World 2011, http://www.itu.int/ITU-D/ict/facts/2011/material/ICTFactsFigures2011.pdf, <retrieved: Feb. 2012>

[2] Vodafoneannual report 2010 : http://www.vodafone.com/content/dam/vodafone/investors/annual_reports/annual_report_accounts_2010.pdf (Mobile penetration : pp. 6-6), <retrieved: Feb. 2012>

[3] Anna C. Gilbert, Yannis Kotidis, S. Muthukrishnan, and Martin J. Strauss, "Surfing Wavelets on Streams: One-Pass Summaries for Approximate Aggregate Queries", AT&T Labs-Research

[4] Corinna Cortes and Daryl Pregibon,"Signature-Based Methods for Data Streams", AT&T Shannon Labs, Florham Park, New Jersey, USA

[5] Munir Cochinwala and Euthimios Panagos, "Near Real-time Call Detail Record ETL Flows", Telcordia Applied Research.

[6] Aniruddha Mukherjee, Punit Diwan, Prasun Bhattacharjee, Debnath Mukherjee, and Prateep Misra, "Capital Market Surveillance using Stream Processing", Tata Consultancy Services Ltd.

[7] Matthias Bröcheler, V.S. Subrahmanian, and Andrea Pugliese,"DOGMA: A Disk-Oriented Graph Matching Algorithm for RDF Databases", University of Maryland, USA and Universit`a della Calabria, Italy

[8] Matthias Bröcheler, V.S. Subrahmanian, and Andrea Pugliese,"COSI: Cloud Oriented Subgraph Identification in Massive Social Networks", University of Maryland, USA and Universit`a della Calabria, Italy

[9] Miran Dylan, "An Analysis of Stream Processing Languages", Department of Computing, Macquarie University, Sydney, Australia

[10] http://www.smartercomputingblog.com/2011/06/10/ibm-infosphere-streams/ , <retrieved: Feb. 2012>

[11] "IBM InfoSphere Streams: Programming Model and Language Reference", Version 1.2.1.

[12] Roger Rea and Krishna Mamidipaka,"IBM InfoSphere Streams : Enabling complex analytics with ultra-low latencies on data in motion", IBM Software Group

[13] Chuck Ballard, Daniel M Farrell, Mark Lee, Paul D Stone, Scott Thibault, and Sandra Tucker, "IBM InfoSphere Streams Harnessing Data in Motion", First Edition, http://www.redbooks.ibm.com/abstracts/sg247865.html, <retrieved: Feb. 2012>

[14] Tom White, "Hadoop: The Definitive Guide", O'Reilly Media, Inc, June 2009: First Edition

ISBN: 978-1-61208-193-9