MOBILITY 2011

The First International Conference on Mobile Services, Resources, and Users

ISBN: 978-1-61208-164-9

October 23-29, 2011

Barcelona, Spain

MOBILITY 2011 Editors

Filipe Cabral Pinto, Telecom Inovação S.A., Portugal

In-Young Ko, Korea Advanced Institute of Science and Technology (KAIST), Korea

Josef Noll, University of Oslo & Movation, Norway

MOBILITY 2011

Foreword

The First International Conference on Mobile Services, Resources, and Users [MOBILITY 2011], held between October 23 and 29, 2011 in Barcelona, Spain, started a series of events dedicated to mobility-at-large, dealing with challenges raised by mobile services and applications considering user, device and service mobility. We welcomed technical papers presenting research and practical results, position papers addressing the pros and cons of specific proposals, such as those being discussed in the standard fora or in industry consortia, survey papers addressing the key problems and solutions on any of topics, short papers on work in progress, workshops and panel proposals.

Users increasingly rely on devices in different mobile scenarios and situations. "Everything is mobile", and mobility is now ubiquitous. Services are supported in mobile environments, through smart devices and enabling software. While there are well known mobile services, the extension to mobile communities and on-demand mobility requires appropriate mobile radios, middleware and interfacing. Mobility management becomes more complex, but is essential for every business. Mobile wireless communications, including vehicular technologies bring new requirements for ad hoc networking, topology control and interface standardization.

We take here the opportunity to warmly thank all the members of the MOBILITY 2011 Technical Program Committee, as well as the numerous reviewers. The creation of such a broad and high quality conference program would not have been possible without their involvement. We also kindly thank all the authors who dedicated much of their time and efforts to contribute to MOBILITY 2011. 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 MOBILITY 2011 organizing committee for their help in handling the logistics and for their work to make this professional meeting a success.

We hope that MOBILITY 2011 was a successful international forum for the exchange of ideas and results between academia and industry and for the promotion of progress in the areas of mobile services, resources and users.

We are convinced that the participants found the event useful and communications very open. We also hope the attendees enjoyed the historic charm of Barcelona, Spain.

MOBILITY 2011 Chairs:

General Chair
Josef Noll, University of Oslo & Movation, Norway

Advisory Committee
Petre Dini, Concordia University, Canada & IARIA, USA
Pekka Jäppinen. Lappeenranta University of Technology, Finland
Abdulrahman Yarali, Murray State University, USA

Industry Liaison Chairs
Filipe Cabral Pinto, Telecom Inovação S.A., Portugal
Xiang Song, Microsoft, USA
Xun Luo, Qualcomm Inc. - San Diego, USA

Special Area Chairs on Video
Mikko Uitto, VTT Technical Research Centre of Finland, Finland
Sandro Moiron, University of Essex, UK

Special Area Chairs on Mobile Wireless Networks
Mohammad Mushfiqur Chowdhury, University of Oslo, Norway
Masashi Sugano, Osaka Prefecture University, Japan

Special Area Chairs on Mobile Web / Application
In-Young Ko, Korea Advanced Institute of Science and Technology (KAIST), Korea

Special Area Chairs on Context-aware, Media, and Pervasive
Brent Lagesse, Oak Ridge National Laboratory, USA

Special Area Chairs on Mobile Internet of Things and Mobile Collaborations
Jörn Franke, SAP Research Center - Sophia Antipolis, France
Nils Olav Skeie, University College Telemark, Norway

Special Area Chairs on Vehicular Mobility
Gianluca Franchino, CEIICP - Scuola Superiore Sant'Anna - Pisa, Italy

Special Area Chairs on Mobile Cloud Computing
Chunming Rong, University of Stavanger, Norway
Josef Noll, Center for Wireless Innovation, Norway

Publicity Chairs
Aline Carneiro Viana, INRIA Saclay - Ile de France - Orsay, France
Sarfraz Alam, UNIK-University Graduate Center, Norway

MOBILITY 2011

Committee

MOBILITY General Chair

Josef Noll, University of Oslo & Movation, Norway

MOBILITY Advisory Committee

Petre Dini, Concordia University, Canada & IARIA, USA
Pekka Jäppinen. Lappeenranta University of Technology, Finland
Abdulrahman Yarali, Murray State University, USA

MOBILITY Industry Liaison Chairs

Filipe Cabral Pinto, Telecom Inovação S.A., Portugal
Xiang Song, Microsoft, USA
Xun Luo, Qualcomm Inc. - San Diego, USA

MOBILITY Special Area Chairs on Video

Mikko Uitto, VTT Technical Research Centre of Finland, Finland
Sandro Moiron, University of Essex, UK

MOBILITY Special Area Chairs on Mobile Wireless Networks

Mohammad Mushfiqur Chowdhury, University of Oslo, Norway
Masashi Sugano, Osaka Prefecture University, Japan

MOBILITY Special Area Chairs on Mobile Web / Application

In-Young Ko, Korea Advanced Institute of Science and Technology (KAIST), Korea

MOBILITY Special Area Chairs on Context-aware, Media, and Pervasive

Brent Lagesse, Oak Ridge National Laboratory, USA

MOBILITY Special Area Chairs on Mobile Internet of Things and Mobile Collaborations

Jörn Franke, SAP Research Center - Sophia Antipolis, France
Nils Olav Skeie, University College Telemark, Norway

MOBILITY Special Area Chairs on Vehicular Mobility

Gianluca Franchino, CEIICP - Scuola Superiore Sant'Anna - Pisa, Italy

MOBILITY Special Area Chairs on Mobile Cloud Computing

Chunming Rong, University of Stavanger, Norway
Josef Noll, Center for Wireless Innovation, Norway

MOBILITY Publicity Chairs

Aline Carneiro Viana, INRIA Saclay - Ile de France - Orsay, France
Sarfraz Alam, UNIK-University Graduate Center, Norway

MOBILITY 2011 Technical Program Committee

Ramon Aguero Calvo, University of Cantabria, Italy
Nazim Agoulmine, University of Evry, France
Ioannis Anagnostopoulos, University of Central Greece - Lamia, Greece
Stefan Arbanowski, Fraunhofer Fokus, Germany
Faouzi Bader, Centre Tecnologic de Telecomunicacions de Catalunya (CTTC), Spain
Atta Badii, University of Reading, UK
Payam M. Barnaghi, University of Surrey, United Kingdom
Sonia Buchegger, KTH - Stockholm, Sweden
Enrico Buracchini, Telecom Italia, Italy
Filipe Cabral Pinto, Telecom Inovação S.A., Portugal
Jian-Nong Cao, Hong Kong Polytechnic University, Hong Kong
Aline Carneiro Viana, INRIA Saclay - Ile de France - Orsay, France
Mohammad Mushfiqur Chowdhury, University of Oslo, Norway
Luis M. Correia, IST - Technical University of Lisbon, Portugal
Klaus David, University of Kassel, Germany
Kevin Deng, Jilin University, China
Marco Fiore, INSA Lyon, INRIA, France
Scott Fowler, Linköping University, Sweden
Gianluca Franchino, CEIICP - Scuola Superiore Sant'Anna - Pisa, Italy
Joern Franke, SAP Research Center - Sophia Antipolis, France
Rosario Garrido Cantos, University of Castilla-La-Mancha - Albacete, Spain
Richard Gunstone, Bournemouth University, UK
Kamrul Hassan, Bangladesh University of Engineering and Technology, Bangladesh
Poul Heegaard, NTNU, Norway
Peizhao Hu, NICTA, Australia
Jiun-Long Huang, National Chiao Tung University, Taiwan
Pekka Jäppinen. Lappeenranta University of Technology, Finland
Eduardo Juarez, Universidad Politecnica de Madrid (UPM), Spain
Vana Kalogeraki, Athens University of Economics and Business, Greece
Vasileios Karyotis, National Technical University of Athens (NTUA), Greece
In-Young Ko, Korea Advanced Institute of Science and Technology (KAIST), Korea
Brent Lagesse, Oak Ridge National Laboratory, USA
Juhani Latvakoski, VTT, Finland
Frank Yong Li, University Agder, Norway
Seng Loke, La Trobe University, Australia
Xun Luo, Qualcomm Inc. - San Diego, USA
Satya Prasad Majumder, Bangladesh University of Engineering and Technology, Bangladesh
Behrouz Maham, UNIK-University Graduate Center, Norway

Copyright Information

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

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

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

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

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

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

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

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

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

Table of Contents

OghmaSip: Peer-to-Peer Multimedia for Mobile Devices

Raimund K. Ege
Department of Computer Science
Northern Illinois University
DeKalb, IL 60115, USA
ege@niu.edu

***Abstract*—Mobile devices are rapidly being accepted as primary vehicle to consume multimedia content. Capable smart phones with high-speed next-generation Internet connectivity are becoming common place. Peer-to-peer content delivery is one way to ensure that sufficient data volume can be efficiently delivered. However, the openness of delivery demands adaptive and robust management of intellectual property rights. In this paper we describe a framework and its implementation to address the central issues in content delivery: a scalable peer-to-peer-based content delivery model, paired with a secure access control model that enables data providers to reap a return from making their original content available. We describe our prototype implementation for the Android platform that uses the session initiation protocol (SIP) for peer communication.**

Keywords-multimedia sharing; peer-to-peer content delivery; session initiation protocol

I. INTRODUCTION

High bandwidth Internet connectivity is no longer limited to reaching PCs and laptops: a new generation of devices, such as netbooks and smart phones, is within reach of 3G/4G telecommunication networks. Smart phones have ushered in a new era in omni-present broadband media consumption. Services such as iTunes, YouTube, and FaceBook are popularizing delivery of audio and video content to anybody with a broadband Internet connection.

In this paper, we describe a framework for multimedia content delivery that is based on peer-to-peer file sharing. Peers communicate with messages according to the session initiation protocol to discover each other and exchange data. We describe the implementation of a video player application for the Android platform that delivers video in a secure and managed way.

Delivering multimedia services has many challenges; the ever increasing size of the data requires elaborate delivery networks to handle peek network traffic. Another challenge is to secure and protect the property rights of the media owners. A common approach to large-scale distribution is a peer-to-peer model, where clients that download data immediately become intermediates in a delivery chain to further clients. The dynamism of peer-to-peer communities means that principals who offer services will meet requests from unrelated or unknown peers. Peers need to collaborate and obtain services within an environment that is unfamiliar or even hostile.

Therefore, peers have to manage the risks involved in the collaboration when prior experience and knowledge about each other are incomplete. One way to address this uncertainty is to develop and establish trust among peers. Trust can be built by either a trusted third party [2], or by community-based feedback from past experiences [3] in a self-regulating system. Other approaches reported in the literature use different access control models [4] [5] that qualify and determine authorization based on permissions defined for peers. In such a complex and collaborative world, a peer can benefit and protect itself only if it can respond to new peers and enforce access control by assigning proper privileges to new peers.

The broader goal of our work is to address the trust in peers which are allowed to participate in the content delivery process, to minimize the risk and to maximize the reward garnered from releasing data in to the network. In our prior work [9] [15], we focused on modeling the nature of risk and reward when releasing content to the Internet. We integrated trust evaluation for usage control with an analysis of risk and reward. Underlying our framework is a formal computational model of trust and access control. In the work reported here, we focus on the implementation aspects of the framework, especially the use of the Session Initiation Protocol (SIP).

Our paper is organized as follows: the next section will elaborate on how the data provider and its peers can quantify gain from participating in the content delivery. It also explains our risk/reward model that enables a data source to initially decide on whether to share the content and keep some leverage after its release. Section III describes our prototype architecture that uses the session initiation protocol to establish a community of peers to share content. No central tracker manages a database of peer and trust

 ISBN: 978-1-61208-164-9

information, but rather peers maintain a distributed database. Peers can serve both as source and as consumer of data. Section IV introduces our prototype client for the Android platform and its implementation in Java. Data is exchanged using the Stream Control Transmission Protocol (SCTP) and is secured using a PKI-style exchange of public keys and data encryption. The paper concludes with our assessment of how peer-to-peer systems can shed their freewheeling image via sensible access control additions.

II. QUALIFYING THE VALUE OF MULTIMEDIA

It is amazing at what rate multimedia data is introduced to the Internet and consumed. Almost any kind of multimedia data has value to somebody. Releasing it to the Internet carries potential for reaping some of the value, but also carries the risk that the data will be consumed without rewarding the original source. In addition to the cost of creating the original multimedia data, there is also a cost associated with releasing the data, i.e., storage and transmission cost.

For example, consider the life of a typical "viral" video found on a popular social media site: the video is captured via a smartphone camera (maybe even accidentally), then is uploaded to the social media site, discussed (i.e., "liked" and "friended"), and viewed by a large audience (measured in millions of hits). The video taker is rewarded with fame, rarely gets a monetary reward, the entity that is getting rewarded is the social media site, which will accompany the video presentation with paid advertising.

Let us first recap our model (described also in [1]) to asses risk and reward, by quantizing aspects of the information interchange between the original source, the transmitting medium and the final consumer of the data. Our emphasis here is on the reward quantity, rather than on how trust in peers affects the outcome.

In a traditional fee for service model the reward *"R"* to the source is the fee *"F"* paid by the consumer minus the cost *"D"* of delivery:

$$R = F - D$$

The cost of delivery *"D"* consist of the storage cost at the server, and the cost of feeding it into the Internet. In the case of a social media site, considerable cost is incurred for providing the necessary server network and their bandwidth to the Internet. The social media site recovers that cost by adding paid advertising on the source web page as well as adding paid advertising onto the video stream. The site's business model recognizes that these paid advertisings represent significant added value. As soon as we recognize that the value gained is not an insignificant amount, the focus of the formula shifts from providing value to the original data source to the reward that can be gained by the transmitter. If we quantify the advertising reward as *"A"* the formula now becomes:

$$R = F - (D - A)$$

Even in this simplest form, we recognize that *"A"* has the potential to outweigh *"D"* and therefore reduce the need for *"F"*. As the social media site recognizes, the reward lies in *"A"*, i.e., paid ads that accompany the video.

Mediation frameworks can capture the mutative nature of data delivery on the Internet (see also our prior work [8]). As data travels from a source to a client on a lengthy path, each node in the path may act as mediator. A mediator transforms data from an input perspective to an output perspective. In the simplest scenario, the data that is fed into the delivery network by the source and is received by the ultimate client unchanged: i.e., each mediator just passes its input data along as output data. However, that is not the necessary scenario anymore: the great variety of client devices already necessitate that the data is transformed to enhance the client's viewing experience. We apply this mediation approach to each peer on the path from source to client. Each peer may serve as a mediator that transforms the content stream in some fashion. Our implementation employs the stream control transmission protocol (SCTP) which allows multi-media to be delivered in multiple concurrent streams. All a peer needs to do is add an additional stream for a video overlay message to the content as it passes through.

The formula for reward can now be extended into the P2P content delivery domain, where a large number of peers serve as the transmission/storage medium. Assuming "*n*" number of peers that participate and potentially add value the formula for the reward per peer is now:

$$R_p = \sum_{i=1}^{n} (F_i - (D_i - A_i)) - F_p$$

D_i and A_i are now the delivery cost and value incurred at each peer that participates in the P2P content delivery. F_i is the fee potentially paid by each peer. F_p is the fee paid to the data source provider. Whether or not the data originator will gain any reward depends on whether the client and the peers are willing to share their gain from the added value. In a scenario where clients and peers are authenticated and the release of the data is predicated by a contractual agreement, the source will reap the complete benefit.

In our model, we quantify the certainty of whether the client and peers will remit their gain to the source with a value of trust. Trust is evaluated based on both actual observations and recommendations from referees. Observations are based on previous

 ISBN: 978-1-61208-164-9

interactions with the peer. Recommendations may include signed trust-assertions from other principals, or a list of referees that can be contacted for recommendations. Our model enables an informed decision on whether to accept a new peer based on the potential additional reward gained correlated to the risk/trust encumbered by the new peer.

III. PEER-TO-PEER ARCHITECTURE

Our prototype of peer-to-peer multimedia delivery aims to deliver multi-media content from a source to a large number of clients. We assume that the content comes into existence at a source. A simple example of creating such multimedia might be a video clip taken with a camera and a microphone, or more likely video captured via a smartphone camera, and then transferred to the source. Likewise the client consumes the content, e.g., by displaying it on a computing device monitor, which again might be a smartphone screen watching a Internet video. We further assume that there is just one original source, but that there are many clients that want to receive the data. The clients value their viewing experience, and our goal is to reward the source for making the video available.

In a peer-to-peer (P2P) delivery approach, each client participates in the further delivery of the content. Each client makes part or all of the original content available to further clients. The clients become peers in a peer-to-peer delivery model. Such an approach is specifically geared towards being able to scale effortlessly to support millions of clients without prior notice, i.e., be able to handle a "mob-like" behavior of the clients.

The nature of the source data will dictate the exact details of delivery: for example, video data is made available at a preset quality using a variable-rate video encoder. The source data stream is divided into fixed length sequential frames: each frame is identified by its frame number. Clients request frames in sequence, receive the frame and reassemble the video stream which is then displayed using a suitable video decoder and display utility. The video stream is encoded in such a fashion that missing frames don't prevent a resulting video to be shown, but rather a video of lesser bit-rate encoding, i.e., quality, will result [7]. We explicitly allow the video stream to be quite malleable, i.e., the quality of delivery need not be constant and there is no harm if extra frames find their way into the stream. It is actually a key element of our approach that the stream can be enriched as part of the delivery process.

In our architecture, peers participate in peer groups. A peer is a network-connected computing device. The purpose of a peer group is to facilitate the dissemination of the multimedia data. Multimedia data, e.g., some video clip, comes into existence at a source. The source tells a single peer about its network location and addressability, i.e., IP and port number. The single peer serves as the "bootstrap" peer, it disseminates the knowledge about the video to the peers in its peer group. The source also advertises the single peer as the "seed" peer on the web. Peers can partake in the video stream either via being told by a peer in their peer group, or by retrieving the "seed" peer from its web advertisement and contacting the "seed" peer and joining its peer group. Peers can do 3 things: (1) they continuously request frames from other peers (the original source is viewed as just another peer) and store them; (2) they may display the frames as video to the user of the peer device; (3) and they make the stored frames available to other peers. Peers don't have to provide all 3 services. A peer that provides only service (1) and (2) is an "edge" peer, i.e., an end user consumer. A peer that provides service (1) and (3) is a "relay" peer. Relay peers are specifically important for peers that have limited access to the public Internet, i.e., peers behind network boundaries, such as a NAT firewall. In addition, peers stay in contact with each other to continuously update the peer group and source data availability.

Peer communication is achieved via session SIP messages. Each message has a message type and carries a payload. The initial message is of type "peer_join" that a new client peer sends to an existing peer in the peer group. The payload of the message contains the peer's public key, which will later be used to enable encrypted media delivery. The peer answers with a list of peers that currently make up the peer group. The "ping" message is sent periodically by peers to each other to establish whether they can reach each other: again, a peer that receives a "ping" message answers with its current list of peers in the peer group and its public key. Peers that have answered to a message are maintained as "neighbor" peers and will always by queried first. Another important type of message is "query_media", which inquires about which media is available and maintained by the peer group. The answer to this message is a list of which peers are able to serve which parts of the available media. The answer also provides communication details such as the IP and port number at which a peer will serve up frames of the media. Every peer constantly monitors the rate of response it gets from the other peers and adjusts its connections to the peers from which the highest throughput rate can be achieved.

Figure 1 shows an example snapshot of a content delivery network with one source, one bootstrap peer, 2 relay peers and one edge peer.

 ISBN: 978-1-61208-164-9

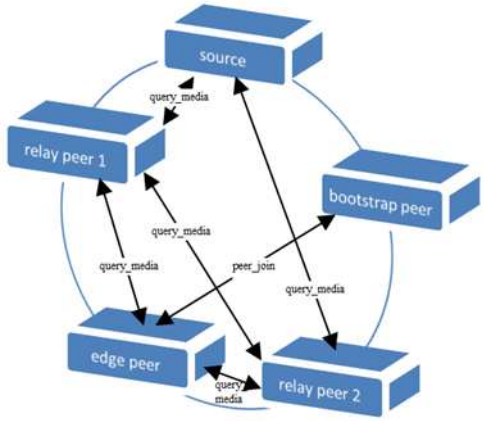


Figure 1. Peer Group with source & active peers

The source is where the video data is produced, encoded and made available. The bootstrap peer knows the network location of the source. Peers connect to the bootstrap peer first and then maintain sessions for the duration of the download: the 2 relay peers and the single edge peer maintain a peer group. The bootstrap peer initially informs the peers in the peer group which source to download from: peer 1 is fed directly from the source; peer 2 joined somewhat later and is now being served from the source and peer 1; the edge peer joined last and is being served from peer 1 and peer 2. In this example, peer 1 and 2 started out as edge peer, but became relay peers once they had enough data to start serving as intermediaries on the delivery path from original source to ultimate consumer.

IV. JAVA IMPLEMENTATION

Our implementation has 3 major components: a typical source application, a typical relay peer, and an edge peer to run on a mobile device. All 3 components are implemented in Java. We chose the Android platform to implement a proof-of-concept client for a mobile device. Android is part of the Open Handset Alliance [10]. Android is implemented in Java and therefore offers a flexible and standard set of communication and security features. The communication among the peers within their peer group uses session initiation protocol (SIP) messaging based on the Sip2Peer library [16]. The actual media exchange uses the Java implementation [13] of the SCTP [14] transport layer protocol. In the following we will first showcase the Android client, and then present details of the relay peer implementation.

Figures 2, 3 and 4 show three sample screen shots taken from the Android system. They illustrate our OghmaSIP media app.

Figure 2. OghmaSip Login Screen

Figure 2 shows the login screen to our OghmaSIP mobile client. It uses OpenID[6] user credentials and allows to establish a connection to a bootstrap peer via a web URL lookup. The client generates a pair of public/private keys and sends “peer_join” message to the bootstrap client.

Figure 3. OghmaSip Available Video Streams

Once the bootstrap peer has authenticated the new peer it will respond with a list of available video streams (Figure 3). After the user has made a selection, the screen shown in Figure 4 appears.

Figure 4. OghmaSIP Video Delivery Screen

Once a sufficient read-ahead buffer has been accumulated, the video stream starts playing on the Android device.

We also provide a Java desktop implementation of a peer. The typical peer is a "Relay" peer, i.e., it will request media frames from the source, potentially show them locally to a user, and then make these frames available to other peers. Peers that wish to participate in the content delivery must first locate media sources. A peer will start by looking up the bootstrap peer via its web advertisement. Like the mobile client, the typical "relay" peer generates a public/private key pair and sends a "peer_join" message to the bootstrap peer. Figure 5 shows the relay peer's graphical user interface that tracks the peers in the peer group: the center of the screen shows peers that have been accepted into the P2P content delivery network; the bottom of the screen shows a log of access requests from other peers. Overlaid is a popup-screen showing the public key information of a selected peer.

At least one source must exist for the content delivery network to get started. The source first advertises its bootstrap peer. It generates a PKI [11] public/private key pair and transmits its public key to the bootstrap peer. It then stands ready for data requests from clients. If a request from a client peer is received, it looks up the client's public key and uses a Diffie-Hellman key agreement algorithm [12] to produce a session key. The session is then used by the source to encrypt all data that is sent to the client. Peers that become "relay" peers use the same method to encrypt frames as they are sent to other peers.

Our prototype uses the Java implementation [13] of the SCTP [14] transport layer protocol. SCTP is serving in a similar role as the popular TCP and UDP protocols. It provides some of the same service features of both, ensuring reliable, in-sequence transport of messages with congestion control. We chose SCTP because of its ability to deliver multimedia in multiple streams. Once a client has established a SCTP association with a server, packages can be exchanged with high speed and low latency. Each association can support multiple streams, where the packages that are sent within one stream are guaranteed to arrive in sequence. Each source can divide the original video stream into set of streams meant to be displayed in an overlay fashion. Streams can be arranged in a way that the more streams are fully received by a client, the better the viewing quality will be. The first stream is used to deliver a basic low quality version of the video stream. The second and consecutive streams will carry frames that are overlaid onto the primary stream for the purpose of increasing the quality. In our framework we also use the additional streams to carry content that is "added value", such as advertising messages or identifying logos. The ultimate client that displays the content to a user will combine all streams into one viewing experience.

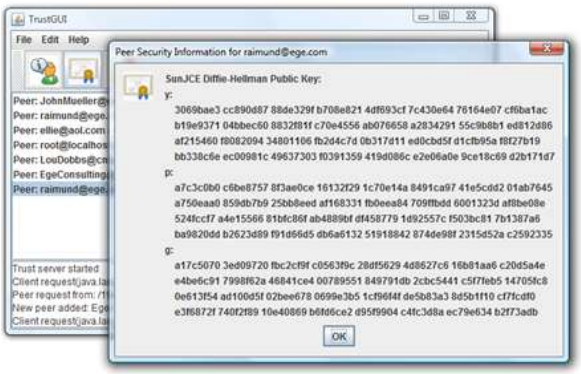


Figure 5. Peergroup listing and security info

V. CONCLUSION

In this paper, we described a framework for new content delivery networks that almost implements access control for its participating peers. We have described a prototype implementation that uses SIP messaging to establish a P2P network, where a group of peers disseminate information on which sources are available to download from, and includes a Java-based client for the Android platform for smart phones. Such P2P content delivery has great potential to enable large scale delivery of multimedia content. Our framework is designed to enable content originators to assess the potential reward from distributing the content to the Internet. The reward is quantified as the value added at each peer in the content delivery network and gauged relative to the actual cost incurred in data delivery but also correlated to the risk that such open delivery poses.

Consider the scenario we described earlier in the paper: a typical "viral" video found on a social networking site: the video is captured on the fly, then uploaded onto the site, stored and transmitted for free and viewed by a large audience. The only entity that is getting a reward is social media site, which accompanies the video presentation with paid advertising. The only benefit that the original source of the video gets is notoriety. Using our model, the original data owner can select other venues to make the video available via a peer-to-peer approach. The selection on who will participate can be based on how much each peer contributes in terms of reward but also risk. Peers will have an interest in being part of the delivery network, much like Facebook and YouTube have recognized its value. Peers might even add their own value to the delivery and share the proceeds with the original source. Whereas in the social media approach the reward is only reaped by one, and the original source has shouldered all the risk, i.e., lost all reward from the content, our model will enable a more equitable mechanism for sharing the cost and reward.

REFERENCES

[1] Raimund K. Ege. Trusted P2P Media Delivery to Mobile Devices. Proceedings of the Fifth International Conference on Systems (ICONS 2010), pages 140-145, Menuires, France, April 2010.

[2] Y. Atif. Building trust in E-commerce. IEEE Internet Computing, 6(1):18–24, 2002.

[3] P. Resnick, K. Kuwabara, R. Zeckhauser, and E. Friedman. Reputation systems. Communications of the ACM, 43(12):45–48, 2000.

[4] E. Bertino, B. Catania, E. Ferrari, and P. Perlasca. A logical framework for reasoning about access control models. In SACMAT '01: Proceedings of the sixth ACM symposium on Access control models and technologies, pages 41–52, New York, NY, USA, 2001.

[5] S. Jajodia, P. Samarati, M. L. Sapino, and V. S. Subrahmanian. Flexible support for multiple access control policies. ACM Transaction Database System, 26(2):214–260, 2001.

[6] OpenID, http://www.openid.net. [accessed September 22, 2010]

[7] C. Wu, Baochun Li. R-Stream: Resilient peer-to-peer streaming with rateless codes. In Proceedings of the 13th ACM International Conference on Multimedia, pages 307-310, Singapore, 2005.

[8] R. K. Ege, L. Yang, Q. Kharma, and X. Ni. Three-layered mediator architecture based on dht. Proceedings of the 7th International Symposium on Parallel Architectures, Algorithms, and Networks (I-SPAN 2004), Hong Kong, SAR, China. IEEE Computer Society, pages 317–318, 2004.

[9] L. Yang, R. Ege, Integrating Trust Management into Usage Control in P2P Multimedia Delivery, Proceedings of Twentieth International Conference on Software Engineering and Knowledge Engineering (SEKE'08), pages 411-416, Redwood City, CA, 2008.

[10] Open Handset Alliance, http://www.openhandsetalliance.com/. [accessed November 19, 2010]

[11] Gutmann, P., 1999. The Design of a Cryptographic Security Architecture, *Proceedings of the 8th USENIX Security Symposium*, pages 153-168, Washington, D.C., 1999.

[12] Network Working Group, Diffie-Hellman Key Agreement Method, Request for Comments: 2631, RTFM Inc., June 1999.

[13] java.net – The Source for Java Technology Collaboration, The JDK 7 Project, http://jdk7.dev.java.net. [accessed September 22, 2010]

[14] R. Stewart (ed.), Stream Control Transmission Protocol, Request for Comments: 4960, IETF Network Working Group, September 2007, http://tools.ietf.org/html/rfc4960. [accessed September 22, 2010]

[15] Raimund K. Ege, Li Yang, Richard Whittaker. Extracting Value from P2P Content Delivery. Proceedings of the Fourth International Conference on Systems (ICONS 2009), pages 102-108 Cancun, Mexico, March 2009.

[16] Sip2Peer, SIP-based API for robust connection and communication among peers, http://code.google.com/p/sip2peer/. [accessed May 9, 2011]

 ISBN: 978-1-61208-164-9

An NFC-based Customer Loyalty System

Jef Smets, Glenn Ergeerts, Rud Beyers, Frederik Schrooyen, Marc Ceulemans, Luc Wante, Karel Renckens
Department of Applied Engineering
Artesis University College of Antwerp
Antwerp, Belgium
glenn.ergeerts@artesis.be

***Abstract*—Customer loyalty systems that use barcode-based cards have gained a lot of popularity in the last decades, resulting in customer wallets that are overwhelmed with barcode-based loyalty cards. In this paper, a solution for this problem is provided. Based on general and market research of customer loyalty systems, a customer loyalty system that uses NFC (Near Field Communication) technology is designed that requires only one NFC medium (e.g., an NFC-enabled smart card or an NFC-enabled mobile device) for each customer, which is capable of holding multiple virtual loyalty cards.**

***Index Terms*—Customer loyalty, NFC, group loyalty, city loyalty, NFC smart card, mobile NFC.**

I. INTRODUCTION

As an old proverb states: the customer is always right, and nothing is more truthfully nowadays. Due to the current economical crisis and the rising expectations of the technology-enriched customers, the variableness and the switching behaviour of customers has increased significantly [1]. For this reason, many companies are forced to turn into customer-centric companies in order to attract new customers and retain the existing ones.

In a customer-centric company, the focus lies on the needs and behaviours of the customers instead of on the companys internal drivers. One of the key factors of a customer-centric company is customer loyalty, a very powerful tool for merchants [2]. By rewarding customers for their purchases, customers are retained. Furthermore, new and existing customers are attracted by publicity campaigns, resulting in an increased turnover. Research [1] has shown a revenue increase of 20% per customer. Other advantages of a successful loyalty system are marketing effectiveness, building true loyalty, increased word of mouth (WoM) marketing, strengthened value and brand proposition and increased long-term profits.

Existing loyalty systems are in need of improvement [3]. In this paper, a system is described that takes customer loyalty to a higher level, giving both merchants and customers new opportunities to enrich their relationship and build true loyalty. Decreased marketing costs and increased marketing effectiveness by using real-time sales data are the main benefits for merchants.

NFC (Near Field Communication) technology is used as enabler-technology to ensure an efficient and convenient usage of the system. This very promising and relatively young technology is slowly penetrating the market and has a number of advantages over traditional loyalty systems technologies (e.g., loyalty points on a barcode-based medium and paper strips for some short-term promotion).

This paper proposes a NFC-based loyalty system which allows the customer to use only one medium containing multiple virtual loyalty cards. From a merchant point of view the main advantage the system delivers is the easy implementation and managing of a loyalty system by joining the platform.

In the following section, we give a description of the concept of customer loyalty and in Section 3 we give a brief overview of NFC technology. In Section 4 we discuss the customer loyalty system. Section 5 handles about the performed interviews. Future works can be found in Section 6 and finally, the conclusion of this paper is located in Section 7.

II. CUSTOMER LOYALTY

The main focus of customer loyalty is to retain customers. In order to achieve customer loyalty, a relationship between a store or brand and the customer has to be built up. This is done by rewarding loyal customers. A customer is considered loyal when the customer actively participates in a loyalty program. Promotions and other forms of publicity are also used to increase the retaining rate of customers and to attract new customers, who, hopefully, will be converted into loyal customers.

Some examples of loyal customers rewards are a reduction on the total price, a cash value, a free product, a reduction on a specific product, a lottery game ticket (of a lottery game organised by the store or company), a reduction in another store or even a reduction in the customers parking ticket price. A larger reward is likely to retain more customers, but also decreases the revenue.

Coupons or promotion codes, which a customer can exchange for a reward, are also often used for attracting new customers or for rewarding loyal customers. When the reward is a free product or reduction, sometimes a catalogue is available from which customers can choose their free product or a product with reduction. A membership card is a kind of loyalty card as the members can be considered as loyal customers that are entitled to a reward.

A distinction can be made between issuers, who want to increase the loyalty of their own products or brand, and merchants, who want to increase the loyalty of their stores, which offer a wide range of products. In this paper, the focus

 ISBN: 978-1-61208-164-9

lies on merchants because issuers only distribute coupons, not loyalty cards.

In order to increase loyalty, customers must be satisfied about the given reward and the received publicity. Because no customer is the same, there is no one-size-fits-all solution for delivering rewards and making publicity. Therefore, the concept of customer segmentation, where customers are divided into groups with each their specific characteristics regarding loyalty, is used. In this way, customers are better targeted by the publicity campaigns and more satisfied about the given rewards, thus increasing the customer loyalty. However, if the customer segmentation is done badly, customers will receive rewards and publicity in which they are not interested, resulting in customer churn. Its therefore important that good analysis of the customers is performed and optimally used in order to make a profitable loyalty system.

Research of Accenture [1] summarizes this in a three-step scheme. The first step is about knowing the customer in order to choose good segmentation criteria. The second step is all about customer-centric marketing (the actual publicity to lure customers to the store). The rewarding of loyal customers is discussed in the third step.

As loyalty can be increased by dividing the customers into groups, we can state that an optimal loyalty level can be reached by examining each customer separately. This approach is called one-to-one marketing [4]. Using one-to-one marketing, fully customized campaigns can be generated to reach the customers optimally and to boost loyalty. There can also be a much faster response to the increasingly switching behaviour of customers.

In order to measure the success of a loyalty system, so called loyalty or marketing factors must be carefully chosen and constantly measured. Those loyalty factors will determine the success of the chosen rewards and incentives, and the chosen customer segmentation.

Marketing agents have to keep in mind to make offers compelling but not too intrusive in order not to spam the customers. It is after all the customer that decides to opt-in to the promotions or not. Also, if customers take advantage of the system and better organise their redemptions, the system will have less impact [5]. Systems like mFero [6] or Puntavista [7], which help the customer to choose the best reward, reduce the complexity of choosing the best reward at the POS (point-of-sale) and could reduce the impact as well.

Todays younger generation uses smartphones intensively but good examples of mobile marketing are still hard to find. Research has shown that people care more of losing their cell phone than their wallet [8] and that 75% of the people take their cell phone with them everywhere [3]. Keeping that in mind, it is no surprise that mobile marketing is a hot topic nowadays. Customers that can be reached everywhere and at any time, using a different amount of technologies, make the customers mobile device an ideal marketing medium. The wide range of technologies that a smartphone offers can be used to for time based and/or location based loyalty at low cost.

III. NFC TECHNOLOGY

NFC (Near Field Communication) technology is a close-range communication technology with a typical operating distance of approximately 10 cm [9]. Its operating frequency is 13.56MHz. In 2003, NFC has been approved as ISO/IEC standard making it a very modern technology.

Next to its shorter distance and other operating frequency, NFC differs from other technologies such as RFID, Bluetooth, ZigBee, IrDA and Wi-Fi by a slower data rate [10]. The maximum data rate of NFC is 424 kB/s. This small data rate is no problem, as the size of the by NFC transmitted information is usually small as well.

NFC is, due to its short operating range, said to be very secure. However, research [11] has proven that some important related security issues like eavesdropping, data corruption, data modification, data insertion and man-in-the-middle-attacks remain. Solutions are available but not always easy to implement.

Typical NFC applications are smart posters, payment or ticketing and loyalty. Many different NFC trials in countries around the world have been held from those applications. However, during the last years the research about NFC has focussed primarily on mobile NFC [8]. Mobile NFC is the integration of NFC in mobile phones whereby mobile payment systems are primarily very popular nowadays.

IV. SYSTEM OVERVIEW

The idea behind the system described in this paper is that customers can replace all their loyalty cards with one NFC medium. This medium could be an NFC card or, in the future, an NFC GSM, which most customers already carry with them. This should be well accepted by customers since they are overwhelmed with paper vouchers and plastic loyalty cards nowadays. Issuers and merchants have lower costs and are more eco-friendly since theres no need for printing and distributing paper vouchers or catalogues.

A customer then only needs one NFC medium that is usable in different stores. Many actions can be automated, reducing processing time at the terminal. NFC is also said to be faster since there is no need for searching for a loyalty card in the customers wallet as this is automatically done by the terminal [5]. The tap and go functionality of NFC devices allows the customer to perform more intuitive actions which makes NFC an easy-to-use technology and results in an increased impulsive buying behaviour.

Since multiple stores use the same medium, the price can be split among them. Also, a medium already owned by the customer could be used, reducing the medium costs to zero. A disadvantage is that the branding aspect of traditional loyalty cards is lost, but the customers interface can be used for better, more personalised publishing resulting in increased customer loyalty.

A. System, terminals and mediums structure

The system consists of a backend which contains the webserver and database, terminals which are located at the

 ISBN: 978-1-61208-164-9

POS (point-of-sale) in the stores, NFC mediums (possessed by the customers) and a website as an online interface for both customers and merchants. This is shown in Figure 1.

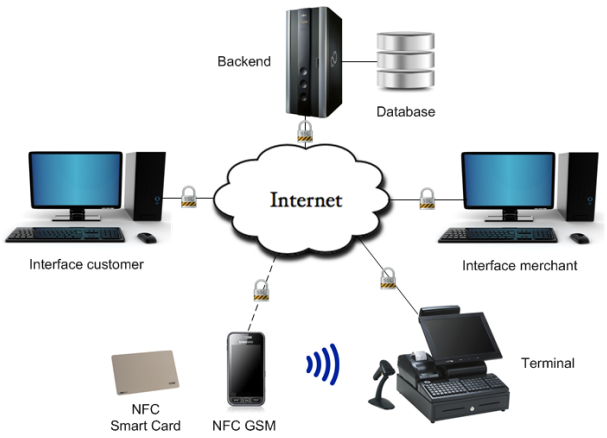


Fig. 1. System overview

An online architecture is used, meaning that all data is stored in the central database and the customers medium is only used for identification. The data is accessible via the online interface of the customers and merchants in real-time, which enables real-time monitoring of the customers (marketing information) on the merchants interface. The online medium and terminal architecture is needed to ensure this real-time data on the backend and website.

All operations (processing of the data) are performed on the backend (alternative is on the terminals) which enables efficient updating of the software. When implementing a new condition (see Section 4.3), only the code on the backend needs to be modified, all terminals can use the new backend code without need for updating. Also, when there are multiple branches of a store, all the data is already centralised; so, no additional synchronisation is needed (in contrary to when all or part of the data would be stored at the terminals).

Choosing an online architecture for the medium makes the system medium independent as only an identification number needs to be saved on the medium and all mediums support this. The most common NFC mediums are NFC smart cards and mobile NFC devices, as discussed in the previous Section. However, mobile NFC devices have not yet penetrated the market sufficiently, but NFC smart cards have (reaching even 95% penetration in Hong Kong) [8]. Therefore, smart cards are currently the most suited medium to use.

With 2011 that is announced as a promising year for mobile NFC, the breakthrough of mobile NFC will not last much longer, so the use of mobile NFC devices is supported by the loyalty system in order to be future proof. Its a small step to port NFC cards to mobile NFC devices. The intermediate solutions towards mobile NFC devices that currently exist, NFC stickers and NFC MicroSD cards, could also be used as a solution for the current absence of mobile NFC devices [12].

Because all data is stored in the database on the backend, an OTA transaction is required when us ing a mobile NFC device to consult this data. A semi-offline mobile NFC medium would not require an OTA transmission but as the new generation of smartphones all support mobile internet and the current price of mobile internet is descending, this should not be a problem.

An online architecture requires no memory storage on the medium (which is limited in size). This leaves room for other applications, it enables the possibility to have multiple users per medium and backups can easily be managed since all data is centrally available on the backend. A backup could be used to restore the data of a stolen medium which can be blacklisted instantly.

The only disadvantage of an online system is that the terminals require a constant connection with the backend to access the database. This is more expensive than a non-constant connection, the used terminals have to support network connections and there is the possibility that the connection is broken. However, using a buffer, in case of a broken connection the system can switch to a limited operation mode where it is still possible for customers to gain loyalty points but not to redeem them (see Section 4.4). In this way, customers still have to come back to the store (where loyalty is all about), so the system is still effective. The connection should be restored as soon as possible to ensure a full operating system in which customers can also redeem items so the customers loyalty experience remains positive.

Considering the terminals, three different terminals were implemented. Firstly, loyalty functionality was added to an online mobile terminal, the ACR880 GPRS Portable Smart Card Terminal. Secondly, an online fixed desktop pay desk was implemented and extended with an NFC reader, barcode reader, eID reader and loyalty functionality. Finally, the same fixed desktop terminal was implemented and extended with loyalty functionality to work in a semi-offline manner, which means processing the data on the terminal and logging all resulting data to the backend system at regular times. Advantages of this approach are no need for a constant connection with the backend and the ability to use product-specific conditions without the need for the product database to exist on the backend. The data on the backend (and the interface) is not real-time. Sections 4.3 and 4.4 dont apply to this terminal type as the data processing and management are not system-specific.

B. Support for different kinds of loyalty

Normal loyalty fits the scenario of stores with possible multiple branches with each one or more terminals per store. The following kinds are supported as well so more customers can be reached. All the different kinds of loyalty can be combined at will.

If a store has already an electronic barcode-based loyalty system in place, this system can be replaced by storing the barcode of the current loyalty card of a customer on the customers medium. An extra module would be needed in the terminals to support the NFC mediums in the system and to

 ISBN: 978-1-61208-164-9

log all data to the backend system. The semi-offline terminal could be used for this purpose. The advantage of this approach is obviously the support for existing barcode based loyalty system, which are very popular nowadays.

Group loyalty fits the scenarios of common loyalty points for a group of stores with each possible multiple branches and one or more terminals per store. The problem arises that the points have a monetary value. For example, customer X gains 499 points in store A after 10 visits, then gains 1 point in store B (1 visit) and receive a reduction of 5 (promotion says: for each 500 points, you receive a reduction of 5). This would mean that store B would pay for 1 visit in store B and 10 visits in store A which isnt correct.

This money mismatch requires all transactions to be logged to make regular money exchanges between all participation stores possible. Another option is that the stores can buy and sell the points from and to the system. Points need to be available on-the-fly and a store should never run out of points. Due the systems online infrastructure, this could be implemented without much effort. The monetary value of points also raises the need for increased security of the system. Advantages for the participating stores are not only raised profits due to the increased foot traffic but shared publicity costs as well.

As an extension of group loyalty, city loyalty additionally has third-party payment devices in the loyalty environment, e.g., a parking meter. Cities are found the most suitable places to perform this kind of loyalty, hence the name, but city loyalty is not restricted to cities at all, e.g., shopping malls are also a perfect location. The advantages of group loyalty remain. In addition, when using this kind of loyalty, customers find themselves in a complete loyalty environment, which they are yet more likely to return to.

Next to store-specific loyalty, product-specific loyalty is also an option as stated in Section 2. When enough stores use the system, cross-store product-specific promotions (branding) could be organised by the manufacturers, bringing new players to the loyalty ecosystem.

C. Promotions configuration

A difference is made between receiving points and receiving a reward (redeeming points). Both are triggered by conditions. There can be different kinds of conditions implemented. Examples of receive conditions are: receiving points for each visit, receiving points as a function of the total price, receiving points as a function of the price of a specific product or receiving a fixed amount of points for each visit. There are also different kinds of rewards as discussed in Section 2. All rewards are triggered by an amount of points.

When a promotion is configured, a merchant chooses which receive and redeem conditions to use. Minimum one of each is required but multiple conditions can be combined at will. Each chosen condition needs to be configured. The system can easily be expanded with new conditions, not requiring updating the terminals due the online architecture as mentioned above.

Multiple promotions are possible on a stores virtual loyalty card just as this is now the case with traditional loyalty systems.

D. Use cases

1) Normal flow: The following steps summarise the general customer flow at the terminal.

1) The cashier enters all products
2) The cashier asks the customer to put the medium on the reader
3) The cashier asks the customers which promotions and items of that promotions the customer wants to receive points for
4) The customers choice is sent to the backend and processed, a list of available rewards is returned. If this is the first time a customer uses the loyalty card in that store, a virtual loyalty card is created automatically. If this is the first time a customer uses the promotion on the virtual loyalty card of that store, a counter is created automatically.
5) The cashier asks the customers which rewards the customer wants to collect. If a customer doesnt select a reward, no points are redeemed. This enables a customer to save for a certain reward, e.g., 100 points give a free magazine and 200 points give a free book, etc. Another possibility is that the reward is e.g., a free hamburger and that the customer already bought food and wants to postpone the reward to the next visit.
6) The customers choice is sent to the backend and processed.
7) Next customer.

Some of the promotion items to receive points and receive a reward are chosen automatically. This is configured by the merchant on the merchants part of the website.

Received points can be added to the points balance before there is checked if enough points are available to receive a reward or afterwards. This is also configured by the merchant on the merchant part of the website. If afterwards, this means the customer will have to return to the store in order to retrieve the award (which is the basic of loyalty, letting customers return to the store).

2) New customer flow: Customers can start saving points immediately after receiving their medium in the store and providing their eID information the first time they use their medium to save points.

The systems goal is to increase loyalty. In order to achieve this goal, merchants must have access to information of their customers. This information is obtained when an NFC medium is used for the first time to ensure every new customer provides this information.

Customers will have the choice between using an eID reader or handover the eID to the cashier which will enter the data manually in a form (in order to ensure correct data). The manual option is for customers who fear their privacy as it is unclear for the customers which data is obtained from their eID.

 ISBN: 978-1-61208-164-9

The advantage of using an eID reader for entering the customers information is the reduced processing time. Instead of paper registration forms that need to be processed (manually entered in a computer system), this is done automatically, the customers eID data is even available in real-time. In addition, the entered eID data has a very low error-rate.

Following steps summarise the new customer flow at the terminal.

1) A new customer asks for and receives an NFC card (or uses an already owned NFC medium)
2) The cashier enters all products
3) The cashier asks the customer to put the medium on the reader
4) The cashier asks the customers which promotions and items of that promotions the customers want to receive points for
5) The customers choice is sent to the backend and processed, a list of available rewards is returned. If this is the first time a customer uses the loyalty card in that store, a virtual loyalty card is created automatically. If this is the first time a customer uses the promotion on the virtual loyalty card of that store, a counter is created automatically.
6) The cashier gets a notification that the customer is a new customer. The customer is asked to handover the customers eID card.
7) The cashier reads out the eID card or enters the data of the eID card manually (if preferred so by the customer)
8) The cashier asks the customers which rewards the customer wants to collect If a customer doesnt select a reward, no points are redeemed. This enables a customer to save for a certain reward, e.g., 100 points give a free magazine and 200 points give a free book, etc. Another possibility is that the reward is e.g., a free hamburger and that the customer already bought food and wants to postpone the reward to the next visit.
9) The customers choice and eID data is sent to the backend and processed. An un-activated customer website account with the eID data linked to is created at the backend.
10) The customers registration code is printed on the cash ticket.
11) Next customer.

After using their medium for the first time, customers can (optionally) register an account on the website interface using their registration code which is printed on the cash ticket.

E. Customer interface

Customers have an online, browser based interface (website) on which they can view the stores of which they have a virtual card and view their points balance and history of each promotion. In their account settings they can opt-in and opt-out to various aspects of the system so at any time their privacy is guaranteed.

All stores that use the system and their promotions can be viewed. Customers are informed about new promotions and other news via advertising on the pages. Finally, each store has also its own information page with the stores address details, a map, opening hours and other information.

Customers can also add additional profile data using the website as an extension to the eID data as this data is not complete (e.g., phone numbers are missing). The customers email address is retrieved during the registration process.

F. Merchant interface

On the merchant interface (website), merchants can manage the promotions of their store. Each store has its own space on the website with information about the store such as the stores address, opening hours and general information.

Merchants have access to the points balance and history of their customers, the (relevant) eID information of their customers and other marketing information of their customers and promotions which allows them to increase loyalty. Different graphs and other tools are available, e.g., a map with all addresses of their customers in order to view the distance to the store or a graph that shows the amount of customers that visited the store during the past week.

V. Security & privacy

Various security levels were implemented in the system. Firstly, all communication between the backend system and clients (website users and terminals) is encrypted using SSL. Secondly, Web pages or web services require authentication (expect the public pages). Finally, each user or terminal is restricted to see only the appropriate data. The system should prohibit data duplication (points). This is obtained by securing the backend, all data is stored there.

Considering the privacy, access to customers data (personal data and loyalty data) is restricted to merchants of stores where customers have a virtual loyalty card. One is considered a customer of a store if a virtual loyalty card of that store exists. By visiting a store, a customer opts-in to that store and a virtual loyalty card is created, giving the merchant of that store access to that customers profile information.

The eID reader brings privacy issues with it as well. Laws exist that protect the customers privacy and that should not be ignored and, as discussed above, an alternative to the eID is provided by giving the option to let cashier manually insert the eID data in to the system.

Confidence agreements between merchants and the system administrators have to be made as well. Those regard the loyalty and sales data of the stores which should not be sold to third-party organisations by the systems administrators.

VI. Interviews

7 interviews were held in 5 different stores, ranging from small local stores to national stores with multiple branches. The currently used loyalty mediums vary from none to paper and barcode cards. Both normal loyalty and group loyalty were encountered.

The system was well accepted in all stores. The same goes for the NFC technology. Merchants find NFC easy to use and

 ISBN: 978-1-61208-164-9

faster than traditional loyalty mediums. The possibility to opt-in to the loyalty system and keep the existing (barcode) loyalty system was found to be a good alternative.

The opinions about the competition between the stores when using a shared medium and the loss of the mediums publicity aspect are divided. However, everyone is enthusiast about the possibilities of the online platform. The customer interface and the merchant interface were found well-thought out and complete.

The layout and usability of the website were rated very high. None of the interviewees found any missing parts in the interface pages. We made the suggestion of adding billing information on the interface but this was found not needed and to complex and confusing for the customers by the interviewees.

The interviewees have, as expected, wishes for some specific information of the customers regarding their preferences (store-specific marketing data). Further, the interviewees liked that the offers and publicity can be personalised using the marketing data on the website and noticed this is already possible with their current loyalty system but not done as it takes to much effort. Loyalty on product level is a requirement of the larger stores. The small stores were only interested in the basic functionality of the system. Social media integration was also well accepted by all stores. See the next Section for more information about those items.

The use of an eID reader to collect customer information was also well accepted. Some of the stores admitted that they never used the collected paper forms because the processing took too long while other stores simply dont gather customer information due the manual processing.

Finally, privacy of the customers was found to be a delicate subject. The interviewees are very aware of the fact that customer privacy should not be ignored.

VII. Future works

A. Backend

Barcode replacement, group loyalty, city loyalty, product-specific loyalty (branding) and combinations of those, discussed in Section 4.2 are not implemented yet and can be added to the system.

Not all conditions, discussed in section 4.3, are implemented yet. Note that the implementation of product-specific conditions will require product information to be stored in the backend database as this data is needed for the processing of the product-specific conditions which takes place on the backend. The product data is also needed on the terminals to ensure a fast barcode/price lookup. Product-specific conditions are currently implemented on the semi-offline fixed desktop terminal, avoiding the need for a product database on the backend.

On the merchant interface managing the advertising (adding and removing ads), new graphs on the marketing pages and a marketing data export functionality, to export the marketing data to a more advanced CRM (customer relationship management) program, could be implemented. The marketing data should be usable to segment customer into groups and ideally to perform one-to-one marketing by generating more personalised advertising (on the customer interface), publishing (via folders and mailings) and promotions based on the customers preferences. This can already be obtained but is not yet fully automated which is required in order to provide a full surplus to the system. When a store has multiple branches, it should be possible to manage those branches and filter the marketing data per branch.

Both on the customer and merchant interface newsletter functionality and/or a news page per store could be implemented. Also, integration with social media is possible. Merchants and customers could opt-in to a loyalty system where the customers social media is used for publishing in exchange for loyalty points (given to the customers). The online interface can also be used to provide support towards the customers and merchants.

The billing part for the merchants needs to be implemented. A flat fee could be used or the billing could be directly based on advertising, server space (database space), the number of terminals, the number of customers the number of website users, the number of distributed NFC cards and/or NFC GSMs and/or other parameters.

B. Terminals

The buffer functionality, discussed in section 4.1 is not yet implemented on some of the terminals. Also, the customer flow is not fully implemented on the terminals. The selection of promotion receive and redeem items (including auto selected items) is not possible.

The customer flow at the terminal and or the website can be extended to show the rewards a customer may expect next visit. It should be possible for the cashier to manually enter a custom amount of loyalty points instead of choosing a predefined promotion in order to make corrections or to make a one-time loyalty receive or redeem transaction.

Modifying the eID data should be done via the eID reader or form on the terminals, instead of via the website, to ensure correct eID data of the customers. Customers should still be able to modify the additional customer information via the website (email address, phone numbers, etc.).

When the system is combined with an NFC payment system, even faster processing times at the terminals can be achieved.

C. Mediums

New possibilities arise when an NFC smartphone is used as medium as discussed in section 4.1. The customers and/or merchants interface can be accessed using a mobile web browser or a smartphone application. A smartphone application is preferred due the higher user-friendliness. Such an application requires OTA application management (downloading, installation, updating and removal).

It is also possible to perform P2P (peer-to-peer) transactions using NFC smartphones. This functionality could be used for exchanging points between customers (if permitted).

 ISBN: 978-1-61208-164-9

The system can be extended by Smart Posters (only work with active NFC mediums such as an NFC smartphone) or Smart Kiosks (also work with passive mediums such as an NFC smartcard). When using a Smart Poster or Smart Kiosk a semi-offline medium is required. Using a mobile NFC device and an OTA connection with the backend, a semi-offline medium could be emulated, enabling the use of a Smart Poster.

Finally, if required, a terminal page on the merchant interface or a terminal smartphone application could be implemented.

VIII. Conclusion

After performing extended market research and literature research an NFC customer loyalty system is developed that bundles all advantages of NFC technology and traditional customer loyalty systems.

An online system architecture is found to be the best choice and an online platform is created. An optimal customer flow is designed to ensure a convenient and user-friendly loyalty experience. During the implementation process there was ensured at any time that the system was as generic as possible. The system has been tested extensively. Convenience, speed, security and privacy were fully considered.

Next to normal loyalty, other kinds of loyalty such as group loyalty and city loyalty are supported. Its also possible to extend existing loyalty systems to interoperate with the system.

The backend framework, the website (customer, merchant and public interfaces) and the web services for the terminals are implemented. The Event Wallet Mifare Desfire NFC smart card is used as medium. Three different kinds of terminals are implemented. The mediums are portable to mobiles NFC devices, making the system future proof.

The interviews learn that the choices that were made during the research and implementation phase are correct. Merchants are waiting for automated one-to-one marketing and an online platform is preferred. This customer and merchant interface is found to be complete and user-friendly, because most users are familiar with web interfaces.

Considered future works, there are a lot of possibilities to go from now, also due the generic structure of the system.

References

[1] W. Driggs and N. Kasolowsky, “Customer acquisition and retention: Creating customer loyalty a customer-centric approach,” 2008, [accessed 10-July-2011]. [Online]. Available: http://www.accenture.com/SiteCollectionDocuments/PDF/Accenture_Creating_Customer_Loyalty_A_Customer_Centric_Approach.pdf

[2] F. R. Marc Hoogenberg and S. Juchnowicz, “Customer centric loyalty programs as key to high performance retailing,” 2010, [accessed 10-July-2011]. [Online]. Available: http://www.sylwiajuchnowicz.nl/wp-content/uploads/2010/01/Customer-Centric-Loyalty-Programs-as-Key-to-High-Performance-Retailing.pdf

[3] W. Driggs and N. Kasolowsky, “Serving the new mobile consumer - jumpstarting a new generation of mobile retail solutions.” [Online]. Available: http://wireless.accenture.com/SiteCollectionDocuments/PDF/Accenture_POV_Serving_the_New_Mobile_Consumer_Web.pdf

[4] D. Klabjan and J. Pei, “In-store one-to-one marketing,” pp. 64–73, 2010.

[5] J. Ondrus and Y. Pigneur, “Coupling mobile payments and crm in the retail industry,” p. 7, 2004.

[6] R. K. Balan and N. Ramasubbu, “The digital wallet: Opportunities and prototypes,” p. 2, 2009.

[7] “Puntavista,” [accessed 10-July-2011]. [Online]. Available: http://www.puntavista.be/nl

[8] J. Fischer, “Nfc in cell phones: the new paradigm for an interactive world,” pp. 22–26, 2009.

[9] K. Grassie, “Easy handling and security make nfc a success,” pp. 12–13, 2007.

[10] S. O. Jr., “Is near-field communication close to success,” pp. 18–20, 2006.

[11] E. Haselsteiner and K. Breitfu, “Security in near field communication (nfc),” pp. 4–9.

[12] S. C. Alliance, “Chip-enabled mobile marketing,” 2010, [accessed 10-July-2011]. [Online]. Available: http://www.smartcardalliance.org/pages/publications-chip-enabled-mobile-marketing

 ISBN: 978-1-61208-164-9

Amazon-on-Earth Library Navigator

Indoor Navigation Using Un-Augmented Mobile Phones

Amnon Dekel, Scott Kirkpatrick, Niv Noach, Barak Schiller
The School of Computer Science & Engineering
The Hebrew University Jerusalem
Jerusalem, Israel
[amnoid, kirk] @ cs.huji.ac.il, [nivnoach, bschiller] @gmail.com

***Abstract*— This paper describes the Amazon-on-Earth project that enables users to look for, navigate to and find objects of interest inside a physical space. We implemented a working prototype system in one of the libraries on our campus and ran a user study to see if there was any advantage to using the system relative to the existing library information services. Results show that subjects using our service were able to find information about a book 35% faster and were able to navigate to a book 52% faster. We also found that the control group made four times as many navigational errors and had to ask for help five times more than the experimental group. Measuring qualitative variables we found that subjects using our system rated the ease of finding the book in the library as easier than the control group and felt more positive towards the service. The results make it clear that the use of such a service can substantially help a user find an object inside a space in a faster, easier and more pleasant fashion. We end the paper by pointing out a number of shortcomings of the system and how they might be dealt with.**

Keywords- Indoor Navigation; Mobile Phone; QR Code

I. INTRODUCTION

For a number of years now, efforts have been made in research and product circles to enable digital information services within the context of real world scenarios. The growing sector of powerful Smartphones with their multiple embedded sensors and networking systems have made them a prime focus in consumer based location based services. These have spawned a number of commercial systems that can point out relevant physical services close to a person's current location [3, 6]. Such systems are mostly used while driving or walking near or within a shopping area. While our system uses location as an important dimension, we focus on using ***indoor*** navigation to enable users to find ***objects*** in the physical world and to interact with them. Thus, our "objects" do not transmit their existence to the world (there are simply too many of them to make this feasible)- but once a user using our navigation maps finds them, the system enables a number of relevant functions to be enacted in relation to the object.

This paper has the following structure: we start with a short section about previous related work. We then describe the system we have built and present the empirical study we have run to test its effectiveness. We then present the results of our study and end with a discussion of shortcomings and how the system can be improved.

II. RELATED WORK

1) Mobile Interaction with the Real World

Smith et al. [15] presented a prototype for mobile retail and product annotation services. Their system enabled the user to scan the object's barcode and receive relevant information about that object which was found on existing web services such as Amazon.com. Their system used a special purpose barcode scanner to decode the object's ID for further querying (since then, 1D and 2D visual tag decoding software have become available for most Smartphone systems). But their system did not help users find an object within a physical space, nor to conduct a transaction to buy the object if the user wished to. Many additional research projects have focused on this space in the last few years, i.e., [3][4][7][13].

Broll et al. [4] present models for tag – service interfacing. The physical tag (be it a visual barcode or a Near Field Communication (NFC) RFID tag) was used as shortcuts to online services. Henze et al. [7] go further by showing how the camera can be used not only to decode visual tags, but also as a tool to create visual (photo based) tags in the real world. Although interesting, these systems only focus on linking a physical device to online services without exploring navigation per se.

Figure 1: Visual Contextual Bookmarks [6].

Rukzio et al. [13] explored different tag based services but focused mostly on how tags can be used to enable phone-based discovery and access to online services. Using NFC tags embedded inside public posters, a user can walk up to a poster, visually identify the NFC enabled functions as landing spots on the poster and activate the services by touching their NFC enabled phone to the landing spots.

More relevant to our work, Serra, Caboni and Marotto [14] showed the use of 2d barcodes as part of an initial indoor navigation system, but they used barcodes only as URL addresses to download maps of an indoor space onto a mobile phone.

 ISBN: 978-1-61208-164-9

2) Indoor Navigation

Nokia Research [9] ran a public trial of their Locate Sensor system in the Kamppi shopping center in Helsinki in 2009. The system enables mobiles phones to track and present the location of special tags on the phone's screen. In this case the use was mostly for advertising- enabling a person to look for a specific store in the shopping center and receive promotional coupons relevant to their location. But their system relied on the use of special purpose hardware.

Puikkonen et al. [12] tested a WIFI based indoor navigation system in the Kamppi Mall also. Their system showed the location of the user on the map, but at very low precision (resolution was about 50 meters- meaning that it could localize a person to the level of a section of the Mall and no more). Although such services show future promise, the low resolution exhibited makes them of limited use for most cases.

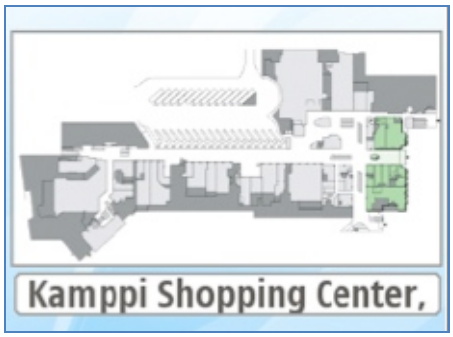


Figure 2: Kamppi Indoor Shopping Service [11].

Mulloni et. al. [8] developed and tested the Signpost indoor navigation system. After implementation they focused on studying end user's feedback about the ease and usefulness of the system. They compared the user ratings for a simple 2D map view with no location information in it, a more advanced model using discrete point localization (using 2D barcodes) that shows the location of the last seen marker and a third and most sophisticated model using a simulation of a real time indoor GPS (using a wizard of Oz approach where an operator would walk behind the subjects and update the location on the application). They found that discrete localization was seen as easier to use and more useful than the simple map approach, whereas the real time indoor GPS simulation was deemed most useful. Since it is not practical to have an operator walk behind every user, the most relevant result of their study in our opinion is that they show that indoor navigation using 2D barcodes is a realistic and useful approach.

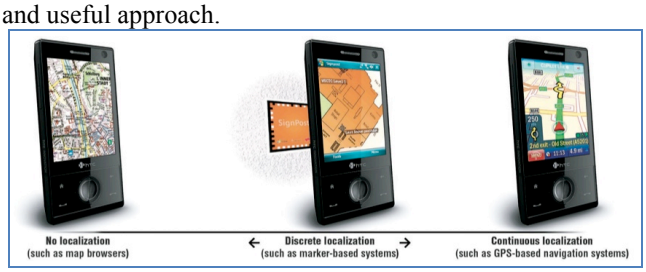


Figure 3: Signpost Indoor Navigation System [8].

Another point of interest in their system was that they kept the camera on at all times and had it continuously search for tags in the environment, automatically updating the location on the map when a marker was seen. Although this seems promising from a user interface point of view, they admit that such an approach was deemed impractical in their real world tests since it consumed too much power and depleted the battery too fast. Additionally, although their system could show the last seen marker location on the map, their system did not generate navigation paths to a target location for the user.

Nokia research recently [10] showed a prototype of their Indoor Navigator system using the Nokia High Accuracy Indoor Positioning system. In demos presented at the Nokia World Conference in 2010 they seem to show very high precision, but we have not seen any research presenting their capabilities systematically. Although simpler to install and maintain than previously mentioned high precision systems, it still needs the installation of special purpose positioning equipment.

III. AMAZON-ON_EARTH LIBRARY NAVIGATOR

Our Amazon on Earth Library Navigator project explores a method of enabling map-based navigation inside a physical space, but with the added value of being able to show the user their last known or current location as well as being able to generate a navigation path to a wanted target, *all without the need for any specialized or expensive localization hardware.* Our service also offers pre and post object finding services. This project is a continuation and improvement of our previous work [5].

A. Description

Our project focuses on enabling a person with the following main capabilities:

- Search for information about an object they are interested in
- Physically find that object in an indoor space via a navigation path drawn on a map
- Receive recommendations about relevant alternative objects
- Pick-up-n-Go: Pick the object up, purchase the object, and carry it out of a store.

1) System Test Location

We implemented the system in one of the libraries on our campus. The reason for this was proximity and ease of access, and should not be taken to mean that we are focusing only on libraries. The opposite is true - a library to us is a representation of a physical retail store. Such a store has stock (the books), a physical space to view the stock and handle it (the book cases and desks), and a checkout counter where people can buy (borrow) the books. To us, such a system is conceptually equivalent to retail stores, while allowing us to explore and test flows and methods without the obvious difficulties involved in using a real store location.

 ISBN: 978-1-61208-164-9

2) Scenarios

The AoE Library Navigator system enables our users to perform the following scenarios:

a) Finding Information about a book

Our user is looking for a specific book they need for their work. They go to our web site and run a search for the book (using key words, author names or ISBN number). They receive an information page about the book and can browse the information that has been gathered from the Google Books and Amazon web sites using their public application programming interfaces (API's). Information about a book can also be accessed by scanning the barcode on the book with our application. This process returns the ISBN number that is then fed into our web-based query system, returning the same information page.

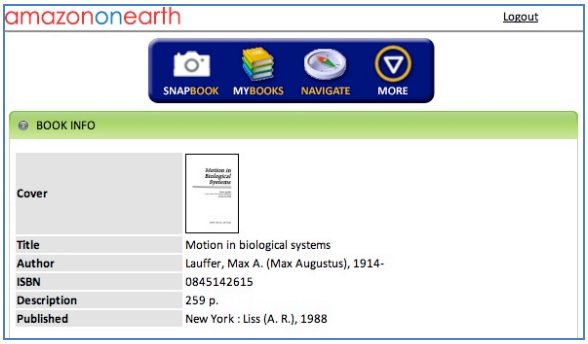


Figure 4: AoE Library Navigator Web Site Book Information Screen

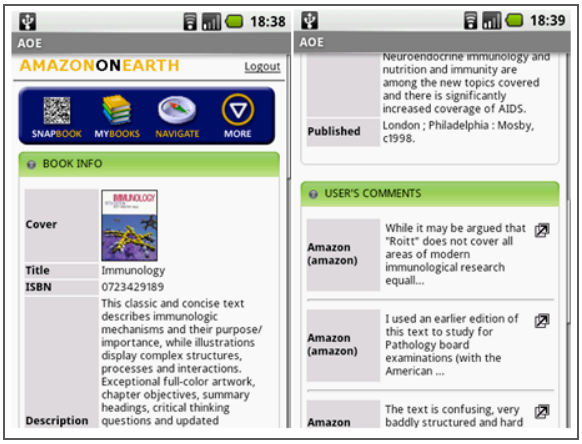


Figure 5: AoE Library Navigator Phone Book Information Screen

Figure 6: AoE Library Navigator: Tag Based Information Search: Scanning the book barcode initiates a web based search

b) Adding a book to their personal list

If they are interested in the book, they can enter it into their book list after signing in to the system. They can now go to the "store" to view the book, and check it out.

c) Navigating to the book in the library

Once they arrive at the library, they launch the AoE Library Navigator mobile application and select the book they are interested in. This brings up a navigation map that shows them the path they need to take in order to reach the book. If they navigate properly, they will reach the bookcase that holds the book they are looking for. If they get lost, they can walk to one of a number of public and centrally located navigation tags and scan them or alternatively, take any book from the shelves and scan its barcode. This will give them a new map with an updated path to reaching the bookcase.

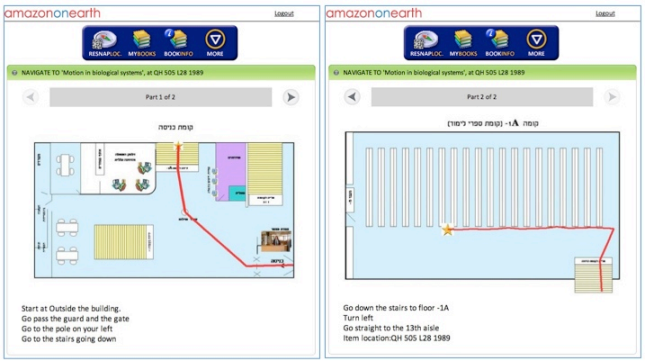


Figure 7: AoE Library Navigator: Navigate Screens

If they find that the book is not there, they can get information about additional books that can be relevant for them. The other books can be scanned using their bar codes and available information can be viewed.

d) Taking the book with them

Lastly, if the user wants to take the book with them, they can select the Check Out option under the book screen and receive feedback that the book has been successfully checked out and that they can take it with them.

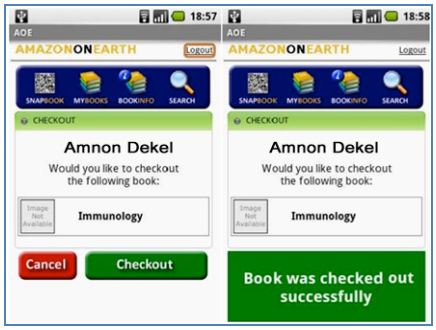


Figure 8: AoE Library Navigator: Check Out Screens

B. Technical Description

Figure 4-18 presents the main Amazon-on-Earth Library Navigator (AoE) system modules:

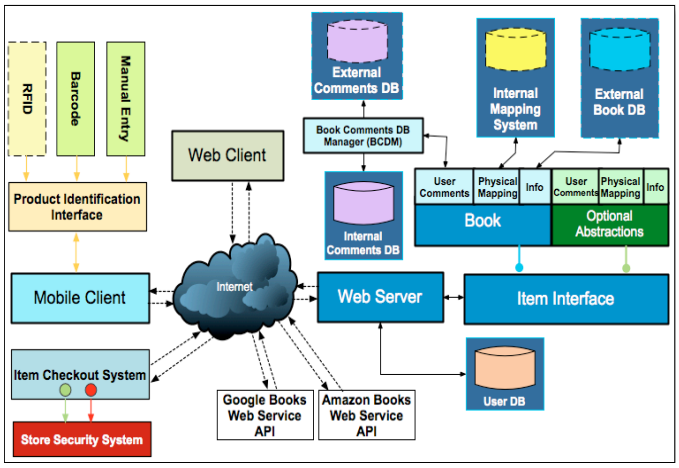


Figure 9: AoE library navigator architecture

1) External databases

- EXTERNAL COMMENTS DB: Uses Google Book information and Amazon's ratings and opinions. Enables us to shows this information for any book in the library.
- EXTERNAL BOOKS DB: Use the library's web site as the main search database. Since the library would not allow us to interface directly to their internal database, we access the book database by sending HTTP queries directly through their public web site.
- MAPPING SYSTEM: A mapping database that interfaces with our Map Middleware and returns results as text (may include links to maps stored on the web).

2) Book Comments Database Manager (BCDM)

An item may have several servers where users' comments are stored. The BCDM layer supplies a convenient abstraction for multiplexing comments from and to several DBs. This allows us to present a single interface where the user can see comments that were created internally by library visitors as well as externally via the Amazon and Google books APIs.

3) Internal comments DB

Since not all DBs are writeable, this DB is used to store local users' comments.

4) Users DB

Stores data regarding users who are eligible to access the system (i.e., User names, passwords, and custom data: "My Books", Preferences and user privileges).

5) Item interface

This is an interface for generalized access to information about an item. The information is separated into 3 sub-categories:

- INFO: General information for identifying and describing the specific item.
- POSITION: Positioning (global and/or local, absolute and relative to a given point)
- COMMENTS: Getting/adding comments capabilities (might use the BCDM interface if many sources for comments are available).

6) Mobile Client Application

The mobile part of the service is enabled via a native Android application. The application is responsible for:

- Preserving the internal state of the user's requests
- Initiating requests to the server(s) based upon requests.
- Parsing data returned from server(s) as response to requests.
- Displaying the incoming data.
- Interacting with Physical World Objects via the Physical World Connectivity Module

7) The Physical World Connectivity (PWC) Module

This module is responsible for acquiring and analyzing information from the world in the following methods:

- 1D/2D barcodes
- Keyboard input
- Optional: RFID and Voice Recognition

Output of this module is unified for all acquisition methods. At this point in time only 1D/2D barcodes and Keyboard input are supported. Adding RFID and Voice recognition is relatively simple. Voice input is already available via the Android input method, but the results will need to be parsed appropriately for our use. RFID input will soon be available via the NFC interfaces that are starting to be used in the newest Android phones (i.e., Nexus S in Dec 2010).

Figure 10: Navigation Tag and Resulting Map

8) Map Middleware System

The system includes a Library Map descriptor and Location Pattern Convertor (what we call the Map Engine). We use the Dijkstra algorithm for finding the shortest path in a directed graph between a start location and the target location. The system converts any possible location string that might be received from an external source (such as web-sites or similar) to one of the targets in the map. In other words, this function connects between any string (from decoded Tags) to the set of strings used as location names/aliases in the map.

We developed a Map Builder desktop application that makes it easy to insert new maps into the system and then connect between a textual target and their symbolic location on the map. The result is a database with relations between a symbolic code and their symbolic location on the represented maps.

 ISBN: 978-1-61208-164-9

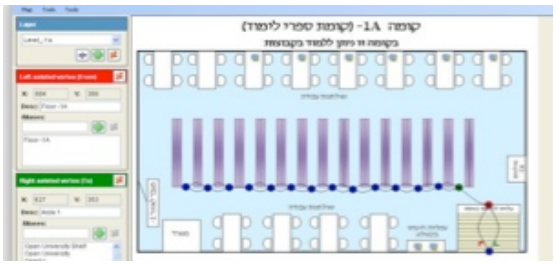


Figure 11: AoE Library map builder

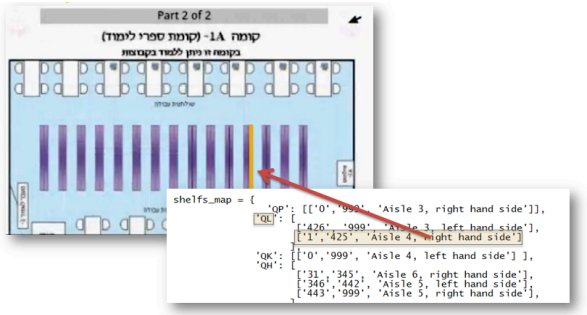


Figure 12: Relating symbolic location to visual map

The Output of the Map Engine is:

- A graphic representation of the Map. Draws a line on top of the map to mark the resulting path.
- An ordered list of Edges. Indicating the x,y coordinates of the starting and ending points relative to the graphics coordinates.
- An ordered list of Strings. These are the textual directions relevant to the layer being shown.

We have created some enhancements to the Dijkstra algorithm. First is **Angle Testing** with built in thresholds to determine if the instructions should be "Go straight", "Turn Right" or "Turn left". We then use **Edge Grouping** to enable us to present simpler instructions to the user: instead of [go straight to shelf 1→go straight to shelf 2 → go straight to shelf 3] we get [go straight to shelf 3].

We initially developed a Java prototype on a Sony Feature-phone but later changed to the Android platform.

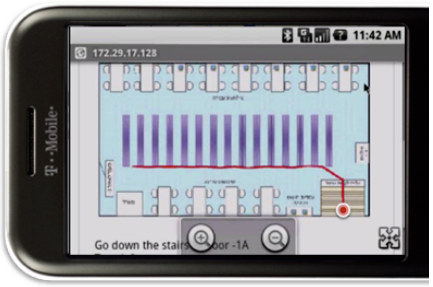


Figure 13: Android based Aoe library navigator

The navigation screen presents the navigation broken up into floor-sized sections. So if a path necessitates moving through more than one floor, the navigation on the first part will steer the user to the stairs and tell them to travel to the additional floor. The second part of the navigation will continue from there. This allows us to break the navigation into easier to understand paths for the user and also leaves more screen real estate free for showing the largest graphic possible.

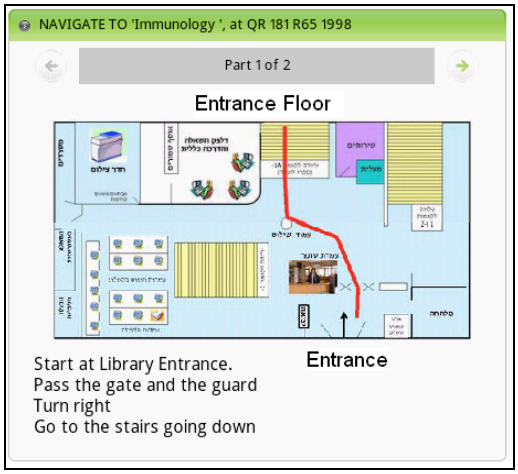


Figure 14: AoE library navigator: navigation map to the target book

C. *Implementation*

We implemented the system using a mixture of web based and native mobile technologies. The server modules were written in Python, and hosted within an Apache HTTP server. The mapping application was implemented in .NET and it created an XML map file for the library, above which the navigation path was drawn at run time with Python. The Databases were implemented in SQL Lite. The check out part of the scenario used HTTP POST to write to a Check out service on the HTTP server. The checkout station used a barcode scanner interfaced to a netbook computer that in turn was connected to an Arduino Microcontroller system for activating the demagnetizer. The check out system worked but because of time and financial constraints was not hooked up to a demagnetizing station- so although a book could be checked on the phone, and the barcode scanning system could check with the service that it was in fact checked out and could be taken out, the book had to be manually demagnetized.

IV. USER STUDY

A. *Method*

We sent out a call for subjects and recruited students from departments at the University that are located on a separate campus at the other side of the city. This was to ensure that they were not acquainted with the physical premises of the library. Subjects were invited at 30-minute increments and were tested alone.

Subjects first read a general explanation about the test procedure in which it was made clear to them that there is no correct or incorrect performance. We asked them to act as naturally as possible and that in the performance of the tasks they were given they could ask for help from anyone in the building except for the tester.

1) *The Tasks*

We used a between-groups design. Both groups were asked to search for information about a specific book and then navigate to and find the physical book in the library.

CONTROL GROUP: This group used the existing IT infrastructure in the library (a number of computer stations using a well known library management and search system). The subjects walked over to a station and ran a search for the book. They then wrote down the library code for the book. This was the first variable we measured: *time to find information* about the book. We then asked them to find the physical book. They were told that they could ask for help from the librarians or other people in the library. This was the second variable we measured: *time to find the physical book* in the library. We also recorded how many navigation *errors* they made and how many times they asked for *assistance* in the process of finding the book. After they found the book we asked them a number of qualitative questions to gauge how they felt about the system and their experience of using it. We then explained the study to them. Ten subjects were in this group, 5 of them male and 5 female.

EXPERIMENTAL GROUP: This group was given the same tasks using our system. The initial information lookup about the book was done on a laptop using the web site we created for the system. After they found the information about the book they were asked to add it to their book list on the system (My Books). This was the parallel step of searching for and writing the book code information by the control group. They were then given a mobile phone (an HTC Nexus 1) with the AoE Library Navigator application running on it and were asked to open the book information in the application and tap the Navigate button. This brought up the Navigation map to the book. We explained to them how to use the application and then told them that they now needed to find the physical book in the library using the map. We explained that they could ask for help from anyone in the library (except for the tester) and also showed him or her how to use the built in feature to scan strategically placed tags or books in the library if they got lost. They then started the search for the book and we timed how long it took them to find it, how many navigation errors they made and how many times they asked for assistance. After they found the book we asked them a number of qualitative questions to gauge how they felt about the system and their experience of using it. We then explained the study to them. Ten subjects were in this group, 5 of them male and 5 female.

B. Results

Figure 15 shows the results of the test. As can be seen, the two main metrics (time to find information about the book and time to navigate to and find the physical book) show a clear advantage for the experimental group. The average time to find information about the book was 77 seconds for the control group but only 51 seconds for the experimental group (35% faster) ($T=1.93$, $P<0.05$). The control group then took an average of 287 seconds to find the physical book, while the experimental group took only 138 seconds on average (149 seconds faster- 52% faster) ($T=5.29$, $P<0.01$).

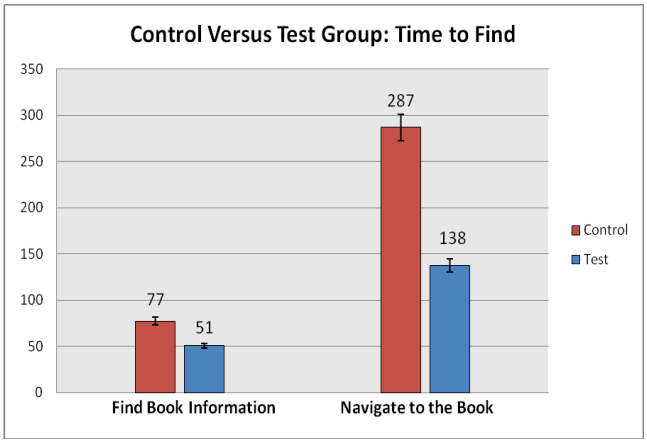


Figure 15: Results of Second Navigation Test (in seconds).

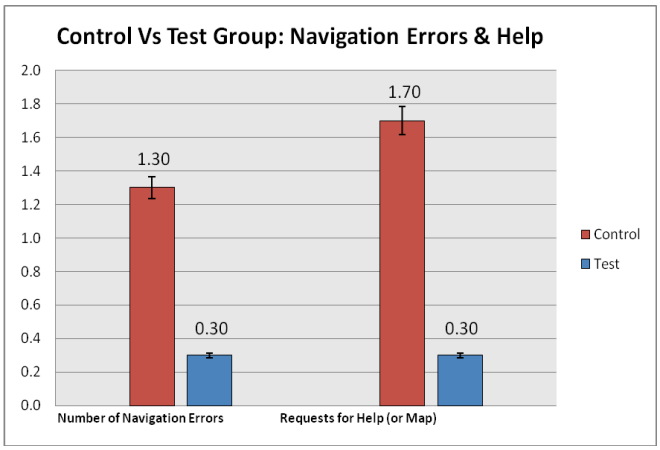


Figure 16: Navigation Errors and Requests for Help.

We also analyzed the amount of navigation errors that were made by the subjects and how many times they requested help in finding the book. We define a navigation error as a situation where instead of getting closer to the target the user's navigation at a certain point in time enlarges the distance from the book. By Assistance we mean asking for help from another person or the librarian or by scanning barcodes to request a new map. As can be seen in Figure 16, once again the experimental group showed an advantage. Whereas the experimental group showed an average of 0.30 navigational errors per task, the control group made an average of more than 4 times as many navigation errors (1.30) ($T=3.87$, $P<0.01$). The control group also needed much more assistance – they asked on average 1.70 times for assistance per task, while the experimental group asked for assistance only 0.30 times on average per task (more than 5 times as much). ($T=6.33$, $P<0.01$).

Figure 17 shows the qualitative questionnaire results.

Ease of searching for information about a book: Both groups rated the ease of searching for information about a book as "Easy" (Control = 4.6, Experimental group = 4.5).

 ISBN: 978-1-61208-164-9

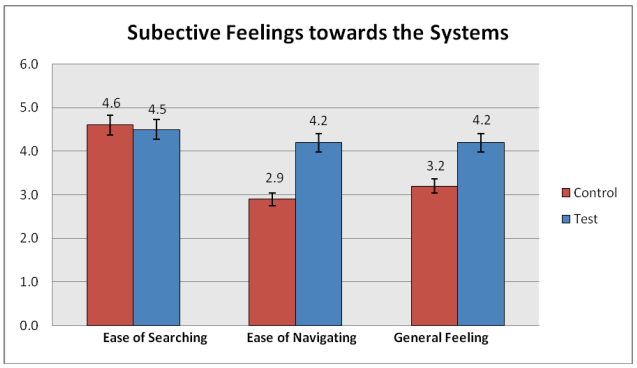


Figure 17: Subjective Feelings towards the systems.

Ease of finding the book in the library: Not surprisingly, the experimental group showed an advantage here- giving on average the rating of "Easy" (4.2) while the control group viewed this on average as neutral (2.9). ($T=4.99$, $P<0.01$).

Feelings towards the method: We found a difference in how positive or negative the subject's feelings were towards the methods - The experimental group showed positive feeling towards the system (average of 4.2) while the control subjects showed a neutral feeling towards the system (average of 3.2). ($T=2.23$, $P<0.05$).

V. GENERAL DISCUSSION

The results show a very clear advantage for using the system. Not only is the time to find a book drastically lowered (by 52%), but also the time to find *information about* a book was lowered by 35%. Additionally, the amount of navigation errors made while looking for the books was lowered substantially, and probably because of that, the need to ask for assistance also dropped substantially. Only one of the experimental subjects made a real navigation error (they went down the wrong set of stairs and found themselves in a different part of the library). That subject then used the navigation barcodes on one of the walls to get a new map generated and was able to reorient themselves and continue the search for the book.

Interestingly, even though the empirical evidence shows that using the system for finding information about a book is faster with our service, on average, the subjects rated the regular library search system to be as easy to use as the experimental system. We think that this has to do more with familiarity than with usability. Most users know and have had experience using the in-library catalog search systems, while none of them have used our system before. This novelty might cause them to think that the existing system is easier to use even while the data shows that it might not be. It seems that a 35% reduction in time is not pronounced enough to be felt by the users as large enough to warrant them to feel that it is better than something they already know and feel comfortable with. For that to happen they must see a much more pronounce performance change.

That is precisely what we think has happened when subjects rated how easy it was to find the book. As shown in the results, the experimental group felt that it was easier to find the book using our system relative to the control group. So in this case, although they were used to searching for books using the standard catalog stations and then physically search using the book coding system, the performance enhancement offered by our system was pronounced enough to make that crossover- and even though it was not familiar to them, they rated it higher than the control group. This "added value" that they received from the new system seems to have caused them to show a general positive feeling towards it, while the control group showed a neutral feeling towards the system they know so well. We see this as meaning that if your system offers enough value to the end user, that added value can overcome their built in bias towards using something they are familiar with.

A. *Future Improvements*

While building this service and testing it we have found a number of things that need to be improved before such a service can be used in the real world.

1) *User Interface Design*

The interface design we have used can and should be improved.

1. We found that users found it initially difficult to understand the map and it was not clear to them where the path starts. This can easily be solved by adding a START HERE tag and placing it in the proper place on the map.
2. We found that users sometimes missed seeing the actual book library code on the map screen- this is important information since they need it in order to identify the actual book on the shelf. This information should be made more pronounced and visible on the screen.
3. The current architecture breaks up the multi- floor navigation into multiple screens. This means that when the user moves from one floor to the next, they need to request the next map from the server by tapping the next button. In hindsight we think it will be better to download all parts of the navigation into one screen and have the user scroll through the screen to reach the later parts of the navigation path. This also ensures that all the data has been downloaded and cached on the phone at the beginning of the navigation and potential WIFI or Cellular Data weak spots will not disrupt the navigation.

2) *Check Out System*

Because of time and financial constraints we did not finish the physical demagnetizer station. We built a physical and working proof of concept but did not add it to the final testing scenario. A system without this feature will still be useful, but allowing the full cycle from search, find, pickup and go will make this an even better system.

3) *Enhanced Social Input:*

Although the current system can access and show comments about a book from Amazon and Google books, its built in social features are limited to allowing a user to add

their own comments about a book into the system. These comments stay in the system and are not published to the Amazon or Google books systems. We think that this service will be enhanced if its users can add and publish information about a book. Thus, users should be able to tweet or publish to their Facebook pages that they have checked a book out, what they think about a book, and also create socially based lists that can help others. This data can then be used as a crowd-sourced collaborative filtering service that can offer an alternative book to the one being searched.

VI. SUMMARY AND CONCLUSIONS

The Amazon on Earth Library Navigator is the first system we have built in our efforts to enhance indoor navigation. We have developed a system that includes a back end and middleware service to map an existing physical space and the locations of objects in it. Working at the Harman library, we were able to create a system that allows a user to search for information about a book, add that book to their book list, and then inside the library, use our service to receive a personal map showing a navigation map to the book. If the user gets disoriented during their walk to the book, they can scan preconfigured navigation tags, or alternatively, use the barcodes on the books themselves, to generate a new map. The new map will show them the path from the current location to the book they are looking for.

We ran a series of tests to explore the utility of the system, and after ironing out some initial problems, we show that the service in fact has promise: Using our service, subjects were able to find information about a book and then navigate to that book substantially faster and with less mistakes than using the existing method of doing this in the library. We found that while users did not feel that the service was easier to use for finding information about a book, they did feel that it made finding the physical book in the library an easier task.

All this makes us conclude that such a service does in fact offer real value for people in such situations. If the system makes it easier and faster to find a book, then the time saved can be used for other purposes- if to find additional books, or to have more time to think about what books are relevant to find. But we view this prototype as representing a more general model in which such services can help people find things within indoor spaces. These things can be books, but they can be merchandise in commercial settings (i.e., products in a store) or objects inside large warehouses. Additionally, such a service can also be used by people moving through unfamiliar surroundings: this can be a new worker in a large building- the Pentagon for example, or an emergency services worker needing to find someone quickly in an unknown building. Many additional use cases can be thought of where such a system can be useful

A. Limitations

Such a system has some very clear limitations. The first one is that the system can only work as long as the target objects being looked for have a known location and they do not move. If an object is moved and its location is not updated in the system, then the utility of the system breaks down, since the map generated will send a person to the last known location of the object. Another limitation is that the system does not provide the user with a real time signifier of his or her own location within the building. Being able to do so will help the end user orient themselves within the space, just as a car GPS system shows the location of the car relative to the path being taken. These issues are dealt with in another project.

REFERENCES

[1] Balakrishnan, H., Baliga, R., Curtis, D., Goraczko, M., Miu, A., Priyantha, N., Smith, A., Steele, K., Teller, S., and Wang, K.. Lessons from Developing and Deploying the Cricket Indoor Location System. November 2003. Retrieved October 1 2011 from http://cricket.csail.mit.edu/

[2] Bandara, U., Hasegawa, M., Inoue, M., Morikawa, H., and Aoyama, T.: Design and implementation of a bluetooth signal strength based location sensing system. In: Proc. of IEEE Radio and Wireless Conference (RAWCON 2004), Atlanta, U.S.A (2004)

[3] Blue Umbrella indoor navigation offers one metre accuracy. Retrieved October 2 2011 from http://news.thewherebusiness.com/content/blue-umbrella-indoor-navigation-offers-one- metre-accuracy

[4] Broll, G., Haarlaender, M., Paolucci, M., Wagner, M., Rukzio, E., and Schmidt, A. Collect & Drop: A Technique for Physical Mobile Interaction. In Advances in Pervasive Computing, Adjunct Proc. of the Int. Conference on Pervasive Computing (Pervasive'08), Austrian Computer Society (OCG), 103-106, 2008. pp 74-81.

[5] Dekel, A, Noach N, and Schiller, B. (2009). Amazon-on-Earth: Wedding Web Based Services with the Real World. MIRW 2009 Sept 15, Bonn, Germany

[6] Eleven Location Based Applications for your phone. Retrieved October 4 2011 from: http://www.leveltendesign.com/blog/colin/11-location-based-applications-your-iphone

[7] N. Henze, R. Reiners, X. Righetti, E. Rukzio, and S. Boll. Services Surround You: Physical-Virtual Linkage with Contextual Bookmarks. The Visual Computer, 2008. pp 847-855.

[8] Mulloni, A., Wagner, D., Barakonyi, D., and Schmalstieg, I., Indoor Positioning and Navigation with Camera Phones. IEEE Pervasive Computing, 2009, 8(2): pp. 22-31.

[9] Nokia Locate Sensor debuts at CES. Retrieved October 1 2011 from: http://conversations.nokia.com/2009/01/12/nokia-locate- sensor-debuts-at-ces/

[10] Nokia Research, 2010: Demo of High Accuracy Indoor Navigation System. Retrieved October 4 2011 from: http://youtu.be/kWMJ_6rQFGY

[11] Papliatseyeu, A., Kotilainen, N., Mayora, O., and Osmani, V. FINDR: Low-Cost Indoor Positioning Using FM Radio. In Mobile wireless Middleware, Operating Systems, And Applications, Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, 2009, Volume 7, 15-2.

[12] Puikkonen, A., Sarjanoja, A., Haveri, M., Huhtala, J., and Häkkilä, J. 2009. Towards designing better maps for indoor navigation: experiences from a case study. In Proceedings of the 8th international Conference on Mobile and Ubiquitous Multimedia (Cambridge, United Kingdom, November 22 - 25, 2009). MUM '09

[13] Rukzio, E., Müller, M., and Hardy, R., Design, Implementation and Evaluation of a Novel Public Display for Pedestrian Navigation: The Rotating Compass. In Proceedings of CHI'09, ACM Press 2006, pp. 113-122.

[14] Serra, A., Carboni, D., and Marotto, V., 2010. Indoor Pedestrian Navigation System Using a Modern Smartphone. In Proceedings of the MobileHCI 2010. pp. 397-398.

[15] Smith, M.A., Davenport, D., Hwa, H., and Turner, T.: Object AURAs: A Mobile Retail and Product Annotation System. In: EC '04: Proc. of the 5th ACM Conf. on Electronic Commerce, ACM Press (2004). pp. 240–241.

 ISBN: 978-1-61208-164-9

On-Demand Service Delivery for Mobile Networks

Fragkiskos Sardis
School of Engineering and Information Sciences
Middlesex University
London, UK
f.sardis@live.mdx.ac.uk

Glenford Mapp
School of Engineering and Information Sciences
Middlesex University
London, UK
g.mapp@mdx.ac.uk

Jonathan Loo
School of Engineering and Information Sciences
Middlesex University
London, UK
j.loo@mdx.ac.uk

***Abstract*—Support for mobility has become a key requirement for computer networks. Users now expect to be connected at any time and from anywhere. At the network level, this will be done using vertical handover techniques across multiple network technologies. However at the service level there is no agreed mechanism by which to support mobile users. Current service delivery techniques that depend on overprovisioning are no longer valid as they are inefficient in terms of network resource management. Furthermore, mobile users now want access to demanding applications such as multimedia services, i.e., iPlayer, YouTube and 3D-TV. These services often have constraints in terms of bandwidth and latency that need to be properly supported in the mobile environment. This paper outlines the challenges involved in the design of a service delivery model for mobile nodes with high Quality of Service requirements. The proposed approach uses service migration techniques that take into account user mobility and network conditions so as to ensure efficient use of network resources. In this paper, we introduce the novel concept of user clustering to help us decide when and where services should be migrated. We also show how this idea can be used to support a video streaming service.**

Keywords- mobile; services; clustering; migration; NMS

I. INTRODUCTION

Technological advancements in recent years have made mobile devices more accessible to a large number of people. Smart-phones and tablets are increasingly becoming more common and users now expect to be connected to the Internet while they are on the move. In addition, these devices now come with multiple network interfaces such as Wi-Fi, High-Speed Downlink Packet Access (HSDPA) and Bluetooth which allow them to have network connectivity at all times. The concept of vertical handovers will allow these devices to stay connected to the Internet as they move from the range of one network technology to another [1]. Transport mechanisms including Mobile IPv6 [2], Stream Control Transport Protocol (SCTP) [3] and Multipath TCP [4] attempt to support ubiquitous connectivity at the network level. However, service delivery issues in mobile environments also need to be addressed.

Legacy networks use an overprovisioning approach to the delivery of services. This approach relies heavily on allocating resources in anticipation of user requirements over long timeframes. Such an approach is unable to adapt, in an appropriate way, to a mobile environment where users are constantly moving around and results in significant waste of network resources. Hence a new approach is needed.

Furthermore, in recent years, the popularity of audio and video streaming, as well as browser applets and HTML5 has made the Internet more multimedia-centric. Multimedia applications have strict temporal and Quality of Service (QoS) requirements that have to be continually supported in this mobile context. Continual service provision for these applications requires that we keep response times to a minimum. 4G technologies such as Long Term Evolution (LTE) offer more bandwidth to the users and higher QoS, which in turn, increase the need for fast service delivery on the server side.

One way to address these problems is by creating on-demand instances of a service. These instances can run on multiple servers and in different geographical locations in order to balance the load and where possible, put a service closer to mobile users. Better load balancing and better QoS can be achieved through this service delivery scheme compared to existing methods.

This paper addresses the issues of Service Migration in the context of delivering services to mobile devices. The concept of User Clustering in which we group users into a cluster is introduced. This cluster is tracked in order to determine when and where a service migration should occur. Bringing these two concepts together allows us to create a service delivery platform capable of creating service resources based on user movement and demand. The rest of the paper is outlined as follows: Section II discusses related work. Section III details the key areas of the problem. In section IV we propose solutions for each area. Section V introduces the test platform we will use to carry out our proposed solution. Section VI demonstrates how we aim to use a working model of this technology to achieve video streaming. The paper concludes at Section VII looking at future aims.

II. RELATED WORK

A service-centric networking platform has been developed at Princeton University. The SCAFFOLD [5] architecture is capable of providing flow-based Anycast with moving service instances. Rather than retaining their addresses as the hosts move, SCAFFOLD allows end-point addresses to change dynamically. This enables hosts to migrate across Layer-2 boundaries. When end-points move,

 ISBN: 978-1-61208-164-9

in-band signaling is performed to update the remote end-points of established flows. Thus, when a service moves, the network automatically directs new requests to the new location. This architecture is aimed at maintaining service availability in the event of server failure but it can also work in our context.

The Y-Comm framework [6] is a network architecture that supports vertical handover. Y-Comm uses two frameworks: the first is the Peripheral framework that manages different functions on the mobile terminal. The second is the Core Framework which deals with operations in the core network to support different peripheral networks. As shown in Fig. 1, these frameworks are brought together to represent a future telecommunications environment that supports heterogeneous devices, disparate networking technologies, network operators and service providers. Although, the two frameworks share the first two layers, they diverge in terms of functionality but the corresponding layers interact to provide support for heterogeneous environments. In the context of this paper, we are interested in the service and application environment layers which are used to support services and their delivery to mobile nodes. In Y-Comm, the service platform layer is independent of the underlying networks but is used to facilitate service level handovers in an integrated fashion.

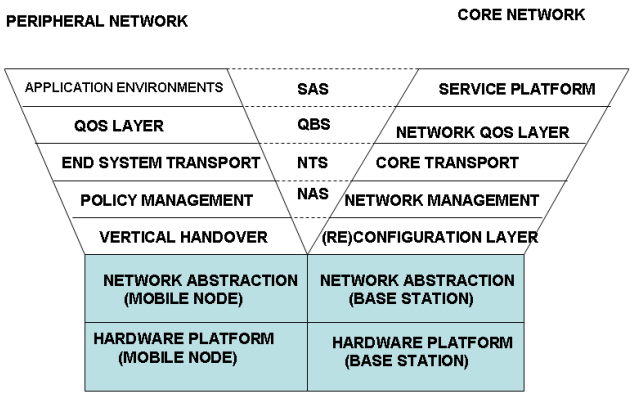


Figure 1. The Y-Comm Framework

III. PROBLEM ANALYSIS

Existing service delivery technologies rely on providing a fixed amount of resources that is greater than the expected service requests within a timeframe. In a scenario where the requests exceed the capacity of the service, some users are not served or in the worst-case scenario, the entire service fails. This does not happen frequently but there are examples due to flash crowds or Denial-of-Service (DoS) attacks. The continuously increasing popularity of always-online mobile devices results in a higher number of constantly connected users. Moreover, as mobile devices gain more processing power, users are able to multitask in the form of connecting to multiple services concurrently i.e. weather updates, video streaming, social networking and file sharing. Hence, a more efficient resource management scheme should be used in the future.

Using a proactive service model allows us to create resources on-demand and gives us the ability to move services closer to users, thus improving QoS and decreasing routing costs. The key challenge is the integration of service migration strategies with models of user mobility. Service migration solutions should also be aware of the network status and available resources before attempting to move a service. Therefore, an efficient service migration strategy must take into account network and server resources as well as user mobility in a scalable fashion.

In terms of user mobility, it is necessary to be able to track the movements and service usage patterns of mobile nodes. To achieve this, we need to be able to track a node's location. In order to efficiently allocate service resources to a geographical area, we need to group similar users into a cluster. The cluster size and its movement will define when service migration is desirable. The reason we are clustering similar users together is so that we can attach a service instance to the group and use it as a mechanism that triggers service migration. An added advantage of this method is that we can treat user clusters as Multicast groups for services such as Internet Television and Internet Radio.

There are also network factors that determine when service migration should take place. Latency, congestion and bandwidth are some of the metric that we should consider before a service is moved between locations. QoS requirements for each service are also an important factor that will determine when migration is desirable and beneficial. Similarly to monitoring users, we need to be able to monitor these resources in real-time. For example, if the network reports higher congestion at a service migration point, there will be little benefit in moving the service to that point. So network statistics are vital to the system.

Another aspect of the problem is the handover at the service level. This needs to be integrated with the network-level handover in order to support seamless connectivity. In order to replicate a service to a new server, we need to transfer relevant user information and service content. There are two types of service-level handoff occurring in sequence. The first type of handoff is between the instances of the service. This transfers user and service data such as active sessions, user files and data caches. The second handoff is between the clients and the service. It binds the client nodes to the new service instance by registering new service IDs and follows with a network-level handover that reroutes all connections from the old service's IP address to the new one's. We assume that a transparent addressing scheme is used in which devices always have the same IP address, even if they move across heterogeneous networks [7].

Finally, node tracking introduces privacy concerns and there may be users who wish to opt-out from tracking but still want to access services. Furthermore, some services may hold sensitive data and may have security requirements such as encrypted file systems and encrypted connections. The system should consider these factors and services must not migrate to servers that do not support adequate security levels.

 ISBN: 978-1-61208-164-9

IV. SOLUTION APPROACH

A. User Clustering

In order to make correct service decisions, we need to be able to track users as they move and the services to which they are currently subscribed. The concept of user clustering attempts to group similar users together by allocating them into clusters based on their location and their service subscriptions. User tracking can be done either by Global Position System (GPS) capabilities on the mobiles nodes or by GSM antenna tracking. Another possibility is Wi-Fi hotspot tracking as used by Google for Street View. Hotspot tracking is similar to GSM antenna tracking in which the mobile node can estimate its location by scanning for nearby Wi-Fi hotspots and comparing the results to an online database that holds hotspot names and locations. These technologies provide varying accuracy in user position ranging from a few meters to the size of a city block but for the purposes of our system, they all provide enough accuracy.

A user cluster is defined as a group of users subscribing to the same service and sharing the same approximate location. In our initial investigation we are proposing the use of simple parameters. For example, we define a cluster as having a Centre (c) and a Radius (r). Furthermore, for the purposes of scalability, we will limit the maximum Cluster Population (p). A cluster is first created when the first user subscribes to a service at a specific location. When a cluster reaches maximum population, another cluster is created in the area. An example of the clustering mechanism is shown in Fig. 2.

The concept of clustered mobility introduces two kinds of dynamics that need to be tracked: The first type of dynamics involves the collective movement of a cluster as its members move from one location to another. This can be defined by Velocity and Acceleration vectors (u) and (a) respectively. The second dynamic involves users joining and leaving a cluster either by subscribing or unsubscribing from a service or by crossing the boundaries of a cluster. This is defined as Cluster Entropy (E). We should also consider the merging of clusters in cases where two clusters attached to the same service come close together and the sum of their members is less than the Maximum Population (p).

The parameters discussed above need to be tracked in real-time so as to form a correct model of the movement of a cluster. Using that model we are able to predict the probable speed, direction and location of a cluster at a point in the future. Finally, all cluster data such as cluster ID (CID) and others mentioned above will be held in a database and updated in real-time so that the service migration logic can process it.

B. Network Resource Monitoring

By using the data from the clustering database and combining it with real-time network metrics we can create an algorithm that will instruct a service to migrate to an appropriate location. Real-time network metrics will include data such as network congestion, latency, and available bandwidth. Additionally, we need to know of candidate servers that can accept a migrating service. Metrics such as available CPU and storage resources as well as security parameters will be used by the migration logic to decide which server is best suited for the type of service that is being moved.

Network metrics are a good indication of the state of a particular network or subnet. Knowing the available bandwidth and latency will allow the Migration Logic to make correct decisions on where to move a service or if it is worth moving the service at all. For example, if a cluster is about to move to an area where wireless signals are not strong and latency is increased while bandwidth is decreased, the Migration Logic will issue an urgent migration of the service to a server as close as possible to the cluster in order to improve QoS as much as possible. In a scenario where network conditions are good the service will be migrated more casually. Similarly, if a server reports an overloaded status, the Migration Logic may decide to move a service to another location in order to balance the load evenly across servers.

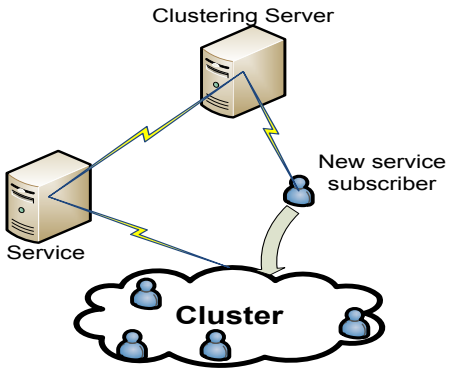


Figure 2: User clustering example.

To gather this data we can use Simple Network Management Protocol (SNMP) and routing protocol information. We can also query participating servers for available resources. This data will be held in a database and updated at frequent intervals. This information allows us to know where best to move a service. In addition to the above, the Network Resources Database (NRD) will also hold Security Level Agreement (SLA) data that define what level of security a participating server offers and whether or not it can accept sensitive services.

C. Migration Logic

The Migration Logic should be able to process cluster mobility data and attempt to predict the future location of a cluster. Eventually the centre of the cluster is going to exceed a distance threshold from the location of the service. At that point, it will instruct a service to replicate itself to a location as close as possible to the predicted location of the centre of the cluster. This will give users lower network latency and will also decrease congestion on a large scale. An additional effect will be decreased costs to the service provider due to decreased packet routing and switching. If the Migration Logic predicts that a service migration is not

 ISBN: 978-1-61208-164-9

going to improve QoS, it will attempt to recalculate the data after a time period. This process will be repeated until the distance threshold is not exceeded anymore, a successful migration occurs or the cluster itself ceases to exist.

If the system fails to predict the future location of a cluster due to erratic movement or other factors, the algorithm will only instruct a service to replicate to a server closer to the cluster's present location. Other parameters that will affect such decisions include the cluster's population, the type of service and whether or not it would be cost effective to replicate an entire service.

When the criteria are met for a replication event, a call packet will be sent by the replication logic to the instance of a service that needs to be replicated. The packet will include a flag for migrating or replicating a service and the address of the target server. If the call is for migration, then the service will make a copy on the target server and delete itself from the initial location. This scenario applies when a cluster moves to a new location as a whole and is demonstrated in Fig. 3 below.

A replication call takes place when a new instance of the service needs to be created without deleting the old instance. In this scenario, the new instance of the service is created with a new service ID and bound to the cluster ID. Replication calls typically apply to load balancing events when more resources are needed. Finally, a kill signal can be sent to a service if it is not needed anymore or a service can be set up so that it shuts down if there are no service requests after a defined time period. If the service is then needed, the requests will be routed to the nearest live service instance and a replication event will occur.

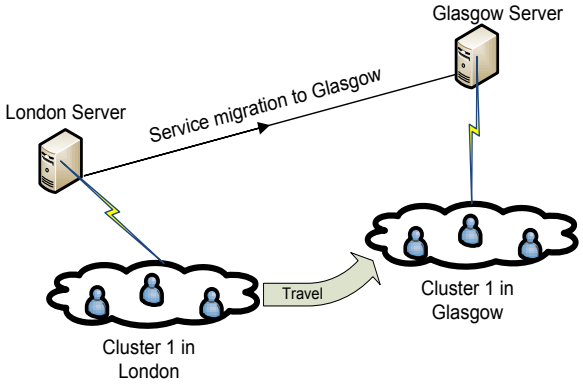


Figure 3: Service migration example.

D. Migration Mechanism

The migration mechanism should be capable of moving a service and its context from the memory of one server to the memory of another server. In addition to moving the service itself, user data should also be moved to the new location. The mechanism should also be capable of launching the service in the new location and terminating the one in the old location. Service level handover should take place once the service has migrated to the new location.

Depending on whether the service in question is stateful or stateless, a handover at the service level may not be needed. Services that rely on stateful user sessions will need to pass session information to the new instances. Otherwise, a stateless service, such as streaming a video, does not need this information passed on, as it is up to the client to request the next block ID of the video. These are called lightweight handovers.

After the service has been migrated, the Cluster Logic can Multicast to all the members of a cluster the new server address. At that point the connection level handover takes place and new connections are initiated while connections to the old service instance are terminated. In order to explore service replication, we will look at replicating instances of the Network Memory Server (NMS) detailed below.

V. NETWORK MEMORY SERVER

A. NMS Features

The Network Memory Server [8] is an example of a simple, stateless service; it stores blocks of data from clients in its memory (RAM). Clients can create, read, write and delete blocks of data. The NMS is primarily made as a storage platform for mobile users. In order to provide support for mobility, the NMS is divided into two parts: The Mobile Memory Cache (MMC) and the Persistent Storage Server (PSS). The MMC initially runs on the same network as the mobile client. If a client moves to another network then the MMC is migrated in order to achieve better performance. The PSS offers permanent data backup for the MMC and there is a level of redundancy implemented so that an MMC can be backed up in multiple instances of the PSS. This is achieved by a multicast call to all the associated PSS. Furthermore, the NMS stores data at the block level in order to provide maximum flexibility in terms of storage and access. The client is responsible for any added-value abstractions, for example, a file-level abstraction. An independent low-level socket interface is used for the server and so the overall interface as seen by applications on client machines is unstructured and can be manipulated as necessary for the needs of the client.

In addition to the above, the MMC provides security to the clients by employing an access rights mechanism over the blocks of data. The owner of a block has full access to it. A user cannot read any blocks that do not belong to them unless given read access to the blocks. Two levels of encryption are also supported in the NMS. A lower level of encryption is used between the client and server as they are on the same network and a higher level of encryption is used between the MMC and the PSS because they are in different networks.

B. NMS Migration

At the moment, the NMS is not capable of migrating from one server to another and only a simple prototype of the PPS is implemented. We will initially explore NMS replication by trying to transfer storage blocks from one MMC instance to another. Because the NMS is stateless, this is a good starting point for our work.

 ISBN: 978-1-61208-164-9

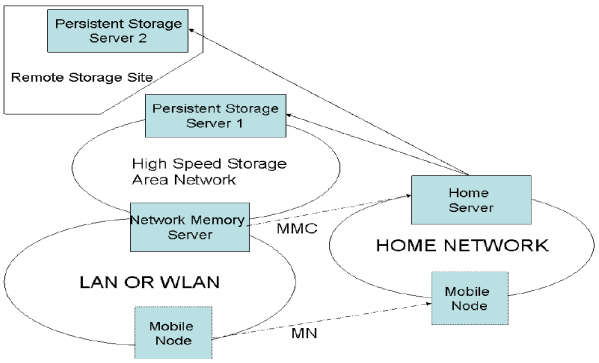


Figure 4: Example of NMS work scenario.

Additionally, we will attempt to implement the PSS in order to have a level of data redundancy during replication. The above will be implemented and tested on a blade server and service migration will be explored across the independent blades. Fig. 4 shows an example of service migration with the NMS using the MMC as the mobile front-end between the WLAN at work and the Home Network. This migration is done using the Context Transfer Protocol (CXTP) [9,10] as shown in Fig. 5 below.

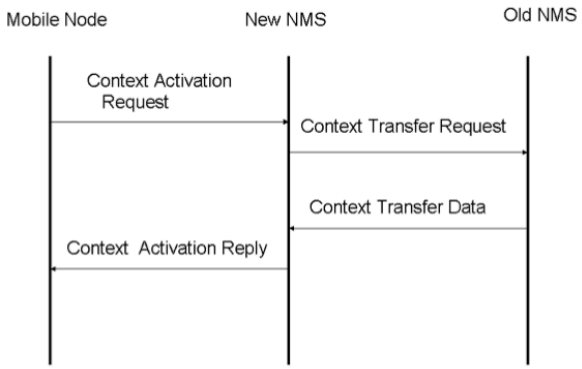


Figure 5: NMS migration model.

VI. MOBILE VIDEO STREAMING SERVICE

The NMS example detailed above is a simple mechanism for network storage delivery and will be the first service we will attempt to migrate. We would like to support a mobile video streaming service in which the streaming service itself, as well as the video file streamed, is migrated across a network. Fig. 6 shows the various components of a framework to support mobile video streaming. The entities and their interactions are described as follows:

The Clustering Mechanism keeps track of User Clusters as they move and the location of the services for each cluster. Once the distance between a user cluster and the service location exceeds a threshold, the clustering mechanism reads the Server Location Database for an available server closer to the location of the cluster. It then sends a signal to the Service Replication mechanism to move the service, suggesting a target migration point.

The Service Replication mechanism queries the target server for available resources and upon a positive response, triggers the replication of the servlet to the new location. Once the servlet is moved to the new location, a signal is sent to the Storage Migration mechanism to copy the relevant video files to an NMS server that is close to the new servlet. The Storage Migration mechanism then uses the CXTP to transfer the cached videos from the old storage server to the new storage server. Once this is completed, it sends a signal back to Service Replication.

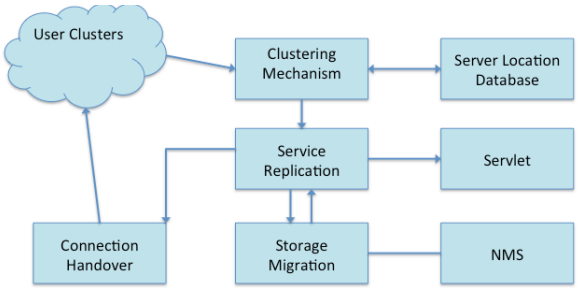


Figure 6: Replication framework for mobile video streaming.

Finally the Service Replication sends a signal to the Connection Handover mechanism, which in turns informs the mobiles nodes within the cluster of the new location of the videos.

In order to make the service efficient and keep it simple, mobile nodes pull the video streams directly from the Storage Server using prefetching algorithms [11]. This means that the Storage Architecture can be stateless, with the relevant state information stored in the mobile nodes. This also means that the actual data is streamed directly between the mobile nodes and the Storage Server.

VII. FUTURE WORK AND CONCLUSION

In order to explore service migration, a test platform will be built here at Middlesex University. We initially aim to use the test platform to explore service migration on the NMS. The first step is to have a working prototype capable of replicating the MMC to a server in the local network. The replication signal will be sent to the MMC manually at first in order to simplify the development process.

The second step is to automate the replication signal using real mobile devices. The initial replication logic will be able to group users into a cluster and track them as they move around on campus. In addition, it will be able to remove users from a cluster if they leave the network or stop accessing the service.

To test this functionality, we are implementing a GSM network on campus that will allow us to use mobile devices to access the NMS. As we move around on campus, the devices will handover between GSM base stations and we will use those signals to trigger service migration between servers. In the long term we are aiming to use NMS service migration in order to achieve video streaming for support of mobile video server applications.

 ISBN: 978-1-61208-164-9

In this paper, we have briefly outlined the challenges presented by user mobility in future networks. Current models of service delivery are inefficient and will not scale to cover the future needs of mobile users. We believe that the combination of User Clustering and Service Migration can bring a better solution to the efficient management of network resources while providing a high quality of experience for users. The authors recognize that there is much to do and would welcome feedback on this paper.

REFERENCES

[1] Mapp, G., Shaikh, F., Aiash, M., Vanni, R., Augusto, M., and Moreira, E. 2009. Exploring efficient imperative handover mechanisms for heterogeneous wireless networks. In: Duresi, Arjan and Barolli, Leonard and Enokido, Tomoya and Uehara, Minoru and Shakshuki, Elhada and Takizawa, Makoto, (eds.) Network-Based Information Systems, 2009. NBIS '09. International Conference. IEEE. ISBN 9781424447466

[2] Johnson, D., Perkins, C., and Arkko, J. RFC 2775 - Mobility Support in IPv6. IETF, June 2004.

[3] Dtewart, R., Xie, Q., Morneault, K., Sharp, C., Schwarzbauer, H., Taylor, T., Rytina, I., Kalla, M., Shang, L., and Paxson, V. RFC 2960 - Stream Control Transmission Protocol. IETF, October 2000.

[4] Ford, A., Raiciu, C., Hadley, M., and Bonaventure, O. RFC 6182 - TCP Extensions for Multipath Operation with Multiple Addresses. IETF, July 2011.

[5] Freedman, J. M., Arye, M., Gopalan, P., Ko, Y. S., Nordström, E., Rexford, J., and Shue D. 2010. Service-Centric Networking with SCAFFOLD. Princeton University, September 2010.

[6] Middlesex University, 2011. Y-Comm Research. [online] Available at http://www.mdx.ac.uk/research/areas/software/ycomm_research.aspx [Accessed: 25 July 2011].

[7] Mapp, G., Aiash, M., Guardia, C. H., and Crawford, J. 2011. Exploring Multi-homing Issues in Heterogeneous Environments. In: Proceedings of the 1st International Workshop on Protocols and Applications with Multi-homing Support in Biopolis, Singapore 22nd- 25th March 2011.

[8] Mapp, G., Thakker, D., and Silcott, D., 2007. The design of a storage architecture for mobile heterogeneous devices. In: Networking and Services, 2007. ICNS. Third International Conference on. IEEE Computer society. ISBN 0769528589

[9] Loughney, J., Nakhjiri, M., Perkins, C., and Koodli, R. RFC 4067 - Context Transfer Protocol (CXTP), IETF, July 2005.

[10] Patanapongpibul, B. L., Mapp, G., and Hopper, A. 2006. An End-System Approach to Mobility Management for 4G Networks and its Application to Thin-Client Computing. ACM SIGMOBILE Mobile Computing and Communications Review, ACM, July 2006.

[11] Thakker, D. N. Prefetching and clustering techniques for network based storage, School of Engineering and Information Sciences, Middlesex University, PhD thesis, May 2010.

 ISBN: 978-1-61208-164-9

Building the Bridge Towards an Open Electronic Wallet on NFC Smartphones

Kevin De Kock, Thierry Van Herck, Glenn Ergeerts, Rud Beyers, Frederik Schrooyen, Marc Ceulemans and Luc Wante
Department of Applied Engineering
Artesis University College of Antwerp
Antwerp, Belgium
glenn.ergeerts@artesis.be

***Abstract*—Many recent initiatives indicate an evolution towards an open electronic wallet to perform all sorts of electronic transactions, like for example micropayments, loyalty, and transport ticketing. Furthermore, the smart phone is emerging as an indispensable tool for many, containing more and more personal information. Hence, it seems like the ideal medium for carrying the electronic wallet as well. The fast and intuitive touch-and-go philosophy and the integration in mobile devices, makes Near Field Communication (NFC) the perfect technology for an electronic wallet. However, the complex ecosystem is holding back the world-wide integration of this technology in mobile handsets, resulting in a low market penetration of NFC smartphones. This paper discusses the use of an active NFC Bluetooth sticker as an intermediate step towards an open electronic wallet on NFC smartphones. Two requirements are set: backwards compatibility with an existing DESFire smart-card solution and support for all the different smart phone platforms. The first requirement will be satisfied through the deployment of a DESFire emulator on the Java Card of the NFC Sticker. The second requirement will be fulfilled by using the Smart Card Web Server functionality of the secure element, which will provide a platform independent interaction with the content of the electronic wallet. Finally, the proposed solution was evaluated in terms of user-friendliness, transaction speed, compliance with the imposed requirements and feasibility of success in the current NFC ecosystem.**

Keywords-NFC; eWallet; NFC Sticker; SCWS; Java Card.

I. INTRODUCTION

The giant leaps of progress in the field of microelectronics since the 1970s have created a solid foundation for the smart card technology of today. An intelligent smart card contains both a microprocessor and data storage, which are integrated on a single silicon chip. This allows cryptographic algorithms to be executed in order to protect sensitive data against tampering and other security threats [1].

This high level security environment has opened the door for applications that resolve around electronic payments (e.g., online/offline debit cards or E-purse cards). However, a lot of the commercial payment systems that exist today are geared towards the general transaction of solely e-money; whereas cases exist that ask for a more specific approach.

The Artesis University College of Antwerp in Belgium has started the Tetra EVENT project [2], which focuses on the development of an open electronic wallet for the event sector. This project will use an online/offline hybrid payment system and several types of items can be stored on the wallet (e.g., vouchers, tickets, coupons, credits, loyalty).

The wallet itself resides on a passive DESFire tag and relies upon terminals equipped with Near Field Communication (NFC) technology to initiate the actual transactions. One of the main advantages in comparison to other existing electronic wallet systems is its inherently scalability aspect in function of big events, because of its online/offline hybrid system. Another advantage is its open character, which allows other 3rd party modules to co-exist on the same physical wallet.

Even though the current market situation imposes the project to focus on NFC tags, a proof of concept will be carried out to test the feasibility of using mobile phones instead. These devices hold a number of intrinsic advantages over tags such as allowing users to view and interact with the contents of the wallet, whereas tags rely solely on terminals instead, which can be deemed to be a shortcoming.

Unfortunately, there is still a low market penetration of NFC enabled mobile phones [3]. A temporary solution to this problem is the use of NFC stickers, which allows any handset to gain NFC functionality through a Bluetooth connection. An extra benefit gained from the use of an NFC sticker is its build-in NFC reader, which allows external DESFire tags to be accessed.

Since current terminals are intended to be compatible with only passive DESFire tags, one of the goals is that the mobile counterpart adopts the current protocols used between terminals and tags. A DESFire emulator will be deployed on a Java Card to ensure backwards compatibility with the currently used system.

Furthermore, there is a lot of differentiation between the various handsets currently available on the market, which means that the wallet implementation needs to be as cross-platform as possible. A Smart Card Web Server (SCWS) servlet [4] will be employed to provide the means of interaction with the contents of the electronic wallet.

Nevertheless, the result of deploying an electronic wallet on a mobile phone could prove to be very useful in the long run. The subsequent sections of this paper will focus on the development of an electronic wallet, which allows the user to interactively work with the contents of the wallet, taking into account the various aspects and hindrances mentioned

 ISBN: 978-1-61208-164-9

before. This paper is structered as follows: the next section goes into details about the research and development done. The third section describes the results and is followed by a concluding section.

II. Research and Development

The research put forth in this paper focuses on carrying out a proof of concept, which will be used as a basis to determine the feasibility of using mobile handsets for the deployment of an open electronic wallet system. The results will be compared with the passive tag version that is currently used in the EVENT project. Furthermore, the project will have to meet certain requirements in order to be accepted as a full-fledged and valid alternative to the use of passive tags. These requirements are the following:

- Finding an intermediate solution to the NFC ecosystem problem, which is currently limiting and/or preventing the further expansion and deployment of NFC services on a larger scale.
- Providing backwards compatibility for the existing passive DESFire Smartcards [5], which are used in the EVENT project for holding the wallet data.
- Allowing the wallet to be used on various different handsets and thus broadening the pool of potential users, by making the system cross platform.
- Taking the aspect of security in account since sensitive data will reside locally on a mobile device.
- Comparing aspects like user friendliness and transaction speed with a passive NFC tag.

A. Project architecture

The software architecture (Figure 1) and hardware used during this project is dictated by the earlier mentioned requirements and can be divided in a number of components. All components are interconnected with each other using either internal or external communication links.

The first component is the interface of the wallet, taking full advantage of the readily available screen and keyboard provided by the mobile device, which is in contrast with the passive tags that rely purely on terminals for the readout of its content.

There is a lot of differentiation however between the mobile devices of different manufacturers (e.g., a variety of different available mobile platforms), which makes creation of a universal interface across platforms currently very hard. A second issue presents itself in terms of secure data storage since the data stored in the memory of the mobile device has some monetary value and the memory itself is intrinsically unsafe (tampering may occur by both the user itself or third parties).

Both the cross platform issue and the secure data storage issue can be solved by the use of a Java Card with a SCWS [6] deployed on it. The Java Card is used as a secure element to prevent any illegal access to the data and an HTML

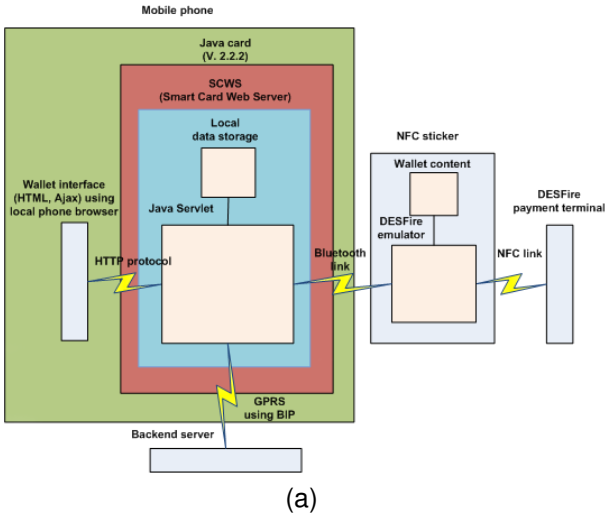


Figure 1. General setup of the system

interface is rendered in the browser of the mobile phone by the SCWS servlet, which will generate the necessary HTML pages. The operation of the Java Card is done by its own local OS and it is thus separated from the general operation of the mobile device OS itself.

The second component of the architecture is the DESFire emulator, which will provide backwards compatibility with the DESFire tags currently used in the EVENT project. This emulator is utilized to link the mobile wallet with the payment terminals through to use of the DESFire protocol. This allows communication links to be setup without having to alter the existing payment terminals.

The third and last component deals with the NFC ecosystem problem, namely the lack of NFC functionality in many currently available mobile devices. We made the decision of using NFC stickers, which are designed specifically for this issue and will provide NFC functionality through a Bluetooth link to a mobile device. The sticker is attached to the back of a mobile phone.

The next sections of this paper will be dedicated towards providing a more in depth discussion regarding the components mentioned before and the actual hardware that has been used.

B. Smart Card & Java Card

The main objective in the use of smartcards is providing the necessary level of security for the storage of sensitive data to an otherwise unsafe environment. This allows applications to be developed around this sensitive data using a subset of the Java programming language [7]. A well-known example is the SIM (Subscriber Identity Module) card, which is used for the identification of a mobile phone user in order to give secure access to the GSM/UMTS network.

 ISBN: 978-1-61208-164-9

Most SIM smartcards are Java Card based these days, which offers the benefit of allowing third-party software to be loaded on the card and executed through a Java Card API.

A high level of security is still maintained on Java Cards by isolating the memory regions of each individual applet from each other through a software firewall, each applet runs in its own context. Furthermore, cryptographic functions can be executed to secure data communication and applets can run in parallel with each other after having been selected by the OS.

However, there are cases where an applet still requires the data and functionality from a foreign applet. A way of acquiring this is through the use of SIO (Shareable Interface Objects) [8][9]. It is important to note that the access mentioned before is only limited in scope. Solely the functions which are defined through the use of one or multiple interfaces are visible outside the applet who grants outside access. Additionally, this access is usually only granted to an outside applet that can identify and authorize itself through the use of its unique AID (Application Identifier).

C. Interface

Mobile handsets offer a very large advantage in terms of user interactivity when compared to passive tags. The former comes equipped with a functional keyboard and screen and grants the possibility of an interactive interface for the mobile electronic wallet. This section will provide the necessary information about the interface part of the project.

The first step in the development of the interface was to compare several types of interfaces with each other and determine which type best suits the basic project requirements mentioned in section II.

The possible interface types are:

- Using an external SATSA MIDlet residing in the memory of the phone itself
- Using SIM Application Toolkit (SAT)
- Using a SCWS servlet residing on the Java Card

Even though SATSA MIDlets [10] offer a lot in terms of the visual interface options and adequate security, these MIDlets lack in terms of the cross platform requirement since they are designed specifically for a J2ME environment. Furthermore, this MIDlet needs to be installed on the handset itself and thus a reinstall is required every time a user decides to switch handsets. Lastly, only very few handsets are currently supporting the JSR 177 API required for using SATSA MIDlets.

The SAT on the other hand offers full interoperability since everything is stored on a SIM card, which is useable by any kind of handset and is furthermore deemed to be very secure. While this may sound tempting to use, the flipside is that it lacks seriously in terms of visual interface options. Moreover, its usability is limited to only SIM cards as SE.

A good compromise is the use of a SCWS installed on a Java Card. A SCWS is a HTTP 1.1 web server embedded on a Java Card and is available since the Java Card 2.2 version, offering a device independent way of the management of personal user data in a secure fashion. This option still puts the application on the SE for security and interoperability purposes. The difference is however that a relatively good HTML interface can be offered by providing static and dynamic content to the browser of a mobile phone through the http://127.0.0.1:3516/ address, which is OS independant.

The last interface type seems to be the most beneficial in terms of the electronic wallet project, because of its advantages in both maintaining a good visual interface as well as the interoperability, security and user friendliness aspects.

The application and data can be additionally managed externally through an Over The Air (OTA) link using web protocols. A secure tunnel will be opened between the SCWS on the Java Card and the OTA platform [11], which is used for the administration of the SCWS.

D. NFC Bluetooth sticker

NFC stickers are contactless cards/tags designed to be glued on the back of a mobile phone and are designed to offer a solution to the current complex NFC ecosystem that prevents a world-wide integration of NFC technology in mobile handsets. A ferrite backing layer prevents distortion to occur between the components of the phone and its radio signal. The sticker also has an internal antenna installed for communication purposes.

An alternative to this is the use of MicroSD cards, which have an embedded chip that will grant NFC functionality to the host device. The antenna itself can be either external or integrated in the package. Additionally, MicroSD cards use similar read/write functions as integrated NFC handsets.

Both have their advantages and disadvantages. An NFC MicroSD card is basically plug and play, thus eliminating difficult setups, but requires the handset to have a MicroSD slot. The NFC sticker on the other hand only needs a Bluetooth radio, which is a very common feature in most of the mobile phones currently produced. The drawback is however that these are less user friendly to setup compared to NFC MicroSD cards.

After carefully weighing both options, we decided to integrate a MyMax NFC sticker from Twinlinx in our project. This will make the project potentially compatible on a much larger variety of mobile handsets compared to the small number of handsets currently available that support a SE on a MicroSD card.

1) NFC sticker characteristics: The sticker has been designed to be as small as possible. A small low voltage battery is used to power the internal Bluetooth chip. This chip is responsible for setting up a connection with the handset and has the capability of making between 300 and

 ISBN: 978-1-61208-164-9

500 connections before the battery is drained. The battery itself can be wirelessly recharged using a specialized USB-charger.

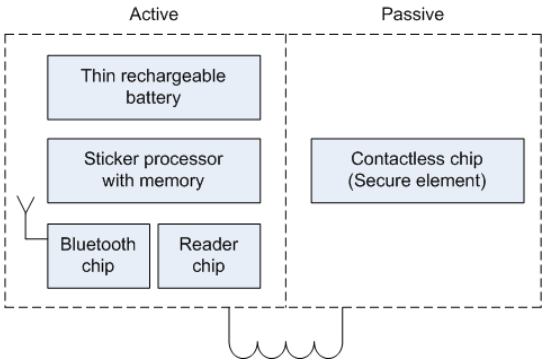


Figure 2. MyMax NFC sticker architecture

The MyMax sticker can act either passively or actively (Figure 2). An active sticker will rely on its own internal power source, while a passive sticker will use an emitted magnetic field from an external reader to draw power from.

Furthermore, the sticker can go into 3 different operation modes; the first mode is a passive mode where the sticker will act as a passive NFC tag. An external reader can be used to read and/or alter the contents of the NFC chip/internal SE on the sticker. It is important to note that a sticker operating in this mode will work completely independent from the handset that it is attached to (i.e., the handset is not required to be powered).

The second mode requires that both the MyMax sticker and its corresponding mobile phone draw power actively from an internal power source. This mode allows a connection to be established between the sticker and the mobile phone. Consequently, the content of the internal SE/NFC chip can be read or changed by the handset through this link.

The third mode takes advantage of the internal reader chip of the sticker, which makes it possible to create a connection to an external tag and allow the mobile phone to read or change the contents of this tag.

2) Testing the capabilities of the sticker: We carried out some preliminary tests with the NFC sticker to determine several key aspects of the sticker that are necessary in the development of the project.

First, we wanted to determine whether the active part and the passive part of the sticker share the same SE. In order to use the active part, we installed the MyMax demo projects on the handset to write new values to the internal NFC tag of the sticker. These values are readable by an external reader, thus proving that the active and passive part of the sticker share the same SE since the reader only reads the passive part.

The second test involved the use of the MyMax SDK to develop a J2ME test project Midlet. The purpose of this project is to read the values of an external tag through the reader chip of the MyMax sticker. DESFire APDU commands are transmitted to the external tag using functions found in the MyMax Library to send APDU.

3) Combining the sticker with the Java Card: Setting up a secure Bluetooth link between the Java Card residing on the handset and the external SE on the sticker is absolutely vital to safeguard the read/write keys that are used in the application for accessing and altering the wallet information from unauthorized use. The purpose of this is to prevent any counterfeiting from occurring.

It is important to note however that the JSR 82 API for Bluetooth is not supported in general by the currently available Java Cards. This obligated us to look for alternative ways to gain this functionality, since the sticker as mentioned before uses Bluetooth technology when accessed from a handset and it is imperative that the overall security of the system is kept at a high level.

A first potential solution for this issue is the use of SATSA (Security And Trust Services APIs) Midlets [12], which are designed specifically for the secure access of Java Card applets. The availability of the JSR 82 API for SATSA Midlets allows data to be passed between the Java Card and NFC sticker in a more indirect fashion.

Two different communication APIs are possible with SATSA Midlets. The first communication API is the SATSA JCRMI (Java Card Remote Method Invocation), which requires a Java Card applet to first extend the java.rmi.Remote interface before any data can be shared with an external application [13][14]. The second communication API is the SATSA APDU, which uses APDU messages to access the on-card objects.

The flipside however in the use of SATSA Midlets is losing the cross platform aspect of the project since SATSA Midlets are J2ME specific, thus narrowing down the number of compatible handsets. Additionally, it turns out that almost no mobile phones support SATSA JCRMI in the first place and only a small number support the SATSA APDU API.

Despite the fact that SATSA Midlets gave good results in terms of programming functionality, they are not viable to be used in our project due to lack of support.

A second potential solution to gain Bluetooth functionality is the use of the BIP (Bearer Independent Protocol) that allows OTA support for the Java Card. While it is theoretically possible to setup a Bluetooth connection using this protocol with the sticker, the lack of publically available practical information for development purposes has prevented us from actually implementing this functionality in our project [15]. The addition of this functionality has been scheduled for future work.

 ISBN: 978-1-61208-164-9

E. DESFire emulator

One of the goals of the project is to maintain backwards compatibility for the existing passive DESFire Smartcards that are used during the EVENT project for holding the wallet data.

The MyMax sticker uses a mifare 1K Classic chip with 1KB of memory, which is a lot smaller compared to its DESFire counterpart that can hold up to 8 KB. The latter can thus hold a lot more wallet data and also uses certain functionality and encryption algorithms (3DES) that the sticker lacks. This results into an inability to port the passive DESFire Smartcard wallet directly to the sticker.

A way to overcome this problem is the deployment of a DESFire emulator on the SE of the sticker to emulate the DESFire wallet functionality. Some additional advantages are its larger sticker SE memory (32KB EEPROM) and it can also be deployed on NFC phones or NFC MicroSD cards.

We will be using a NFC MicroSD card instead of the sticker for testing purposes during this project because of the problems with initiating a secure Bluetooth link from the SE as mentioned earlier.

III. RESULTS

This section will provide some deeper insight regarding the inner workings of the application. This application (Figure 3) exists out of an interface and two Java Card applets, which are responsible for providing the actual wallet data.

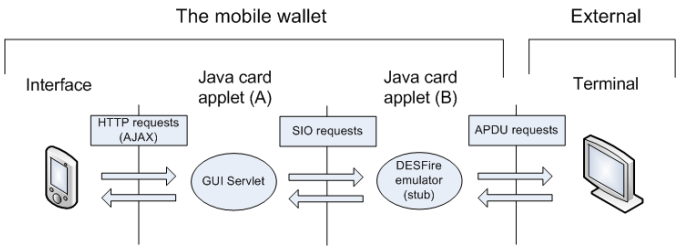


Figure 3. Application components

The content of the event wallet is displayed in the interface on the mobile phone of the user. This HTML based interface can be divided into a number of interface components that are individually requested from a SCWS residing on the Java Card through various HTTP requests.

The servlet that is responsible for providing the interface with the necessary data is running on the SCWS and will send specific SIO requests to the DESFire emulator on the Java Card. This emulator is another applet residing on the Java Card and will (after authentication) provide the servlet with the desired data. The servlet will then, based upon the collected data, give an answer to the previously made request by the interface.

Updates to the wallet content on the emulated DESFire card are done mainly through external points of sale, which are available on either the event itself or through an OTA purchase.

A. Application interface

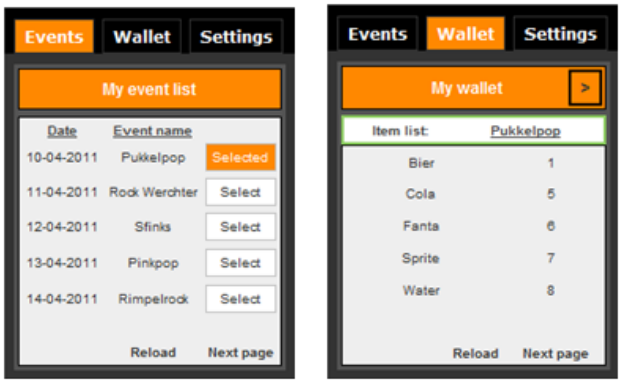
The design of a mobile interface requires some special attention, more so than its workstation counterpart.

First, there is a large differentiation between different cell platforms, each whom presents its own interface. Secondly, physical obstacles such as different screen sizes, aspect ratios and physical buttons need to be taken into account as well. Lastly, there is the user aspect, which demands that an interface needs to be intuitive and easy to learn.

We made the decision to use a SCWS to provide a consistent interface by taking full advantage of the mobile phone browser capabilities, which is tasked for the rendering of an HTML based interface for the wallet application. This interface is theoretically universally applicable to any mobile phone that supports Java Cards. Recent developments like jQuery Mobile and PhoneGap make it possible to build native looking and responsive HTML and javascript-based applications.

The design of the interface of the application itself is based upon known design principles [16], which dictate how information on a page is to be presented to its user. This includes but is not limited to bringing information to the top of the interface by limiting the amount of links a user has to go through thus, minimizing navigation or bringing a collection of relevant information together, based on the desired intent of the user.

1) Interface structure: The wallet interface consists of three main tabs that allow a user to browse through a number of subtabs. The main interface tabs are the ”Events”, ”Wallet” and ”Settings” tab.



(a) The event tab (b) The wallet tab

Figure 4. The interface

The ”My event list” subtab (Figure 4(a)) is used to list all the possible events that a user currently has access to. This

 ISBN: 978-1-61208-164-9

list is kept up to date through a connection with a backend server. A number of possible items are linked to each specific event and are shown in the ”My wallet” subtab (Figure 4(b)). Items can be bought and/or spend using the local terminals or through OTA functionality [11].

The settings tab allows for various options, including setting the number of listed items and/or events on the events and wallet tab pages. The interface is resolution independent and can thus be used on a number of different handsets. The layout can additionally be changed altogether depending on the preferred style of the user.

Finally, the wallet tab includes the option of purchasing items OTA. The purchased items will be listed on the ”My wallet” subtab as ”reserved items”, meaning that they still need to be synchronized by specialized terminals called ”sync points”. The main advantage of this is that wait time for the user at a terminal will be cut down significantly since a user only has to touch the terminal and through a NFC link will the previously purchased items be made available for use.

B. Data flow

The next part of the application consists of two elements, namely the GUI Servlet and the DESFire emulator, which are both tasked with the provision of the actual data to the interface mentioned earlier.

A GUI Servlet is a Java Card application that runs on the SCWS and will act both as a server and a client when data is requested by the interface, since it will serve information to the interface after it has requested the necessary data from the DESFire emulator.

The DESFire emulator is also a Java Card application and consists of an emulated DESFire card with additional communication logic. This emulated card follows the same wallet structure of the passive tags used in the Tetra EVENT project, which is responsible for holding the wallet content. The additional communication logic of the DESFire emulator allows a mobile phone to communicate with the terminals that are used for monetary transactions during an event.

As a result of our project still being in a relative early stage, we have opted to replace the DESFire emulator with a temporary stub in order to fabricate a working prototype. This stub will hold the same functionality from the point of view of the servlet compared to the actual DESFire emulator.

1) Retrieval of data: A backend server forms the backbone of the system, because all the event dates, event names and item names are requested from this server. Providing a page in the interface with values requires the servlet to request a list of AIDs and FIDs first from the DESFire emulator.

These IDs represent the various events and items tied to an event respectively and are required to be translated by the back-end server to their actual corresponding names. The item values shown on the wallet page are collected by using an AID to select a specific event on the DESFire emulator and then request the value of a specific item of that event using its FID.

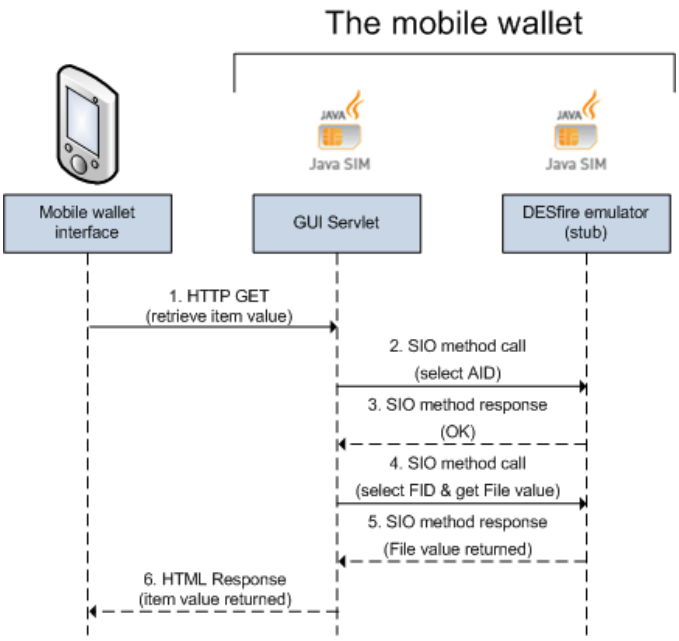


Figure 5. Retrieval of an item value

Every request done by the interface will thus trigger a series of steps in the background of the application (Figure 5). The GUI servlet will perform a SIO method call to select the AID that is linked to the desired event. The emulator will respond with a confirmation message and the actual file value will then be requested next by using the FID of a specific item. The File value is passed back to the servlet, which will pass it to the interface for visualization.

The backend server and its data have been temporary replaced in our project by a number of vectors that store the different names and dates mentioned before in a hardcoded manner. The reason for this is identical to the stub in that it is caused by the fact that the EVENT project is still in a relatively early stage.

C. Installation on a Java Card

This section describes the installation part of the application on an actual Java Card and any problems that occurred in the process [17].

The first step of the deployment on a Java Card is the conversion of the class files of the project, along with any additional required export files, into an executable binary CAP-file (Converted Applet File). This type of file format is designed specifically for Java Cards and is used by an on-chip installer to install the applet and link it with the classes that are already available on the card [18].

 ISBN: 978-1-61208-164-9

1) Used hardware and installation problems: In terms of hardware we used a G&D Sm@rtCafe Expert 5.0 microSD Java Card in combination with the Gemalto Developers Suite for the development of the project. The assumption was made that our project, which is developed for a Java Card version 2.2.2, would run on any physical 2.2.2 Java Card without problems, but this original assumption turned out to be false.

Different vendors use different implementations on their Java Cards [19], which resulted in our case in the use of libraries that are specifically made for Gemalto Java Cards and thus are unavailable on the G&D card that we are using. It should still be theoretically possible to deploy an applet in a cross platform manner by following only the actual global standards put forth by the Global Platform [20].

In order to test this hypothesis, we decided to rewrite the DESFire emulator stub using the Java Card Development Kit V2.2.2 from Sun and were able to successfully install it on the G&D Java Card, thus proving our earlier made point.

IV. Conclusions

The goal of this research was to determine how the functionality of a contactless smart card wallet on a mobile device can be incorporated and improved; using an intermediate step towards NFC enabled devices and taking aspects such as security, usability, backwards compatibility and interoperability into account.

By using a Java Card, a high level of security for safeguarding the sensitive data residing in its memory is maintained. The combination of a Java Card and the SCWS allows the wallet to be deployed on a wide range of mobile devices since no application is required to be installed on the device itself.

The structure of the interface is based upon known design principles to create an intuitive interface for the end-user. Furthermore, the interface has been designed to take aspects such as scalability into account, which is important for overcoming physical obstacles such as different screen sizes of the mobile device.

The current price for a MyMax sticker is 20 EUR, but its price is expected to drop relatively fast, to around 10 EUR, as a result of mass production. The length of the battery life is long enough for it to be used on daily basis. Moreover, it is easy to pair the sticker with a mobile phone.

These factors in combination with our research have proven the MyMax sticker to be a good intermediate solution in terms of the NFC ecosystem problem. However, we were unable to setup a working Bluetooth connection between the Java Card of the mobile handset and the sticker due to the lack of publically available BIP documentation.

The backwards compatibility requirement of the project can be fulfilled through the deployment of a DESFire emulator on the MyMax sticker. Since a working emulator was unavailable during the research phase, we developed a DESFire emulator stub to offer temporary functionality to prove the backwards compatibility of the mobile wallet by installing the stub on the SE of the MyMax sticker.

Our conclusion is that a contactless smart card wallet on a mobile device can be developed, in spite of the current NFC eco system problems, while also taking the fundamental requirements of the project into account.

References

[1] W. Rankl and W. Effing, *Smart Card handbook*. Wiley, 2010.

[2] elab, “Tetra EVENT project,” [accessed 19-September-2011]. [Online]. Available: http://event.e-lab.be

[3] N. Forum, “Whitepaper: Essentials for successful nfc mobile ecosystems,” [accessed 10-July-2011]. [Online]. Available: http://www.nfc-forum.org/resources/white_papers/NFC_Forum_Mobile_NFC_Ecosystem_White_Paper.pdf

[4] SIMalliance, “Whitepaper: Smart card web server, how to bring operators’ applications and services to the mass market,” [accessed 10-July-2011]. [Online]. Available: http://www.simalliance.org/en?t=/documentManager/sfdoc.file.supply&e=UTF-8&i=1185787014303&l=0&s=QcEgCcAUIYrk9KQfX&fileID=1234200160560

[5] B. F. Council, “Mifare desfire specification,” p. 20, 2009.

[6] SIMalliance, “Smart card web server stepping stones,” [accessed 10-July-2011]. [Online]. Available: http://simalliance.org/en?t=/documentManager/sfdoc.file.supply&e=UTF-8&i=1185787014303&l=0&s=K5Aqx7C9UCsQoShk&fileID=/en?t=/documentManager/sfdoc.file.supply&e=UTF-8&i=1185787014303&l=0&s=K5Aqx7C9UCsQoShk&fileID=1261058498628

[7] S. M. Inc, “Java card applet developer’s guide,” [accessed 10-July-2011]. [Online]. Available: http://www.oracle.com/technetwork/java/javacard/overview/index.html

[8] M. Montgomery and K. Krishna, “Secure object sharing in java card,” p. 10, 1999.

[9] D. Perovich, L. Rodriguez, and M. Varela, “A simple methodology for secure object sharing,” p. 7, 2000.

[10] gemplus, “Whitepaper: Integrating the sim card into j2me as a security element,” [accessed 10-July-2011]. [Online]. Available: http://whitepapers.zdnet.com/abstract.aspx?docid=175614

[11] T. C. Vilarinho, “Trusted secure service design: Enhancing trust with the future sim-cards,” p. 167, 2009.

[12] K. Mayes and K. Markantonakis, “Smart cards, tokens, security and applications,” p. 416.

[13] S. Chaumette, A. Karray, and D. Sauveron, “Secure collaborative and distributed services in the java card grid platform,” p. 8, 2006.

[14] J. Andronicj and Q.-H. Nguyen, “Certifying an embedded remote method invocation protocol,” p. 8, 2008.

 ISBN: 978-1-61208-164-9

[15] N. Aini, “Joomla authentication using smart card web server,” p. 48, 2008.

[16] K. Holtzblatt, “Customer-centered design for mobile applications,” p. 11, 2005.

[17] Z. Chen, *Technology for Smart Cards: Architecture and Programmer’s Guide*. Prentice Hall, 2000.

[18] Gemalto, “Java card & stk applet development guidelines.” [Online]. Available: http://developer.gemalto.com/fileadmin/contrib/downloads/pdf/Java_Card_STK_Applet_Development_Guidelines.pdf

[19] J.-F. Dhem and N. Feyt, “Hardware and software symbiosis helps smart card evolution,” p. 19, 2001.

[20] GlobalPlatform, “Globalplatform card specification 2.1.1,” [accessed 10-July-2011]. [Online]. Available: http://www.globalplatform.org/specificationscard.asp

 ISBN: 978-1-61208-164-9

Situation-based Energy Management System

Seung-Won Lee
Future IT Convergence Lab
LG Electronics Advanced Research Institute
Seoul, Korea
seung.lee@lge.com

Se Heon Choi
Future IT Convergence Lab
LG Electronics Advanced Research Institute
Seoul, Korea
ruben.choi@lge.com

Minkyung Cho
Future IT Convergence Lab
LG Electronics Advanced Research Institute
Seoul, Korea
minkyung.cho@lge.com

Jungsu Lee
Future IT Convergence Lab
LG Electronics Advanced Research Institute
Seoul, Korea
jungsu.lee@lge.com

***Abstract*—Energy related technology such as Smart Grid has become one of the main interests in modern industry. Deciding whether to supply or store energy by predicting the amount of energy consumption rate is the core technology in Smart Gird. In this paper, we present a situation-based prediction (SBP) system that not only utilizes user's usage history data but also makes use of the user's situation that is derived from various sensor data. According to our experiment, situation-based prediction system has about four times better performance compared to the time-based prediction system which solely relies on user's usage history.**

Keywords-situation; context; energy; Bayesian Networks

I. INTRODUCTION

A large number of researches have been done on energy management system. At the initial stage, most of the systems were based on simple rule-based system and have evolved to use some contexts (e.g., location) to predict whether more power will be required or not [1][2].

However, most of these context-aware prediction systems are based on simple sensor data and past usage history. Past usage history is not sufficient to detect unordinary conditions and sensor data can only distinguish simple circumstance unless we draw a meaningful significance from those sensor data.

We anticipate that the next stage of energy management system is to understand the gathered contexts and derives meanings from them. In our system, we call it a situation. One possible way to analyze and derive situation from contexts is to use probabilistic network such as Bayesian networks (BNs), also called belief networks, Bayesian belief networks, which are widely used to model uncertain and complex domains [3]. Our situation-based prediction (SBP) system is based on this Bayesian Network, where a cell phone acting as a main device loaded with BN. Cell phone gathers contexts from itself along with other electronic devices at home or office (e.g., TV, light, PC) and uses these contexts as evidence for running Bayesian Network [4]. The main idea of situation based prediction system is to keep track of user's energy consumption rate based on user's situation. We believe recording and analyzing situation can eventually tell user's life pattern and adjust the energy management system to specific user or group of users.

This paper is about our research and development of situation-based prediction algorithm for energy management system and its potential benefit compared to other energy prediction systems.

The organization is as follows: Section 2 starts with the definition of the term "Situation". Section 3 explains the process of gathering and utilizing the usage history data associated with situation. Section 4 covers retrieval and application of feedback data. Section 5 gives a short description of the experiment, and Section 6 presents the advantage of situation-based prediction system. Lastly, we present our conclusion along with possible future works.

II. DEFINITION OF SITUATION

We define situation as interpretation of various sensor data. Sensor can be any device that can give information (context). For example, heater and air conditioner can tell us the temperature of that area. Motion detector, light sensor can present the location of the users in that area. Situation is derived by analyzing these raw sensor data.

One sample situation can be "Sleeping at room". This situation can be derived by using time, light sensor, and other electronic equipments in room. If light sensor and other electronic equipments are off and the time is late at night, we can assume that the user is sleeping or about to get to sleep. In our system, we classified 5 different situations; Sleeping, Watching TV, Eating, Working, and Resting. In addition, the system detects user movement to decide whether to derive the situation or not, since all 5 situations we classified occur while user staying in one position. The sensors we have used to derive situation were personal cell phones, TVs, lights, and personal computers.

III. RECORDING AND PREDICTING ENERGY CONSUMPTION RATE

SBP records user's energy consumption rate every two hours. However, if the system detects that user is moving, it

 ISBN: 978-1-61208-164-9

does not react until the user's movement is finished. We defined moving as a movement from one isolated place to another isolated place. For instance, moving from a living room to a dining room would be considered as moving, whereas a movement within the living room is not. When recording user's energy consumption rate, SBP first derive the current situation in probabilistic matter by running the Bayesian network. A sample result of running the network can be: Sleeping (87%), Watching TV (50%), Eating (10%), Working (15%), and Resting (60%). We only record the highest probable situation which would be "Sleeping" in this case and the amount of consumed energy at that point. In order to derive reliable data, we only record the situation and the consumed energy amount only if the probability is higher than 75%. Once a pair of situation and amount of consumed energy data is collected, we store these data to the database which is composed of time slot, average amount of consumed energy and the number of occurrence. Each situation has its own database table. (Fig. 1)

The number of occurrence data is used to update the average amount of consumed energy and to apply feedback data which will be discussed in Section 5. When an update occurs, we increase the number occurrence by one, and update the average amount of consumed energy. (Fig. 2)

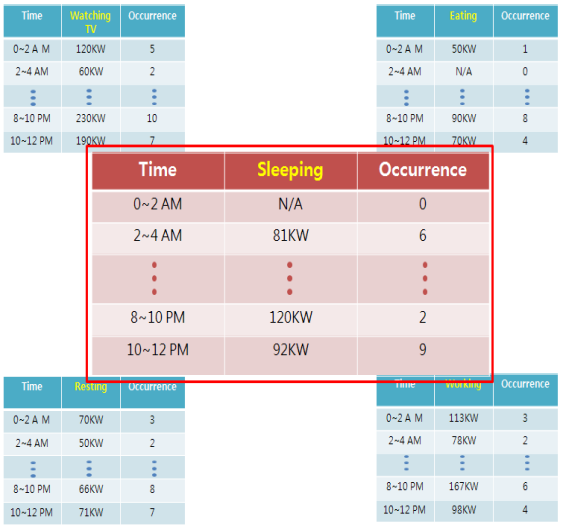


Figure 1. Database table for each situation

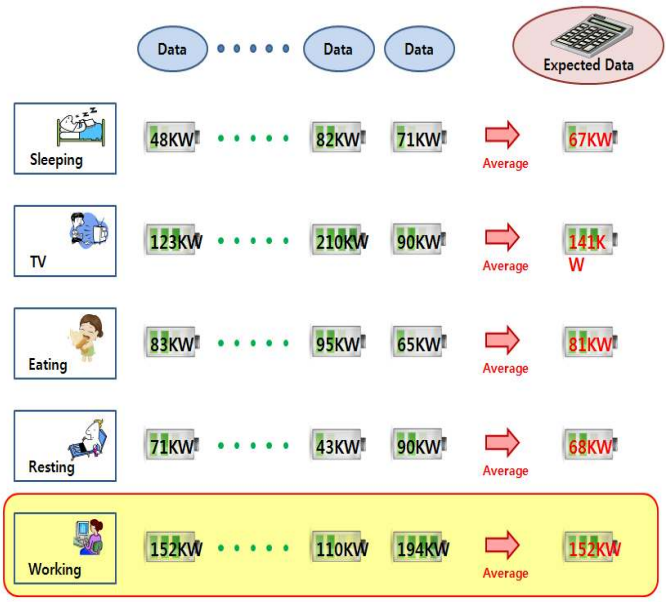


Figure 2. Calculating average amount for each situation

One subtle issue in the system is to decide when to derive a situation. As we mentioned above, the SBP detection occurs every two hours, when no user movement is found. We believe energy consumption rate varies if user moves from place to place. For example, move to dining room to watch TV, move to kitchen to cook, etc. Therefore, if user movement is detected, the system waits until user settles at certain place. However, if the system urgently needs the energy prediction rate, it forecasts the next possible situation by comparing the "Occurrence" field of each situation database. For example, if current time is 2:00 AM, the system retrieves the occurrence field of "2~4 AM" from each database and selects the one that has the highest number. Since the main object of our experiment was to compare situation-based prediction to time-based prediction, the system tries to detect the movement of the user every two hours. We are also considering deriving a situation whenever a user movement is detected regardless of time. In this way, the size of the situation database will decrease by one twelfth but may require more updates if user frequently moves from place to place.

Once the database is filled with enough data, the system utilizes these data to predict user's energy consumption rate. We believe the prediction accuracy increases as the size of the database grows and converges at certain point. Discovering the convergence point would require an experiment in real world. As for now, we gathered a month period data before prediction.

Following are the two steps that SBP uses to predict the amount of energy consumption rate.

1) SBP first runs the Bayesian network and derives the current situation by picking the highest probable situation. Since the probability threshold is set to 75%, SBP does not give its prediction value if the probability is less than 75%.

 ISBN: 978-1-61208-164-9

When a tie occurs on situations, although it rarely happens, we randomly pick one situation.

2) Once a situation is selected, the system retrieves the database of the selected situation, and gets the average amount of consumed energy of the current time slot. For example, if current time is 12:30AM and selected situation is "Sleeping", the system retrieves energy consumption data of "0-2 AM" from "Sleeping" database for its final prediction rate.

IV. APPLYING FEEDBACK DATA

As we mentioned before, SBP is based on probabilistic network such as Bayesian Network and is used to discover user's current situation. In addition to discovering situation, we built a separate database that keeps track of energy consumption rate. Once the system gives its prediction, it observes the actual amount of consumed energy. The difference between the actual amount of consumed energy and the predicted amount is used as the feedback data, and affects future prediction.

In our system, there are two kinds of feedback data, positive feedback and negative feedback. We call it a positive feedback when the predicted consumption rate is higher than actual consumption rate and negative feedback is the opposite where predicted amount is less than the actual one.

	"Sleeping"	"Watching TV"	"Eating"	"Working"	"Resting"
12-2PM	N/A	520KW	78KW	90KW	45KW
2-4PM	N/A	550KW 539KW	110KW	85KW	40KW
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
.10-12PM	24KW	470KW	N/A	99KW	30KW

Figure 3. Applying Feedback Data

After receiving the feedback data, we reflect this data to our future prediction by dividing the difference between the predicted amount and the actual amount by the number of data we have used to draw our prediction (# of occurrence). For example, let's say the current situation is "Watching TV" and the received feedback is +110 KW (predicted amount = 550 KW, actual amount = 440 KW, number of occurrence = 10). We divide 110 KW by 10 which give us 11 KW and then subtract this number from our prediction amount. In other words, our next prediction amount of that situation at that time slot will be 539 KW instead of 550 KW while the number of occurrence remains as 10. (Fig 3)

The feedback data we have mentioned above are only used to modify the prediction amount of consumed energy. In addition to that, we also allow user to give feedback to the derived situation. The Situation-Based Prediction system makes it prediction assuming the derived situation is accurate. However, the accuracy of the Bayesian Network is not always 100% and needs to adjust its network to a specific user or a group of users. Since cell phone is working as a main device by actually running the Bayesian Network. Whenever a situation is derived (once every two hour), the cell phone outputs the resulting situation and asks whether the situation is correct or not. This feedback data is only used to reinforce the Bayesian Network.

V. EXPERIMENT

At this point, our experiment is being done in a simulation environment. The simulator consists of a dining room, four living rooms, and two bath room where each room equipped with TV, light, and personal computer. The experiment is targeted to verify two main issues. First is to check the competitiveness of our Bayesian Network by analyzing the accuracy of the derived situation. The other issue is to see whether SBP indeed gives better performance compared to other energy prediction systems. Unfortunately, it is hard to present the results without experimenting in real environment; however, according to our simulation SBP does have high accuracy when the derived situation is accurate. The details of our simulation are following. We built a simulator that simulates user's daily home life. User randomly acts one of 6 situations; Sleeping, Watching TV, Eating, Working, Resting, and Moving. As we mentioned before, when "Moving" occurs, the system waits until user enters a certain space and detects one of the situations. We recorded 2 hour based energy consumption rate for 30 days in simulator environment. Fig. 4 shows the results of the experiment.

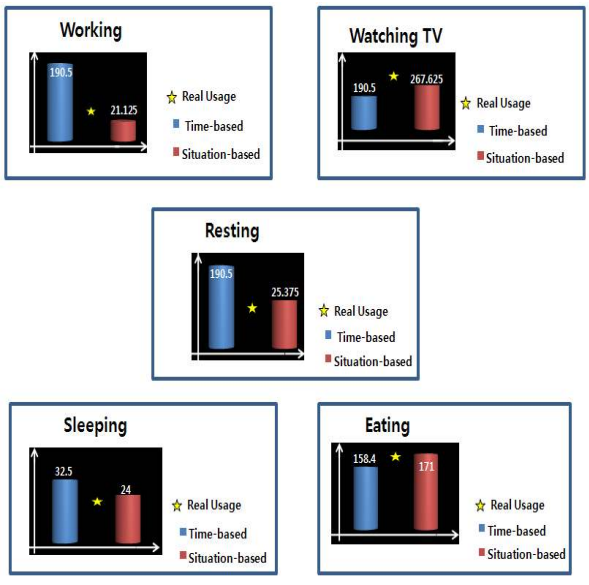


Figure 4. Comparison between Time-based prediction and Situation-based prediction.

For comparison, each graph presents average prediction value of time-based system, situation based system, and the

actual usage rate of the user. Situation-based prediction system gave much better performance when multiple situations occur in certain time slot. For instance, if user either watches TV or sleeps in certain time slot, there is a big difference in consumption rate between these two situations and time-based prediction cannot reflect this circumstance.

In average, time-based system had 54% average error rate when multiple situations occur in certain time slot whereas, situation-based system had only 12% average error rate. We need to acknowledge that simulation was run upon assumption that situation was accurately derived at all time. The accuracy of our Bayesian Network needs to be measured in real environment, and for now we leave it as our future work.

VI. ADVANTAGE OF SITUATION-BASED PREDICTION

A. Accuracy

While most of the modern energy management systems rely on past energy consumption history, SBP tracks history by analyzing user's situation. Energy consumption happens when activity occurs and situation is one of the best ways to predict user's activity. We are certain that SBP's accuracy is at least higher than the system that solely relies on past usage history, since SBP is also based on history data. The biggest difference is that, SBP segments these usage data and retrieves the most relevant history.

B. Specialization

As we mentioned earlier, SBP is based on probabilistic network such as Bayesian Network. One of the biggest strength of Bayesian Network is that it evolves as it receives feedback data from users. There are plenty of ways to receive feedback data from users directly or indirectly. However, we will not get into details of receiving and analyzing feedback data of Bayesian Network since it is beyond the scope of this paper. The essential point is that SBP keeps evolve as the size of the usage history data grow, and as it receives more feedback data from users. In other words, SBP evolves to get customized for a specific user or a specific group of people.

VII. CONCLUSTION AND FUTURE WORKS

In this paper, we presented a Situation-Based Prediction system for energy management. The system is based on Bayesian Network, and derives user's situation in order to associate it with energy consumption rate. We are certain that SBP is more efficient than other energy managing system that solely relies on users' past usage history and is capable of adjusting the system to a specific user or group of users.

SBP is currently targeted for a single user. Our future work will be focused on expanding our SBP to support multiple users in a distributed environment [5]. Each user will use his/her cell phone acting as a main sensor, and share information with one another. The biggest challenge of our future work will be detecting correlation among users. For example, when SBP detects that multiple users are watching TV, it needs to find out whether all of the users are watching the same TV or not. In addition, creating a database that holds individuals' past situation and associating one another to draw a reliable prediction will be the key challenge.

REFERENCES

[1] C. Harris and V. Cahill, "An Empirical study of the potential for context-aware power management," UbiComp 2007: Ubiquitous Computing (2007), pp. 235-252, doi:10.1007/978-3-540-74583-3_14.

[2] R. K. Harle and A. Hopper, "The Potential for Location-Aware Power Management", UbiComp 2008: Ubiquitous Computing (2008), pp. 302-311, doi:10.1145/1409635.1409676.

[3] L. Uusitalo, "Advantages and challenges of Bayesian networks in environmental modelling", Ecological Modelling, vol. 203, No.3-4 (10 May 2007), pp. 312–318.

[4] P. Korpipää, J. Mäntyjärvi, J. Kela, H. Kernen, and E.-J. Malm, "Managing context information in mobile devices", Pervasive Computing, IEEE In Pervasive Computing, IEEE, vol. 2, no. 3, pp. 42–51, doi:10.1109/MPRV.2003.1228526.

[5] G. Pavlin, P. de Oude, M. Maris, J. Nunnink, and T. Hood, "A multi agent systems approach to distributed Bayesian information fusion", INFFUS, vol. 11, no. 3, pp. 267-282, doi:10.1016/j.inffus.2009.09.007.

 ISBN: 978-1-61208-164-9

A Storyboard-based Mobile Application Authoring Method for End Users

Jun-Sung Kim, Byung-Seok Kang and In-Young Ko
Dept. of Computer Science, Korea Advanced Institute of Science and Technology (KAIST)
291 Daehak-ro, Yuseong-gu, Daejeon, Korea
{junkim, byungseok, iko}@kaist.ac.kr

***Abstract*—Mobile computing focuses on supporting everyday activities of users by providing services that utilize mobile computing resources. More diverse types of users and computing resources have been engaged in mobile computing environments. Considering these characteristics, it is essential to support making end-users actively participating in mobile computing environments based on high-level goals of the users. To meet these requirements, we have devised a storyboard-based application authoring method. The main elements of this approach include a storyboard model to structure the mobile applications and a semantically-based abstraction method to represent complex applications in terms of abstracted scenes in a storyboard. This approach improves the existing visual programming paradigms which mostly focus on visually composing fine-grained programming elements. We have developed a prototype implementation of the application authoring tool and tested it with a group of users to prove the effectiveness of allowing end-users to create and manage mobile applications.**

Keywords - Mobile Application Authoring; Application Storyboard; End-user Software Engineering; Visual Programming.

I. INTRODUCTION

To achieve the user-centricity goal in providing mobile applications [1], new computing paradigms such as service-oriented computing (SoC) and task-driven computing (TDC) have emerged [2][3]. These approaches focus on separating the concerns of selecting and coordinating specific services from the concerns of recognizing users' high-level computing goals. These allow end-users to more easily interact with numerous computing resources in a mobile computing environment with their high-level task goals. However, SoC and TDC in mobile computing have been considered mostly within the context of "use of applications" rather than "authoring of applications". Therefore, it is normally difficult for end-users to define their task goals and to arrange and access mobile computing resources that are necessary to accomplish their goals.

In this paper, we propose a storyboard-based mobile application authoring method by which end-users can intuitively specify their task goals as a storyboard and easily generate a mobile application from the storyboard. In a storyboard, users can specify the necessary functionalities and structure of an application as high-level activities (scenes). Once a mobile application is created, it can be validated, executed, evaluated, and personalized.

Our approach also allows users to identify and reuse common patterns of defining storyboards to accomplish a type of goal. The main elements of our approach include a storyboard model to structure the mobile applications and a semantically-based abstraction method to represent complex applications in terms of abstracted scenes in a storyboard. Fig. 1 shows the overview of the storyboard-based mobile application authoring method.

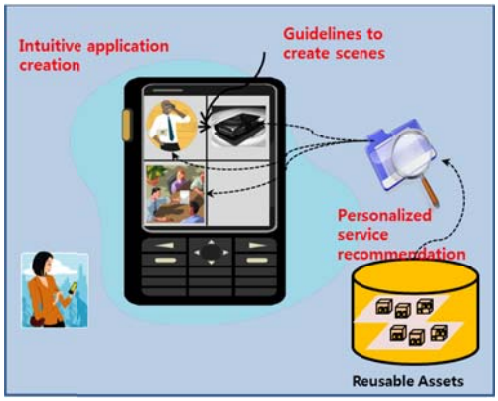


Figure 1. Overview of the storyboard-based mobile application authoring

Scaffidi *et al*. reported that most of the computer users these days are non-professionals who do not know much about conventional programming [18]. In mobile computing environments, there is even a bigger proportion of non-professional end-users.

End-user software engineering is a paradigm of allowing end-users to develop software to meet their own needs while bridging the gap between their high-level requirements and detail system capabilities [4]. Our approach provides an end-user software engineering framework that covers the overall lifecycle of software development including requirement specification, application generation, evaluation, and evolution. In comparison to the existing visual programming paradigms which mostly focus on visually composing fine-grained programming elements, the storyboard-based authoring method enables end-users to represent their requirements in a higher-level abstraction.

This paper is organized as following. We introduce the major requirements of end-user mobile application authoring and explain the related works in Section II and Section III, respectively. Section IV and Section V describe the core approach of the storyboard-based authoring. Section VI shows the evaluation results. Finally, Section VII concludes the paper by explaining the main contributions and future works.

 ISBN: 978-1-61208-164-9

II. REQUIREMENTS OF END-USER MOBILE APPLICATION AUTHORING

A. User-centricity – Task-driven abstraction and visualization

In a mobile computing environment, it is especially crucial for end-users to access mobile computing resources based on their task goals without considering any technical details. End-users need to be able to focus on describing what they need to achieve their task goals rather than expressing the detail structure and functions to be implemented in their applications. In addition, there must be a visual aid to allow users to intuitively represent their task requirements and to understand the core activities to be supported by a mobile application.

B. Efficiency – Automated and non-error-prone authoring and instantiation processes

Most end-users are non-programmers who do not have enough technical skills to create, recognize and compose services and to monitor applications [5]. Therefore, it is essential to minimize such technical burdens of users by automating the process of selecting and combining component services that are necessary to accomplish a task goal. In addition, there must be a mechanism of bridging the gap between a high-level task representation and a set of services available in a mobile computing environment, and making automated bindings between those two different abstractions.

End-users' authoring activities are normally error prone. Therefore, the authoring process should support the ways of resolving mismatches between required capabilities and available service functions, and validating an application against a user's task goal.

C. Reusability and Evolovability – Reuse of common application patterns and support of application evolution

In mobile computing environments, there is a wide spectrum of applications to be supported for diverse types of users. In these environments, reuse of common application patterns and evolution of applications based on users' feedback are critical to reduce development efforts of applications and to improve the quality of applications [11][12]. Therefore, it is necessary to provide a mechanism to identify a common structure and functionalities to support a similar set of user tasks and to enable these common application patterns to be refined and extended based on users' feedback. The application instantiation process also needs to meet these requirements by providing a mechanism of reusing successful task-service bindings for similar situations, and by making these binding patterns evolvable.

D. Mobile Usability – Usability support in user-interface-constrained mobile computing environments

Mobile usability is about allowing mobile users to effectively interact with an application by using User Interface (UI)-constrained computing environments such as smart phones and tabular PCs. The end-user mobile application authoring environment should support mobile usability such that users can effectively represent and recognize the core structure of task activities by using their mobile devices. The granularity of UI elements that comprise a task-driven abstraction of an application needs to be coarse-grained enough to be efficiently visualized on a small screen, and to be controlled by using constrained input methods. Especially, it is essential to make the high-level UI abstraction of an application consistent with the detail application integration structure [6].

III. RELATED WORK

End-user software engineering is a paradigm to allow end-users to create and manage software applications without having deep programming knowledge and skills [4]. Many approaches have been developed to help end-users conduct various development activities throughout the software development lifecycle.

A. Flowchart-based Approach

In this approach, end-users can draw a flow of component services and conditions by using a common format or template provided by developers [13][14]. Although this allows users to structure an application based on the main flow of activities and events that are important in a specific domain, it is often difficult for end-users to learn and understand detail notations (branches, loops, etc.) and options (e.g., sequential structure vs. parallel structure) to represent a flow.

The activities in a flow are normally represented at the same abstraction level as component services. Therefore, users need to associate each activity to a specific component service to be used. It is usually a difficult job for end-users to recognize, select and compose component services with understanding their functionality and other technical factors such as interfaces, preconditions and post-conditions. In addition, the detail flow structure cannot be shown effectively on a small screen of a mobile device.

B. Wizard-based Approach

In this approach, end-users can create and customize applications by creating forms and representing dialogues that are needed to be used in user interactions [15][16]. The step-by-step dialogue sequence and appropriate forms to provide at each step can be specified in the application definition. The wizard-based approach relatively does not need users to understand complex notations and technical factors. However, this approach requires careful modeling of the forms and dialogues. The forms need to be modeled such that the users can easily understand the desired inputs to be provided at a step. In addition, detail conditions and branches of the steps cannot be easily programed by end-users. In this approach, some error-toleration features can be incorporated to ensure the quality of data filled in a form.

C. Spreadsheet-based Approach

This approach allows end-users to create applications by putting values and assigning computations to designated cells in a spreadsheet [19]. Since spreadsheets are widely used by people, end-users can easily learn how to make applications by editing cells in a spreadsheet. In addition, some features to improve the dependability of application can be supported by adding interactive and dynamic testing capabilities described in [19]. However, the types of applications that can be developed by using this approach are limited to the ones that require management and computation of data in a tabular form. In addition, computational rules are normally hidden behind the

 ISBN: 978-1-61208-164-9

visual representation of a spreadsheet, and novice users may have difficulty of creating and validating them.

D. Storyboard-based Approach

A storyboard is a series of visual illustrations sequentially arranged and displayed. It has been widely used in the movie, advertisement, and multi-media domains. There have been some attempts to use storyboards for designing and creating applications [9][10][17]. In this approach, service functions are abstracted to 'scenes' that can be arranged into a storyboard. A storyboard visualizes the structure and semantics of an application, and provides users with a series of interfaces to interact with the application. To create an application, end-users firstly identify available service functions via predefined mockup scenes, and then select and arrange the scenes in a storyboard template. Users can understand the semantics of an application by interpreting the sequence of the scenes selected.

This storyboard-based approach provides end-users with an intuitive interface to select and compose component services. In addition, a mockup scene can be used to effectively visualize the essential functionality of a component service. In addition, a storyboard that is composed of multiple scenes can be visualized and browsed effectively on the small screen of a mobile device.

However, similar to the wizard-based approach, most of the storyboard-based approaches cannot visualize and control detail application structures. In addition, large-scale applications that need to be composed of many scenes arranged in various structures cannot be efficiently created and managed by using storyboards. Some researchers have tried to overcome the limitations by providing detail structure representations on storyboards. However, the complex storyboard representations increased the difficulties of understanding and controlling component services to be accessed to accomplish a user task. In other words, although we can represent some control structures such as loops, branches and parallelism in a storyboard, a mobile application that is presented in a storyboard becomes too complex to understand and manage by end-users.

In each scene of a storyboard, users need to be able to represent their computational needs in their own perspective. However, most storyboard-based approaches lack the ability of hiding the technical details of component services and hardware devices in an environment. In addition, most of them do not support the reuse of existing storyboards to make new mobile applications based on previously defined compositional patterns. Without the support of finding and reusing existing storyboards, it is hard for the end-users to represent a service composition from scratch with considering different candidates of services in a local environment and the dynamically changing context of using an application. As we discussed in Section II, it is necessary for end-users to find most appropriate patterns of making storyboards for their needs, and generate an application by simply extending and customizing them. The previous storyboard approaches also do not provide facilities to validate the correctness of a service composition generated from a storyboard.

IV. Storyboard-based Mobile Application Composition Model

We have developed our storyboard-based application composition model based on the task-oriented application framework [20]. As depicted in Fig. 2, the model is composed of three different views: visualization model (called U-board), task meta-model, and instances.

The *visualization model* is for allowing end-users to describe their computing tasks as a story which is composed of multiple activity scenes. The *task meta-model* defines a decomposition structure of user-centric applications. A task is created for each computational goal of a user, and composed of unit tasks each of which defines a compositional pattern of services for performing an action in the task. A unit task defines a set of necessary component services and the interconnection structure among them. The *instances* are actual instances of applications and services that can be run in a local environment. Instances of a task are generated by considering system specific characteristics and environmental conditions.

As shown in Fig. 2, the task meta-model maps the user-centric visualization into the elements of an application instance. All the entities in our models are described and managed by using ontologies. In other words, the semantics of task stories, scenes, tasks, unit tasks, and services are described by using domain specific ontologies, and their semantic relationships and similarity can be inferred by using a reasoning engine [20]. In this section, we focus on explaining the content of the visualization model.

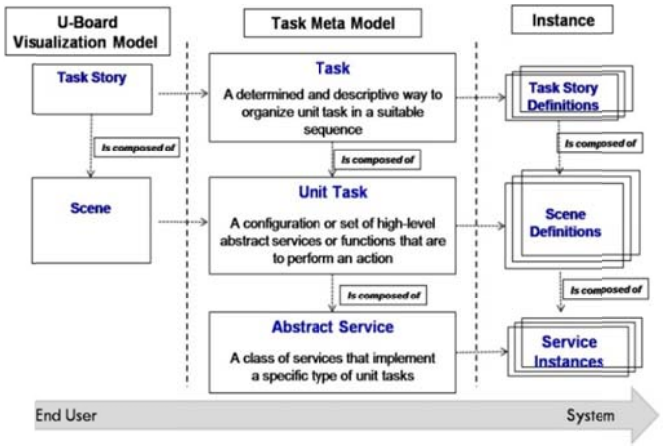


Figure 2. Three views of the storyboard-based mobile applcation composition model

A. Scene Visualization Model

As we discussed in Section II, each scene of a storyboard needs to be highly readable even on mobile devices. A shown in Fig. 2, a scene is composed of three parts: activity name, representative image, and context information. The activity name is a textual name of a scene, and the representative image visualizes the essential characteristics of a scene such as an object, movement, and operations. The context information represents the temporal and spatial context in which the scene is activated and becomes valid.

 ISBN: 978-1-61208-164-9

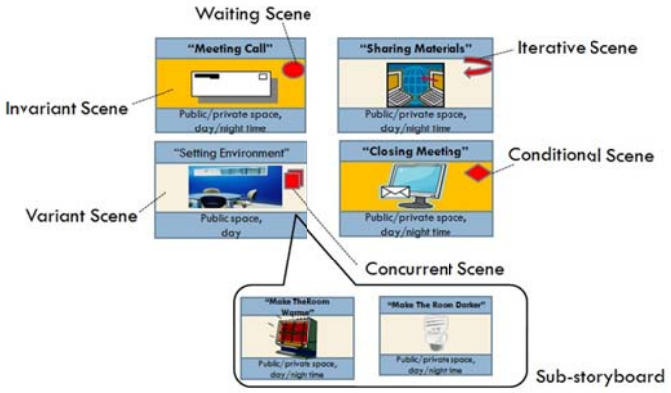


Figure 3. Examples of scene visualization (the scenes defined in the task story called 'Having a Meeting')

B. Scene Types

A task story can be composed of scenes that are either invariant or variant scenes. *Invariant scenes* are static parts of a story that are not changed over time and are common across different situations. At the example story shown in Fig. 3, 'Meeting Call' and 'Closing Meeting' are the scenes (marked with the yellow background) that are common in every meeting tasks, and defined as invariant scenes. *Variant scenes* are the ones that can be replaced by an alternative scene based on user preferences and environmental conditions. In the example above, 'Sharing Materials' and 'Setting Environment' are variant scenes that can be substituted to an environment-specific or customized scene. For example, 'Shareing Materials' can be replaced by a scene of exchaning secure emails to handle critical information. In most of the cases, variant scenes are replaced by personalized or more specialized scenes automatically based on user preferences or environmental conditions. This is to meet the efficiency requirement explained in Section II.

Scenes are also categorized into four groups based on their control structures: sequential, concurrent, iterative, and conditional scenes. The scenes that are arranged in a storyboard are *sequential* in default. The *concurrent scenes* that need to be executed in parallel can be grouped together into a scene as depicted in Fig. 3. The compound scene can be expanded into multiple, concurrent scenes by clicking on the control icon (overlapped rectangles). Scenes with a curved arrow are *iterative scenes*, which services are executed iteratively while a condition is met. The iteration condition of a scene is represented as a set of properties, and can be checked by double clicking on the curved-arrow icon. A diamond icon that is shown on a scene means that the scene is a *conditional scene*. A conditional scene is activated when a condition, which is represented as a set of properties, is met.

In our approach, control structures can be imposed only on each scene rather than across multiple scenes in a storyboard. Although this limits the representation power of controls, our simple and graphical controlling mechanism makes the authoring and management of storyboards much simpler and easier for end-users.

V. IMPLEMENTATION

We have implemented a prototype of our end-user mobile application authoring tool. Fig. 4 shows screen shots of the tool. The main screen shows a storyboard canvas and a task radar. The task radar allows users to visually control some contextual aspects such as spatial, social, temporal and personal aspects to narrow down the candidate storyboard to reuse for a task. The tasks found are shown in a hierarchical structure (based on their ontological relationships) in the task browser. When a task story is selected from the task browser, the sequence of scenes is displayed in the storyboard canvas, and the detail list of scenes and their decomposition structures are shown in the scene browser.

Our tool is developed by using Java SWT. We used jfreechart 1.0 for implementing the task radar. This client is installed in two ultra-mobile PCs (SAMSUNG Q1, Fujitsu U2010). In addition, the semantic descriptions of unit tasks and task stories are made by using the Protégé ontology editor [7]. The semantic reasoning of finding storyboards and scenes are implemented by using the Jena library [8].

By using this tool, a user can define an initial task story by finding and selecting a storyboard template that is most appropriate for his or her task. Then, the end-user can customize the initial task story by rearranging, adding, and deleting scenes on the storyboard. A task story is defined by arranging a set of scenes each of which has its URI, name, control structure type, variability conditions, and contextual properties.

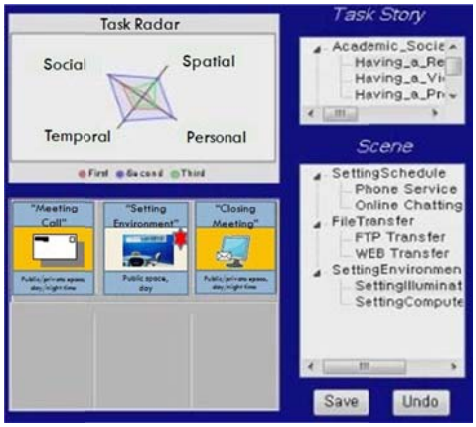


Figure 4. Layout of the mobile application authoring tool

While a task story is defined, the application authoring tool analyzes the relationships among the scenes and automatically suggests necessary modifications on the task story if it detects mismatch between inputs and outs of consecutive scenes or contextual inconsistency among scenes.

Once all authoring steps are finished, the end-user can hit the 'save' button to initiate the process of converting the storyboard representation into a concrete service composition that can be executed in the local environment. The storyboard is saved with some semantic description and can be found and reused for similar purposes.

 ISBN: 978-1-61208-164-9

VI. EVALUATION

To evaluate our approach whether it meets all the requirements to enable end-users to create mobile applications, we conducted a user test. We recruited sixty users and provided them with an operation manual (Table I) of the application authoring tool. The users are mostly graduate students at our school, but only few of them are expert programmers. To compare our approach against the existing storyboard approaches, we divided the sixty users into three groups. Each group participated in evaluating one type of approach to avoid learning effects because we measured the time to finish the customization of a task story according to a given scenario and counted the number of steps to accomplish the job.

TABLE I. OPERATION MANUAL OF THE APPLICATION AUTHORING TOOL

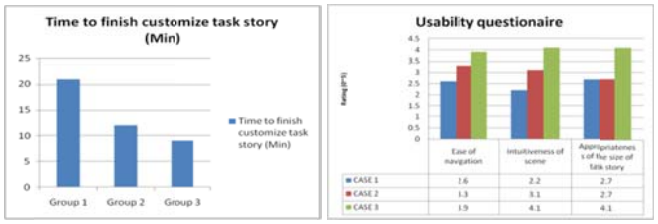
Operation	Descriptions
Add	Drag and drop scenes onto U-Board among the recommended scenes
Delete	Press delete key on the scene
Move	Drag and drop scenes onto the any cell in U-Board
Retrieve	Input the functionality of scenes to the retrieval window
Save	Press save button on task authoring tool
Undo	Press undo button to rollback to the original task story

TABLE II. EVALUATION CRITERIA AND METHODS

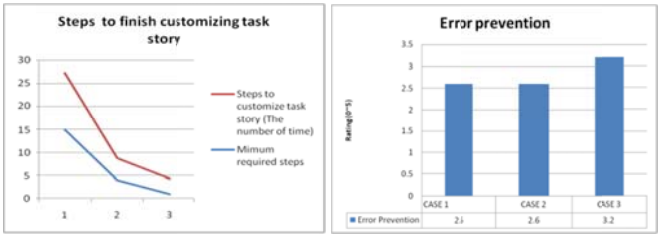
Requirements	Evaluation Methods
Usability	Easy to entry (measure time to finish customizing a task story according to a given scenario)
	Intuitiveness of scenes and task story Q) Difficulties to understand scenes and task story
	Appropriateness of the size of user interface considering mobile device Q) Appropriateness of the size of each scene
	Appropriateness of the volume of information Q) Do you think we provide too much information to bother the use of tool?
Efficiency	User efforts (measure the number of steps to customize a given task story)
	Error prevention Q) Do you think the guidelines are helpful to solve difficulties in customizing task story
Reusability	Ease of Add/Delete/Move/Save Scenes Q) Do you think Add/Delete/Move/Save functions are working properly
	Accessibility to existing scenes and task stories Q) Do you think it is easy to access to existing scenes and task stories

To the first group of users, we provided a tool that shows only a monolithic image for each scene and does not support the abstraction and mapping of scenes into task activities. To the second group of users, we provided a tool that visualizes scenes based on our scene visualization model, but does not support the task-oriented abstraction of scenes. Our task-oriented storyboard authoring tool was given to the third group of users.

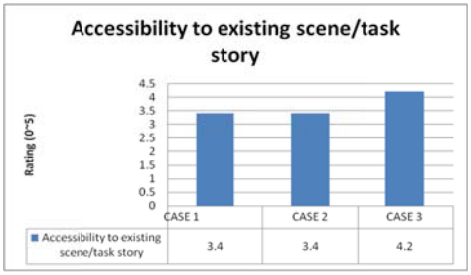
We evaluated the three groups to check whether they meet the usability, efficiency and reusability requirements that we explained in Section II. To measure the usability, we counted the time taken by the user groups to finish customization of a task story, and asked a series of questions to check various usability factors. Efficiency is measured by counting the number of steps to finish the customization of the task story, and by checking the effectiveness of error prevention facilities supported by the tools. Reusability is measured by asking a couple of questions that are about the easiness of managing scenes, and the effectiveness of finding and accessing existing scenes and task stories. Table II summarizes these evaluation criteria and methods.



(a) Usability



(b) Efficiency



(c) Reusability

Figure 5. Evaluation Results

Fig. 5 shows the evaluation results. As shown in Fig. 5(a), our approach (Case 3) lowered the barrier to entry to the application authoring job by reducing the time to learn how to customize and manage task stories. The answers to the usability questionnaires also show that our approach is much more intuitive and satisfactory than other two approaches.

In terms of efficiency, our approach contributed to reduce the number of steps to customize task stories and to prevent errors during the application authoring process. As shown in Fig. 5(b), by using our approach, the users had to perform 9 steps in average where the first approach required average 22 steps.

 ISBN: 978-1-61208-164-9

Finding appropriate storyboards to reuse is crucial to improve the reusability of service compositions. The survey result shown in Fig. 5(c) proves that our approach helped the users to find and access useful storyboards to reuse.

VII. CONCLUSION

In this paper, we proposed the storyboard-based end-user mobile application authoring method. The main goal of our approach is to allow end-users, who do not have sophisticated technical knowledge about developing mobile applications, to easily create and customize those applications. We identified three essential requirements (usability, efficiency, and reusability) of mobile application authoring for end-users to successfully represent their task goals and required contexts in an application description.

Our task-oriented storyboard approach provides an environment in which end-users can develop mobile applications without having technical knowledge. Users can visually browse through existing storyboard templates by controlling multi-dimensional aspects, and easily extend and customize them to generate mobile applications to achieve their computational goals. By using our tool, users can represent compound scenes and essential control structures that are effective to manage and dynamically instantiate storyboards according to the changes of environmental conditions and user requirements.

We are currently in progress on testing our application authoring tool by applying it to the public application domains in our campus and conducting a research to make the application authoring more evolvable by accepting and reflecting feedbacks from end-users. The accumulated feedbacks are analyzed in spatial, temporal, personal, and social perspectives. These can be automatically reflected in selecting or composing unit tasks and services for application authoring.

ACKNOWLEDGMENT

This work was supported in part by the IT R&D program of MKE/IITA [KI001877, Location/Societal Relation-Aware Social Media Service Technology]. This research was also supported by the National IT Industry Promotion Agency (NIPA) under the program of Software Engineering Technologies Development.

REFERENCES

[1] Hansmann., Pervasive Computing: The Mobile World. Springer, 2003, ISBN 3540002189.

[2] Wang, Z. and Garlan, D., Task-Driven Computing. Technical Report, CMU - CS -00-154, 2000.

[3] W.T. Tsai and Yinong Chen., Introduction to Service-Oriented Computing, Arizona State University, http://www.public.asu.edu/~ychen10/activities/SOAWorkshop. <retrieved: July, 2011>

[4] Andy Ko., The State of the Art in End-User Software Engineering, http://www.sei.cmu.edu/interoperability/research/approaches/upload/Lewis-SEEUP2009-Workshop-20Review.pdf. <retrieved: July, 2011>

[5] David Garlan, Dan Siewiorek, Asim Smailagic, and Peter Steenkiste., "Project Aura: Toward Distraction-Free Pervasive Computing", *IEEE Pervasive Computing*, vol. 1, no. 2, pp. 22-31, Apr.-June 2002.

[6] Florian Daniel, Jin Yu, Boualem Benatallah, Fabio Casati, Maristella Matera, and Regis Saint-Paul., "Understanding UI Integration: A survey of problems, technologies, and opportunities," *IEEE Internet Computing*, vol. 11, no. 3, pp. 59-66, May/June 2007.

[7] The Protégé Ontology Editor and Knowledge-base Framework, http://protege.stanford.edu/. <retrieved: July, 2011>

[8] The Jena Semantic Web Framework, http://jena.sourceforge.net/. <retrieved: July, 2011>

[9] Yang Li and James A. Landay., "Activity-Based Prototyping of Ubicomp Applications for Long-Lived, Everyday Human Activities," Proc. *Twenty-sixth annual SIGCHI conference on Human factors in computing systems (2008)*, pp. 1303-1312.

[10] Agnes Ro, Lily Shu-Yi Xia, Hye-Woung Paik, and Chea Hyon Chon., Bill Organiser Portal: A Case Study on End-User Composition, Proc. *WISE 2008 Workshops*, Springer Berlin / Heidelberg, vol. 5176, pp. 152-161, 2008.

[11] Ivar Jacobson, Martin Griss, and Patrik Jonsson, Software reuse: architecture, process and organization for business success, ACM Press/Addison-Wesley Publishing Co., New York, NY, 1997.

[12] Nam-Yong Lee and Charles R. Litecky, "An Empirical Study of Software Reuse with Special Attention to Ada," *IEEE Transactions on Software Engineering*, vol. 23 no. 9, pp. 537-549, September 1997.

[13] W. M. Johnston, J. R. Paul Hanna, and R. J. Millar., "Advances in Dataflow Programming Languages, *ACM Computing Surveys (CSUR)*, vol. 36, no. 1, pp. 1-34, March 2004.

[14] Cycling 74 Max, http://www.cycling74.com/products/max.html. <retrieved: July, 2011>

[15] D. Draheim and G. Weber, Form-Oriented Analysis., A New Methodology to Model Form-Based Applications, Springer, October 2004, ISBN-10: 3540205934

[16] Zhiming Wang, Rui Wang, Cristina Aurrecoechea, Douglas Brewer, John A. Miller, and Jessica C. Kissinger., Semi-Automatic Composition of Web Services for the Bioinformatics Domain, http://cs.uga.edu/~jam/home/theses/z_wang_dissert/thesis/wsbiojournal/workflow-journal29.pdf. <retrieved: July, 2011>

[17] M.Haesen, J.Meskens, K.Luyten, and K. Conix., Supporting Multidisciplinary Teams and Early Design Stages Using Storyboards, Springer Berlin / Heidelberg, Human-Computer Interaction. New Trends, Volume 5610, pp. 616-623, 2009.

[18] Scaffidi, C. Shaw, C., and Myers, B., An Approach for Categorizing End-user Programmers to Guide Software Engineering Research. Proc. *First Workshop on End-user Software Engineering (WEUSE)*, pp. 1-5 at the 27th International Conference on Software Engineering (ICSE 2005), St. Louis, Missouri, USA, May 15-21, 2005.

[19] Margaret Burnett, Curtis Cook, and Gregg Rothermel. End-user software engineering. Commun. ACM 47, 9 (September 2004), pp. 53-58.

[20] In-Young Ko, Hyung-Min Koo, and Angel Jimenez-Molina. User-Centric Web Services for Ubiquitous Computing. J.D.Vel´asquez and L.C. Jain (Eds.): Advanced Techniques in Web Intelligence – 1, SCI 311, pp. 167–189, Springer-Verlag Berlin Heidelberg 2010.

 ISBN: 978-1-61208-164-9

Influence Factors in Adopting the m-Commerce

Francisco-Javier Arroyo-Cañada
Department of Economics and Business Organization
University of Barcelona
Barcelona, Spain
e-mail: fjarroyo@ub.edu

Jaime Gil-Lafuente
Department of Economics and Business Organization
University of Barcelona
Barcelona, Spain
e-mail: j.gil@ub.edu

***Abstract*—The development of mobile devices and wireless communications networks is an opportunity for the companies that want to achieve their customers anywhere and anytime. This work in progress explores the factors that influencing the adoption of m-commerce and proposes a methodology to aggregate the total influence of the incentives on the intention of use. The knowledge of these effects should improve the sale promotions policy, based on incentives, and can influence the adoption of m-commerce. The regression analysis confirms the relationship between the studied variables.**

Keywords–Incentives; m-commerce; Technology Acceptance Model (TAM); theory of the forgotten effects.

I. INTRODUCTION

The wireless communication networks are spreading worldwide at a rate ever recorded to date by any other communication technology [1]. It has become an everyday technology that is changing the relationships at work, family, personal relationships and use of leisure time available anywhere and anytime.

To date, the m-commerce had been to buy ringtones, screensavers, games, videos, etc. This is a stagnant market in which there is to think about developing new kinds of content that appeal to the majority, which is a challenge for marketing departments of content providers as stated [2]. The development of mobile technology and the growth of access to navigation via mobile devices open up new possibilities for m-commerce. New kinds of e-commerce transactions, conducted through mobile devices using wireless networks and other wired e-commerce technologies. Is possible group the m-commerce [3] in transactions services (e.g., mobile shopping, ticket purchasing, stock trading), information services (e.g., news, location/traffic information) and entertainment services (e.g., download rings tone, download movies).

The literature on the adoption of the m-commerce is centered in the study of the intrinsic factors that influencing the intention of use of the m-commerce, but the companies need external factors that they can manipulate to influence the users. This research introduces the incentives, as external factor, and evaluates the total effect, direct and indirect, on the intention of use of the m-commerce. We consider that the knowledge of the real impact of the incentives improves the policy of promotions with the objective to influence the m-commerce adoption.

This work is structured in five sections: the first section presents a review on adoption of m-commerce; the second section presents the research model and explains the used factors; the third section is detailing the data collection and the measurement of variables and proposes the analysis methodology; the fourth section shows our preliminary results of the research, and finally, expose the main contributions for the business and scientific community, and comment future research.

II. ADOPTION OF M-COMMERCE

Many authors have defined the concept of m-commerce as the extension of electronic commerce that use wireless devices and telecommunications networks to accede anywhere and anytime to the exchange goods, services and information.

A review of the concept of m-commerce can be obtained from [4]. A base for studying the determinants of the intention of use is the Theory of Reason Action (TRA) [5] Behavior and attitude toward are subjective norm related to behavioral intention. Theory of Planned Behavior (TPB) [6] adding the perceived behavior determinant of the behavioral intention. Numerous studies take the Technology Acceptance Model (TAM) [7], based on TRA, to study the determinants of technology adoption. Many authors have used the TAM model in the study of acceptance of different technological innovations: Internet and mobile Internet, software, laptops, etc. For the purpose of this work should highlight the work in the fields of m-commerce [8][9][10]. Although there are various studies on the potential of wireless technology and mobile services it is necessary to explore new factors that affect the use intention of m-commerce.

III. THE RESEARCH MODEL

A. Objective

The purpose of this work is to understand the influence of the factors that affect the use intention of m-commerce adding the incentives to the TAM. The incentives are external factors that can be manipulated by the enterprises. So the main objective of this research was to determine the overall effect of incentives on the use intention of m-commerce, direct and indirect effect through other factors.

 ISBN: 978-1-61208-164-9

B. Proposed research model

The context and the incentives have significant impact on users' behavior. We introduce in the model the incentives (I) as antecedent of the intention of use of the m-commerce (BI_m).

H1: The incentives have a direct positive effect on intention of use of the m-commerce.

The incentive strategies influence the attitude toward the technology and the perceived usefulness (PU) in the self check-in service. Therefore we introduce in the model the incentives as antecedent of the attitude and PU.

H2: The incentives have a direct positive effect on attitude toward m-commerce.

H3: The incentives have a direct positive effect on perceived usefulness.

Considering the uses and gratifications theory a person who receives incentives will be most happy. Greater happiness can influence the perception. Based on the uses and gratifications theory, several studies [7] show the influence of intrinsic motivation in the decision to use mobile services. We introduce in the model the incentives as antecedent of the perceived enjoyment (PE).

H4: The incentives have a direct positive effect on perceived enjoyment.

Perceived ease of use (PEOU) as "the degree to believe that which a person using a particular system is free of effort" and observed a positive interaction of PEOU with the PU [7]. PEOU has been considered as an important determinant in adoption of past information technologies such as intranet [11], 3G [12], online banking [13][14] wireless internet [15], Internet commerce [16] and m-commerce [17][18][19][20][21]. We introduce in the model the PEOU as antecedent of the PU.

H5: The perceived ease of use has a direct positive effect on perceived usefulness.

An application of the m-commerce easier to use will also more funny to use [22][23][24][25][26]. We introduce in the model the PEOU as antecedent of the PE.

H6: The perceived ease of use has a direct positive effect on perceived enjoyment.

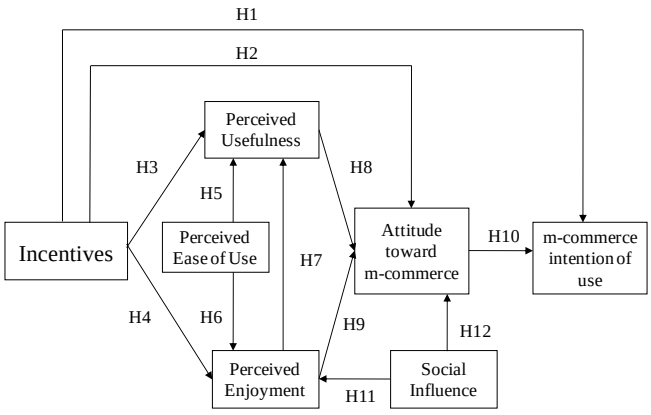


Figure 1. Reseach model.

PU as "the degree to believe that which a person using a particular system would enhance his or her job performance" and observed a positive interaction of PE with the PU [7]. Similarly, [27][28] and [29] working this hypothesis. We introduce in the model the PE as antecedent of the PU.

H7: The perceived enjoyment has a direct positive effect on perceived usefulness.

PU and the PE as antecedent of attitude applied to technology acceptance [7][24][25][30]. Similarly, the PE is present in the studies of [31][32]. We introduce in the model the PU and the PE as antecedents of the attitude.

H8: The perceived usefulness has a direct positive effect on attitude toward m-commerce.

H9: The perceived enjoyment has a direct positive effect on attitude toward m-commerce.

TAM propose the attitude toward particular system as antecedent of the behavioral intention of use following the TRA and the TPB. Then we introduce in the model the attitude as antecedent of the intention of use of the m-commerce (BI_m).

H10: The attitude toward m-commerce have a direct positive effect on intention of use of the m-commerce.

The social influence is present in the TRA [5] and the TPB [6] as antecedent of behavioral intention through the subjective norms. We introduce in the model the social influence (SI) as antecedent of the PE and the attitude toward the m-commerce (Attm).

H11: Social influence has a direct positive effect on perceived enjoyment.

H12: Social influence has direct positive effect on attitude toward the m-commerce.

IV. METHODOLOGY

A. Data collection

We made a survey with students because they are more susceptible to use this kind of services. A total 367 questionnaires have been validated. The sample is formed of 60.3% of women and 39.7% of men, with ages between 19 and 37 years. As for the experience in online purchase, 95.6% have occasionally bought on the Internet and 14.7% have bought some product through the mobile devices.

B. Measurement of variables

A literature review facilitates the scales for measure the constructs. According to [33][34] a factor analysis allows checking the validity of measurement scales. All items should be higher of 0.6. A Cronbach's alpha analysis was done in order to analyze the reliability of the used scales and to test internal consistency of the scales. All factors exceed 0.7, as recommended [35].

Likert scales (ranging from 1 to 7), with anchors ranging from ‘‘strongly disagree’’ to ‘‘strongly agree’’ were used for all questions.

 ISBN: 978-1-61208-164-9

TABLE I. THE SCALES' FIT

Sacales	Items	KMO	Total Variance Explained	Cronbach's Alpha
PE	1-9	0.835	58.316%	0.907
PU	10-14	0.864	82.529%	0.946
PEOU	15-20	0.807	66.822%	0.896
SI	28-31	0.676	86.896%	0.949
A	32-37	0.87	75.006%	0.933
BIm	38-40	0.733	86.089%	0.918
I	41-45	0.85	83.842%	0.947

We use the software SPSS 17 to make the factor analysis and the reliability test. We confirm, in a preliminary analysis the validity and the reliability of all scales considered keeping in mind the adjustment measures previously discussed.

C. Factors influences

To calculate the influence of each factor on the proposed constructs (perceived usefulness, perceived enjoyment, attitudes toward the m-commerce and intention of use of the m-commerce) has made four multiple regression analysis. The table II shows the preliminary results.

D. Incidences of first order and higher

The incidence between two variables is the cause-effect relationship of one variable on another. The simplest case is the incidence of first order (e.g., A affects B) where A is the cause of B. Second order is considered where the impact is not direct (e.g., A affects B through C).

In the the model there are incidences of first order: I → A Incentives (I) affect the attitude (Attm). But there are also incidences of second order: I → A → BI_m, Incentives (I) have an indirect effect on the intention of use of m-commerce (BI_m) through attitude toward the m-commerce (Attm).

With the intention of determining the overall effect of incentives on the intended use of m-commerce use the theory of the forgotten effects. To calculate the maxmin convolution of the impact of A on C from the impact of A on B and the incidence of B on C using the formula [36]:

$$\mu(a_i, c_k) = \bigvee(\mu(a_i,b_j) \wedge \mu(b_j,c_k)) \quad (1)$$

For all a_i, b_j y c_k $i = 1, 2, 3,\ldots$ $j = 1, 2, 3,\ldots$ $k = 1, 2, 3,\ldots$

Maxmin convolution can be represented with the symbol (^{o}) so that the process can be summarized as follows:

$$m_{ac} = m_{ab} \circ m_{bc} \quad (2)$$

In the case of an incidence of order 3:

$$m_{ad} = m_{ab} \circ m_{bc} \circ m_{cd} \quad (3)$$

Given that the maxmin convolution satisfies the associative property:

$$m_{ab} \circ (m_{bc} \circ m_{cd}) = (m_{ab} \circ m_{bc}) \circ m_{cd} \quad (4)$$

In the case of incentives (I) there are incidences of first (I → BI_m), second (I → A → BI_m), third (I → PU → A → BI_m) and fourth order (I → PE → PU → A → BI_m) on the intention of use of m-commerce (BI_m). As develop the following:

$$m_{I\,BIm} = m_{I\,PE} \circ m_{PE\,PU} \circ m_{PU\,A} \circ m_{A\,BIm} \quad (5)$$

V. RESULTS

An exploratory analysis has been made with 67 questionnaires of the sample. The results of regression analysis can be seen in the following table:

TABLE II. REGRESSION ANALYSIS OUTPUTS

Dependent variable: Intention of use of the m-commerce.
Adjusted R^2: 0.859 SE: 0.483

Independent variables	B	SE	t	Sig
Attitude	0.698	0.110	6.367	0.000
PE	0.208	0.113	1.837	0.071
Incentives	0.087	0.047	1.852	0.069

Dependent variable: Attitude toward the m-commerce.
Adjusted R^2: 0.833 SE: 0.505

Independent variables	B	SE	t	Sig
PE	0.495	0.101	4.892	0.000
PU	0.251	0.067	3.768	0.000
Incentives	0.142	0.048	2.932	0.005
Social Influence	0.049	0.012	3.981	0.000

Dependent variable: Perceived usefulness.
Adjusted R^2: 0.671 SE: 0.839

Independent variables	B	SE	t	Sig
PE	0.454	0.131	3.457	0.001
PEOU	0.383	0.088	4.365	0.000
Incentives	0.321	0.070	4.605	0.000

Dependent variable: Perceived enjoyment.
Adjusted R^2: 0.501 SE: 0.678

Independent variables	B	SE	t	Sig
POEU	0.166	0.068	2.457	0.017
Incentives	0.141	0.053	2.646	0.010
Social Influence	0.070	0.014	5.075	0.000

All variables in the regression models are significant (Sig<0,1). Then is possible to confirm the hypothesis of the

research model. Although is necessary develop a confirmatory analysis with the 367 questionnaires of the sample.

Confirmed the hypothesis we calculate the impact of incentives on the set of constructs studied in the model with the proposed methodology based on the theory of forgotten effects.

VI. CONCLUSION AND FUTURE WORK

The development of the methodology proposed will allow study the aggregate effect of one factor on another, keeping in mind direct and indirect effects. It is important to aggregate the effects to know the importance of the factors in the acceptance of m-commerce process. Besides the incentives is a controllable factor for the company, so it can be used as a tool to improve the acceptance of m-commerce.

The combination of the regression analysis and forgotten effects provide a solution to aggregate all influences, direct and indirect, that exists between two variables. Other kind of analysis such as structural equation models obtain result from a group of factors on others but does not offer solution to find the total effect of one variable on another.

If the responsible of decision-making knows the total effects of the incentives on the intention of use of m-commerce then can influence to the users. The findings of this research can incorporate in the sales promotion policy to accelerate the acceptance of this new channel and augmenting sales.

REFERENCES

[1] M. Castells, "Comunicación móvil y Sociedad. Una perspectiva global", Ariel, Barcelona, 2006.

[2] S.W. Campbell, "El mundo en mi mano. La revolución de la telefonía móvil", Universidad UNIACC, Santiago de Chile, 2008.

[3] P. Harris, R. Rettie and C.C. Kawan, "Adoption and usage of m-commerce: A cross-cultural comparison of Hong Kong and the United Kingdom". Jounal of Electronic Commerce Research, nº 6, 2005, 210-224.

[4] H. Feng, "Exploring the Critical Success Factors for Mobile Commerce", unpublished.

[5] M. Fishbein and I. Ajzen, "Beliefs, attitude, intention and behavior: An introduction to theory and research", Reading, Addison-Wesley, 1975.

[6] I. Ajzen, "The theory of planned behavior", Organizational Behavior and Human Decision Processes, Vol.50, 1991, pp.179- 211.

[7] F. D. Davis, "Perceived usefulness, perceived ease of use, and user acceptance of information technology". MIS Quarterly, Vol.13, nº 3, 1989, 319-339.

[8] G. C. Bruner and A. Kumar,"Explaining consumer acceptance of handheld Internet devices". Journal of Business Research. Vol. 58, nº 5, 2005, pp. 553-558.

[9] J. Lu, C. S. Yu, C. Liu, and J.E. Yao, "Technology acceptance model for wireless internet", Internet Research: Electronics Networking Applications and Policy, Vol. 13 No. 3, 2003, pp. 206-22.

[10] H. Nysveen, P.E. Pedersen and H. Thorbjornsen, "Intentions to use mobile services: Antecedents and cross-service comparisons", Journal of the Academy of Marketing Science., 33(3), 2005, pp. 330–346.

[11] P. V. Chang, "The validity of an extended technology acceptance model (TAM) for predicting intranet/portal usage", Master thesis, University of North Carolina, Chapel Hill, NC, 2004.

[12] C. H. Liao, C. W. Tsou and M. F. Huang, "Factors influencing the usage of 3G mobile services in Taiwan", Online Information Review, Vol. 31 No. 6, 2007, pp. 759-74.

[13] P. Guriting, and A. O. Ndubisi, "Borneo online banking: evaluating customer perceptions and behavioral intention", Management Research News, Vol. 29 Nos 1/2, 2006, pp. 6-15.

[14] N. Jahangir and N. Begum, "The role of perceived usefulness, perceived ease of use, security and privacy, and customer attitude to engender customer adaptation in thecontext of electronic banking", African Journal of Business Management, Vol. 2 No. 1, 2008, pp. 32-40.

[15] J. Lu, "Technology acceptance model for wireless internet", Internet Research: Electronic Networking Applications and Policy, Vol.13, No.3, 2003, pp.206-222.

[16] D. Y. Cho, H. J. Kwon, and H. Y. Lee "Analysis of trust in internet and mobile commerce adoption", Proceedings of the 40th Hawaii International Conference on System Science, USA, 2007.

[17] H. H. Lin and Y. S. Wang, "Predicting consumer intention to use mobile commerce in Taiwan", Proceedings of the International Conferences on Mobile Business (ICMB'05), Sydney, Australia, 2005.

[18] S. Wang and S. Barnes, "Exploring the acceptance of mobile auctions in China", Proceedings of the Sixth International Conference on the Management of Mobile Business, Toronto, Canada, 2007.

[19] S. Kurnia, S. P. Smith and H. Lee, "Consumers' perception of mobile internet in Australia", e-Business Review, Vol. 5 No. 1, 2006, pp. 19-32.

[20] N. Mallat, M. Rossi, V. K. Tuunainen and A. Oorni, "The impact of use situation and mobility on the acceptance of mobile ticketing services", Proceedings of the 39th Hawaii International Conference on System Science, USA, 2006.

[21] P. Luarn and L. Hsinhui, "Toward an understanding of the behavioral intention to use mobile banking", Computers in Human Behavior, 21, 2005, pp. 873-891.

[22] M. Koufaris, "Applying the technology acceptance model of flow theory to online consumer behaviour", Information Systems Research,13(2), 2002, pp. 205-223.

[23] H. van der Heijden, "Factors influencing the usage of web sites: the case of a generic portal in The Netherlands". Information & Management, Vol. 40, nº 6, 2003, pp 541-549.

[24] J. Yu, I. Ha, M. Choi and J. Rho, "Extending the TAM for a t-commerce". Information and Management, nº 42, 2005, pp. 965-976.

[25] C. L. Hsu and H. P. Lu, "Why do people play online games? An extended TAM with social influences and flow experience", Information and Management, 41, 2004, pp. 853-68.

[26] J. H. Wu, Y. C. Chen and L. M. Lin, " Empirical evaluation of the revised end user computing acceptance model", Computers in Human Behavior, 23(1), 2007, pp. 162-174.

[27] M. Y. Yi, and Y. Hwang, " Predicting the use of web-based information systems: Self-efficacy, enjoyment, learning goal orientation, and the technology acceptance model", International Journal of Human–Computer Studies, 59(4), 2003, pp. 431-449.

[28] D. Cyr, M. Head and A. Ivanov, "Design aesthetics leading to m-loyalty in mobile commerce. Information & Management, 43(8), 2006, pp. 950–963.

[29] H. Sun and P. Zhang, "Causal relationships between perceived enjoyment and perceived ease of use: An

 ISBN: 978-1-61208-164-9

alternative approach", Journal of the Association for Information Systems, 7(9), 2006, pp. 618-645.

[30] M. K. O. Lee, C. M. K. Cheung and Z. Chen, "Acceptance of Internetbased learning medium: The role of extrinsic and intrinsic motivation", Information & Management, 42(8), 2005, pp. 1095-1104.

[31] B.H. Sheppard, J. Hartwick and P. R. Warshaw "The theory of reasoned action: a meta-analysis of past research with recommendations for modifications and future research". J Consum Res Dec. 1988, pp. 325-343.

[32] T. L. Childers, C. L. Carr, J. Peck, and S. Carson, "Hedonic and utilitarian motivations for online retail shopping behaviour". Journal of Retailing, Vol.77, 2001, pp. 511-535.

[33] R. Bagozzi and Y. Yi, "On the evaluation of structural equation models", Journal of theAcademy of Marketing Science, Vol. 16 No. 1, 1988, pp. 74-94.

[34] J.F. Hair, R. E. Anderson, R. L. Tatham and W. C. Black, "Multivariate Data Analysis". Prentice Hall, 1998.

[35] J. Nunnally and I. H. Bernstein, "Psychometric Theory", 3a ed., Mc Graw-Hill, 1994.

[36] A. Kaufmann, "Modèles mathématiques pour la stimulation inventive", Albin Michel, Paris, 1979.

 ISBN: 978-1-61208-164-9

Mobile Services through Tagging Context and Touching Interaction

Gabriel Chavira, Elvira Rolón, Eduardo Alvarez, Salvador W. Nava, Jorge Orozco
Graduate Division, Faculty of Engineering "Arturo Narro Siller"
Autonomous University of Tamaulipas
Tampico, México
{gchavira, erolon, ccalvar, snava, jorozco} @uat.edu.mx

***Abstract*— One of main objectives of Ambient Intelligence is the reduction to a minimum of the user's interactive effort, the diversity and quantity of devices with which people are surrounded with, in existing environments, increase the level of difficulty to achieve this goal. The mobile phones and their amazing global penetration, makes it an excellent device for delivering new services to the user, without requiring a learning effort. An NFC-enabled mobile phone will allow the user to demand and obtain services, by touching its different elements in the environment. In this paper we present a proposal where we analyze the scope associated with touch interaction, and where a model to perceive touch interaction through the tagging context is designed.**

Keywords-tagging context; touching interaction; mobile services;context aware.

I. INTRODUCTION

One of the first definitions of smart environment [1] [2] [3] arises from Ubiquitous Computing, which "created a new field of computer science, one that speculated on a physical world richly and invisibly interwoven with sensors, actuators, displays and computational elements, embedded seamlessly in the everyday objects of our lives and connected through a continuous network" [4], although Mark Weiser does not define it explicitly.

The vision of Ambient Intelligence [5], which is an evolution of Ubiquitous Computing, proposes a new way of thinking about computers, which will disappear in the environment, meaning that this perceiving and responding automatically to the presence of people is creating a smart environment.

The Ambient Intelligence (AmI) paradigm visualizes environmental management by applications, which will perceive in a continuous way the characteristics of the entities that comprise it and the natural interaction between them, thereby enabling applications to offer services either proactively or with the smallest possible interactive effort. Another characteristic of this type of environment is that, even with a strong technology, it is "invisible" to people; this disappearance can be obtained by embedding it in daily objects in the environment.

The final objective of a smart environment is to satisfy users' needs by providing services that require minimum interactive effort from them (the ideal service is one which the user receives without explicitly demanding it).

A smart environment must be able to perceive all the interaction techniques that people can develop. From these methods, the interaction of contact (or touch interaction) represents an opportunity area, justifying the design of a system that perceives. This interaction technique is simple, requires minimal effort and it is part of the natural reactions of people when they want to use an element of the environment, if it is within reach (if the device we want to use is near, we say we touched it).

Touch interaction can replace complex techniques and even intricate learning processes, for example: sending a photo to a new device, printing a document on a printer that has never been used, just as an older person will be able to request a meal from a company merely by touching a picture of it, etc. This feature and its ease of use by older people will play an excellent role in certain current issues, such as dependency.

Touch interaction is a simple technique; when it is developed between persons or between a person and an environmental element, it may involve a large amount and flow of information, which would represent a significant contribution to a smart environment.

A system that perceives and administers an environment's touch interaction would be able to offer services to users they could not have imagined and will be fundamental to the construction of the ideal smart environment.

Our work proposes perception of the touch interaction, which will be used to demand services at the moment of interacting with the environmental elements or entities. In order to obtain this perception, the "tagging" of the environment's entities will be necessary. The intention is that, when perceiving this interaction, the application that manages this environment will obtain information on the entities involved that, properly combined with the information in the application's databases, will enable services to be delivered.

II. USER SCENARIO

In the following scenario, we describe some activities at a research group; in these the users obtain services through "touching interaction". This scenario is the support for the application we are developing and testing.

John arrives at the building door, where his office and other workspaces (laboratory, other members' offices, and meeting room) of his research group are located. With his NFC-enabled mobile phone he touches the tag at the side of the main door of the building and the NFC-enabled mobile phone reminds him that he has an important comment for

 ISBN: 978-1-61208-164-9

George who is already working at his desk. For this reason John decides to go to the laboratory (where George is). At the moment John touched the tag, all the members of the research group, who are working in a computer, receive a message indicating that John has entered the building.

In a corridor, John can observe (on a public display) a summary of the research group's current work, such as deadlines of the congresses in which they will participate, the last versions of the papers being written, the identity and location of each person working in the building, etc.

When John arrives at the door of the laboratory he can observe who is in inside by looking at a little display. He can also see the degree of progress of the different activities (along with notes on projects, programs, articles, etc.) that the members of the group are developing. Before entering he touches the tag of the next door. Inside the laboratory he can observe a reminder of all "notes to comment on", on a public display at the laboratory, which has been stored in his mobile phone. Meanwhile, all users who have "notes to comment on" to John, can see a reminder indicating that John entered the laboratory on their computers.

While John talks to George, John places his mobile phone near the tag on the display of George´s computer to show a file. After commenting on it, they decide to show it to everyone in the laboratory. To this end, John touches the public display with his mobile phone.

Before John leaves the laboratory, George decides to send him a paper for checking, but, due to its large size, it does not fit in the mobile phone's memory. He therefore, decides to send the file to John so that it can be checked from any computer in the AmI environment.

When John leaves the laboratory, he runs the exit service in his mobile phone to aware the AmI environment that he is coming out of the laboratory. When John arrives at his office and touches the tag in the door, his mobile phone shows the list of people who came to see him while he was out, as well as the messages left for him.

III. ENVIRONMENT ENTITIES

The model we are proposing would endow an environment with the capability of perceiving the touch interaction between environment entities, which we define as: The intentional approach of two entities in order to obtain a service. This implies that when an entity approaches another one, the touch interaction arises. The objective of this model is limited to the touch interaction, which involves only two elements, of which one is a person. This is the reason why the touch interaction of interest to us is defined as: A person's deliberate touching of an environmental entity (the latter can be another person) for the purpose of obtaining services.

One of the most popular and referenced context and entities definition is given by Anind K. Dey, who states that "any information that can be used to characterize the situation of an entity. An entity is a person, place or object that is considered relevant to the interaction between a user and an application, including the user and application themselves" [6].

Owing to the fact that we are interested only in a specific interaction technique between a person and an environment element or another person, we have adapted Dey's definition so as to define the context limited to touch interaction as: any information on the involved entities that is required to deliver the services offered by the touch interaction.

A. *Entities*

The PICTAC (it perceives touch interaction through tagging context, in Spanish) objective is to develop a system that manages the touch interactions of an environment in which at least one person takes part. This is the reason why the entity "application" will not be considered by us.

Since the capacity to participate in a touch interaction that is perceived by the environment is not a person's natural ability and so that a system can take into account a person, their data must have been captured previously. If we add these capabilities to the "person" entity, we create the "user" entity (which would be a subset of the entity proposed by Dey), as used in our model.

Dey's "object" entity has proved to be too general for the purpose of the model, since from the standpoint of the service it can provide, we can distinguish two categories of objects: Devices, those whose service can be demanded through a touch interaction (basically electronic or computer equipment), and Objects, those whose service is not "suitable" or cannot be demanded through a "touch" (e.g., a desk, furniture). However, in the case of the latter, through their location and continued use, we can take advantage of them to integrate some in the model we are proposing and offer services. This is the reason why we have broken down the object category into two types of entities: "object" and "device". It must be remembered that an object does not even have the capacity to process and communicate which is the reason why the capacities that it could have will depend on the available computer device that allows it to participate in a touch interaction and on the user's capacities.

TABLE I. COMPARISION BETWEEN DEY ENTITIES AND OURS

Dey entities	**Adaptation**	**PICTAC Model**	
		Entities	**Description**
Person	Only the person or equipment are considered	**User**	Person with the capacity to participate in a touch interaction
Object	They are divided into two, depending on the form given to the services	**Object**	Those whose service is not "suitable" or cannot be requested through "touch"
		Device	The service can be demanded through a touch interaction
Place	Idem	**Place**	Represents a part of a smart environment or even the entire environment
Application	Not relevant	Not used	

 ISBN: 978-1-61208-164-9

TABLE II. SUMMARY OF TAGGING CONTEXT PROPERTIES

Tagging Context Properties	
Contact	As it is the user which makes the touch, it must have contact property to carry out the touch interaction.
Identification	This will be given to other non-user entities, which will be responsible for responding to the touch made by the user
Context	This can be one of two types: the context limited to the touch interaction, which is the information about the entities involved in the interaction, or the environmental context.
Services references	So as not to keep all the information in the touched entities, only these are placed.
Memory	Although our model will focus primarily on data, the memory property is implicit and necessary.
Processing	It processes the services references.
Communication	Linking the entities involved with the environmental infrastructure.

"Place" is an entity that will remain and will allow us to represent a part of a smart environment or even the entire environment, since it will have the capacity for self-inclusion, in which the place may contain entities (or even another place). Although, in the first instance, a user has the capacity not to associate the service offered by a place, this could be considered because, on touching the place it will provide the service, entering the smart environment and the user could even obtain the opening of the door. A summary and comparison of PICTAC and Dey entities is shown in Table 1.

Whenever a user touches any of the four entities described above, it will generate one of the four classes of touch interaction managed by PICTAC: user-place, user-device, user-object and user-user, as shown in Figure 1.

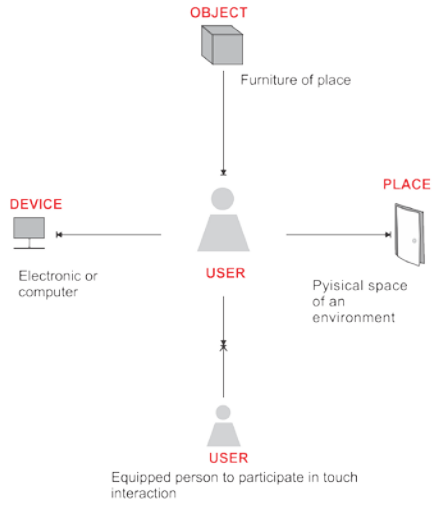


Figure 1. The four types of the PICTAC Model Touch Interaction.

IV. THE MODEL TO PERCEIVE TOUCH INTERACTION THROUGH THE TAGGING CONTEXT

The objective of the model that we are proposing is to be able to develop a system that endows the environment with the capacity to perceive the touch interaction using the *tagging context*.

The model consists of three parts: the properties that the environment must have, the tagging context and the services that can be offered.

A. *Environmental Requirements*

The PICTAC Model is designed in such a way that the system developed can easily integrate with other applications in a smart environment or it can be the only one of its type operating in the environment. To develop the system that manages the touch interaction and maximize the use of the information flow generated for the delivery of services, the environment must have two properties or infrastructures:

1. Processing and storage (considered together so that they can be offered by one technology)
2. Communication

Although these properties are essential to the development of the vision of an environment that manages the touch interaction, we believe that they will almost certainly exist in any environment in which an application of this type will need to be developed. This is partly due to the fact that the majority of workplaces have them and because the touch interaction is ideal for operating the electronic and computer devices in today's workplaces.

Both infrastructures are the common "technology level", which is shared by all the applications that will make up the ideal smart environment.

B. *Tagging Context*

If tagging is to put a tag on something and context is any information about entities, tagging context could sound incongruous but the intention of creating this paradox is to give emphasis to the idea.

Tagging will be necessary for two reasons: we will have to indicate the place where the entity must be touched and we must augment the entity with the capacity to perceive another entity's "touch".

The word context is used because we will augment the entities with the data needed to deliver the services originated by the touch interaction.

The tagging context requires generic properties to be developed: perception of touch, containment of information, processing and communication. These capabilities will be provided by different properties, of which all or some of them will be put in the entities. We define the tagging context as:

"Augmentation of the environment's entities with the necessary properties to participate in the touch interaction perceived by a system"

Two properties are necessary for the touch interaction's perception: contact and identification. As it is the user who makes the touch, it must have the contact capability. This is just a way of distinguishing the roles performed by each one

 ISBN: 978-1-61208-164-9

of the entities that touch. The identification capability will be given to other entities (non-user) which will be responsible for responding to the touch made by the user. In other words, the entity that is touched is the one which identifies (responds to the touch).

For an entity to contain information, it will require memory capability. Although our model will focus primarily on data, the memory capability is implicit and necessary. The data involve two properties: context and services references.

Context is the information needed to deliver the service and can be of two types: the context limited to the touch interaction, which is the information about the entities involved in the interaction, and the environmental context. The first is saved in the same entities as those involved in interaction and the second in the environmental infrastructure.

In order to deliver a service, some processing of information must be carried out and, because we are not certain where such processing is carried out (in the user, in the environment infrastructure or in a device with processing capability), it was decided that, instead of the entity that is touched containing all the necessary instructions for processing the data, it will contain the necessary services references. The services references are all the necessary information to process and deliver the service to the user.

C. Services

The reason for any device or object in the environment is to offer users, as a minimum, an intrinsic service and in order to obtain it the user must interact with the object. The object of PICTAC is to offer services from the environmental elements by means of the touch interaction and, together with these services, to deliver others in an implicit way.

The classification established for the services is based on the manner in which their execution is originated or who has put them. The intrinsic services of the elements in the environment (which can be natural to the item or established by the system) are the default services, which are received by the user who touches. The implicit services are those that accompany the above and can be received either by the user who touches or other related users. The optional services are those created by the PICTAC system that can be received by any user and put in any entity. Table 3 gives a summary with the characteristics of each one of the service categories.

1) Default Services

All the environmental elements with which people are surrounded have at least one function or main use for which they were designed. The model that we propose is designed so that such use or service will be delivered to the user at the touch of the entity. These services are called default services and will be delivered to the user automatically whenever it touches an element.

In a device, the default service is intrinsic to it and is easily identifiable (e.g., the printing of the printer, the display of the monitor, etc.).

In an object, default services do not exist because they are not inherently associated with any services that can be obtained through touch interaction (e.g., the desk). The object will only contain optional services.

In a place entity, the ideal default service, which the user would expect to receive, is to open the door, but its implementation will depend on the available technology; however, touching the place is essential for the user to "enter" or "exit" the application that manages the touch interaction and, should it fail to do so, it will not receive the service.

2) Implicit Services

These services are one of the benefits of implementing a PICTAC system. Sometimes and when the system has just been installed, they may not be easily associated by the user who carries out the touch. In other words, they are services under or attached to the default services.

Implicit services are developed to take advantage of the information flow that is generated when a default service is demanded and, like this one, can be delivered automatically.

All the services that are received either directly or indirectly by the related entities and the entities that participate in the touch interaction and/or are in the place where touch interaction is carried out will be implicit services.

3) Optional Services

These services will be put on entities by the users, depending on their requirements; they were originated to take advantage of the foreseeable and daily use of some environmental entities.

These services allow those elements that do not have an intrinsic service (the objects) to be regarded as part of the PICTAC system and offer services to the touch.

They can also be placed in any type of entity, the use of which can be considered structured. For example, when a user reaches its workplace, it will first touch the door of the building before entering its office; any user that pays a visit to the office must touch the door, etc. Thus, optional services can be placed in the entities that take advantage of such use.

Another example of an optional service is to leave a message to any or a specific user on the door of a building, on its desktop or any entity in respect of which we can be certain that it will be touched by the user to whom we wish to communicate something, or a note for a specific user when visiting the office in our absence, etc.

TABLE III. DESCRUIPTION OF SERVICES

PICTAC Services	
Default	The entities' intrinsic services
Implicit	These services accompany the default services
Optional	These will be put on entities by users in the case of elements that do not have intrinsic services

 ISBN: 978-1-61208-164-9

TABLE IV. PROPERTY-TECHNOLOGY CORRELATION

	Properties					
Technology	I	M	Cc	P	C	Interaction technique
Barcode	✓	×	×	×	×	Touch
Passive RIFD	✓	✓	✓	×	×	Not necessary
Active RFID	✓	✓	✓	×	×	Not necessary
NFC	✓	✓	✓	×	×	Touch
Bluetooth ≈ ZigBee	✓	×	✓	×	×	Not necessary
Wi-Fi	✓	×	✓	×	×	Not necessary
Infrared	✓	×	×	×	×	Movement

I=Identification
M=Memory/Context/Services References Cc=Contact
P=Processing C=Communication

V. MODEL TECHNOLOGY CORRELATION

The first method of validating the PICTAC model is to establish a correlation between its properties and the existing technologies with which they could be implemented. As some of these properties are matching, in this section we decided to consider them jointly in two cases: contact-identification and memory-context-services references. The four categories of model-technology correlation are explained below.

A. *Information Flow*

The automatic identification technologies make it possible to provide the environmental elements with contact and identification capabilities. Some of these technologies (voice, OCR, biometrics, etc.) do not have a relationship with this model, that is the reason why we decided to analyze three: Barcode, RFID and NFC.

Barcode technology has contact and identification properties, meaning that it can be used to add these properties to an entity that could be used in the contact interaction.

Near Field Communication (NFC) technology can establish a link when an initiator-reader is within two inches or less of a tag. This short distance gives the impression that the user is touching the tag (some would touch without problems). So the NFC is appropriate for providing these capabilities to entities in the environment.

B. *Memory-Context-Services References*

The memory property is essential for storing information whose significance is essential for our model because it will contain two properties: context and services references.

It is important to mention that although technologies exist, that by themselves do not provide these properties, they could be embedded in a device that has them. This is the reason why a technology cannot be discarded only on the basis of not having them and why devices in which it is possible for them to be embedded must be examined. Another important feature is that the memory must be re-writeable since the context is constantly changing.

The barcode is not an appropriate technology for this because of two reasons: saving information in a barcode is unworkable and it would not be possible to modify it.

The RFID and NFC technologies have the memory property by default in the tag, while in the reader the memory can be obtained from the device that controls it.

C. *Processing*

This property must be obtained from the device (or computational device that controls it) where the technologies that provide the other properties are embedded.

The power of this ability will be what determines the complexity of the services that can be offered, without depending on the environmental structure's processing power.

D. *Communication*

As in the case of the processing property, communication must be provided by technologies that are embedded in gadgets that provide other properties. To meet this requirement, we considered: Bluetooth and Wi-Fi. The major drawback of Bluetooth is that each connection requires the user's participation, whereas Wi-Fi technology only requires configuration of the first time that the gadgets are linked.

A summary (Table IV) of the analysis made in the three previous sections allows us to observe that none of the technologies considered has processing properties, which turns this into a requirement of the gadget that we use.

The bar code-contact only provides identification, so this can be ruled out.

NFC technology is the solution to the Bluetooth drawback, as it enables the automatic link via Bluetooth.

RFID technology is a good option for implementing the PICTAC system; unfortunately, although the costs of RFID and NFC tags are equal, prices of the antennas and the complete set of RFID make this unaffordable, although trends indicates that these will fall and once this is the case it will be possible to reconsider this option.

E. *Technological Suitability Model*

In order to adapt the PICTAC model to a proposal that combines different technologies for its implementation, we divided it into two main areas: the "tagging context" and the infrastructure environment. The latter is divided into two parts: storage-processing and communication. These and the tagging context are the three sections of the technological model, which can be summarized as shown in Figure 2 and is explained as follows:

- The gadget technology to implement the tagging context will endow the entities with the properties to participate in the touch interaction perceived by the PICTAC system: contact, identification, memory (which contains the context and services references), processing and communication.
- The communication section, which will allow entities to link with the processing and storage infrastructure in the environment.
- Processing and storage, defined as the distribution and operation of the computer equipment that will execute

 ISBN: 978-1-61208-164-9

the services, similar to the way in which they store the information from the environment (context) and process some services.

The "tagging context" properties can be satisfied by an NFC-enabled mobile phone and NFC-tag; the NFC-enabled phone also offers different communication alternatives (some models have Bluetooth, GSM, SMS, MMS, XHTML, SMTP, POP3, IMAP4 EGPRS and/or GPRS) and a large memory (up to 2 Gb).

The PICTAC model will use computer technology (with the communication, processing and storage properties) that can be found in most of the workplaces where people operate. This is established as a requirement for the environmental infrastructure because we consider it unattractive for the PICTAC system to be developed in an environment that does not have them.

VI. PICTAC VISUALIZATION

To formalize the relationships and elements that have been explained in previous sections that comprise the PICTAC model, we use UML class diagrams.

In the conceptual PICTAC model shown in Figure 3, the largest element is the smart environment that contains the PICTAC system, the normal environment, the communications infrastructure, processing, data storage, context and intelligent interfaces.

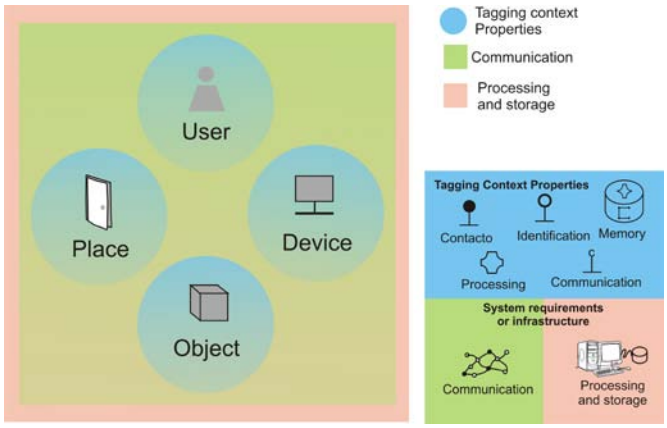


Figure 2. The Technological Sections of the PICTAC Model.

The intelligent interfaces are those involved in user interaction with environmental elements. The environmental elements include objects, places and devices, which together with the user compose the PICTAC model entities.

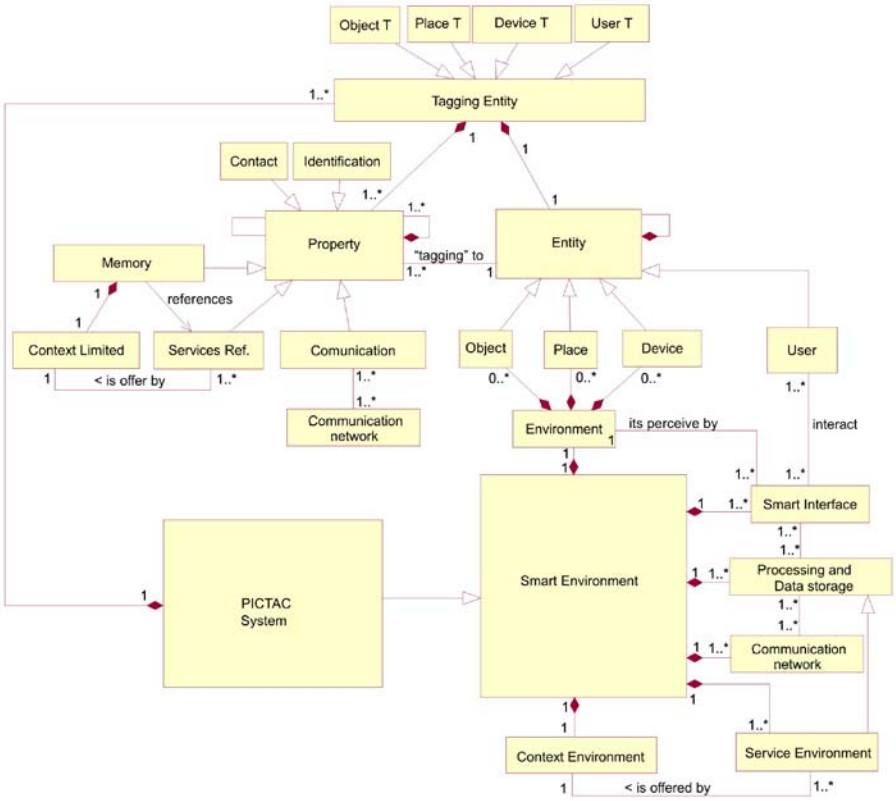


Figure 3. UML Description of the PICTAC Model.

A system that perceives touch interaction through tagging context consists of tagging entities in which a user touches the tagging entity for the purpose of obtaining services. This contact is perceived through properties that obtain and process the necessary data to deliver the service. This information and processing property may be contained in gadgets that can deliver services without needing the environmental infrastructure. However, the quality and quantity of these services depend on the technical characteristics of the technologies used to provide the properties. Within a communications infrastructure, processing and data storage are essential if the goal for the PICTAC system is to be part of an intelligent environment.

VII. CONCLUSIONS

The investigation was possible due to the technological innovations that are currently available; however, technology itself does not include the innovative uses that are involved in our research.

With the PICTAC model, when implementing a managed system, the touch interaction does not only benefit the user involved in the touch interaction but the benefits extend to all users of the smart environment.

NFC technology is an excellent tool to provide services through touch contact, with the real prospect of easier integration with the environmental infrastructure (owing to the fact that an NFC mobile with WiFi is being developed).

The incorporation of contextual information in the tag and the offer of various services is the main difference between our research and from others working with NFC technology.

We have developed the first phase of PICTAC system at the MAmI research group; the TIS (Touch Interaction Services). In the daily use of the system we have observed that putting several services in tags provides advantages but cannot be used in all situations.

Users gladly accept all the services if they save efforts but there are those who are obstinate in accepting the tags where there is no perception of savings. In TIS, when we put a tag in the door so that users may get into or leave the system; for users this represents an over-exertion and sometimes they forgot to do it. This perception of over-effort disappears when we install an NFC-enabled electronic door lock.

ACKNOWLEDGMENT

This work has been financed by Programa para el Mejoramiento del Profesorado (PROMEP) from the Secretaría de Educación Pública, México, and Fondo Mixto de Fomento a la Investigación Científica y Tecnológica CONACYT – Gobierno del Estado de Tamaulipas, México.

REFERENCES

[1] Cook, D. and S. Das, *Smart Environments : Technology, Protocols and Applications*. Wiley Series on Parallel and Distribued Computing, ed. A.Y. Zomaya. 2004. Hoboken, New Jersey: John Wiley & Sons, Incorporated.

[2] Song, C.-H., J. Wu, D.-H. Seo, and W.D. Lee. *Solving multi-sensor problem with a new approach*. in *First International Conference on the Applications of Digital Information and Web Technologies, ICADIWT 2008.* pp. 348-353. Ostrava, Czech Republic.

[3] Klima, M., L. Lhotska, V. Chudacek, M. Huptych , and M. Husak. *Assistive Technologies: New Challenges for Education* in *4th European Conference of the International Federation for Medical and Biological Engineering*. pp. 2730-2733. 2008. Antwerp, Belgium.

[4] Weiser, M., R. Gold, and J.S. Brown, *The origins of ubiquitous computing research at PARC in the late 1980s.* IBM Systems Journal, 1999. 38(4): pp. 693-696.

[5] Aarts, E. and L. Appelo, *Ambient intelligence: thuisomgevingen van de toekomst.* IT Monitor, 1999(9): pp. 7-11.

[6] Dey, A.K.: *Providing Architectural Support for Building Context-Aware Applications*. Phd Thesis, Georgia Institute of Technology (2000).

 ISBN: 978-1-61208-164-9

Context as an IMS Service

Filipe Cabral Pinto[1,2], António Videira[1], Manuel Dinis[3]

[1]Portugal Telecom Inovação S.A., R. José F. P. Basto, Aveiro, Portugal
[2]Queen Mary University of London, Mile End Road, London E1 4NS, UK
[3]Inovetel, Rua Amilcar Cabral nº 54-c 1º dto., Luanda, Angola
{filipe-c-pinto,antonio-p-videira,mdinis}@ptinovacao.pt

Abstract — **The worldwide spread of mobile phones and its increasing trend makes them the main vehicle for mobile communications. Mobile Operators can provide valuable services to their clients by using context information to personalize multimedia content distribution. The knowledge of users' situations shall guide applications to select the most helpful content in each specific instant. This paper proposes IMS (IP Multimedia Subsystem) as a context service enabler allowing Mobile Operators to offer more useful services to their clients by applying the Internet of Things vision into the telecommunications world.**

Keywords: Context; IMS; IoT; Mobile; Services.

I. INTRODUCTION

A core challenge to Mobile Operators is to afford clever services based on the user's current situation in order to make them cheaper and useful. The mobile devices explosion and the wireless networks diffusion allow users to be always connected enabling a ubiquitous access. Furthermore, it is expectable that sensor networks technologies will be fast widened all over the world fostering the Internet of Things (IoT) accomplishment. The information gathered from these sensors networks can be treated as context information, which enables the usage of intelligent models that can be employed by context-aware services to improve their utility. Mobile Operators require an effective communication framework that allows context information to flow from their providers towards the context consumers making possible to get better services and personalized contents distribution.

This paper devises IMS (IP Multimedia Subsystem) as a context service facilitator enabling Mobile Operators to provide valuable services to end-users. It is here proposed context as an IMS service that instead of delivering multimedia content makes available context information to context-aware services allowing Mobile Operators to offer the right content to the right client.

The rest of the paper is as follows: in Section II the main motivations for context usage are referred; Section III presents the devised framework that allows Mobile Operators to use context information to provide useful services to their clients; in Section IV is introduced the procedures that enable the context exchange between interested entities by showing the proposed demonstration scenario. Finally, Section V summarizes the main conclusions.

II. MOTIVATION

A. Context in Mobile Services

Context can be understood as sensed information that changes over time, which can be used by context-aware systems to improve their performance. Nowadays, Mobile Operators are already offering some context-aware services based on users' location by selecting the content that best fits the user position. But Mobile Operators can go further since they are in a privileged position to collect information about their clients' situation. They can bring together statistics on their clients most accessed web pages or their favorite programs and channels; they can even collect figures about their online shopping profile. Furthermore, following the IoT fashion, through the use of external sensoring, it is possible to collect extra context information, like noise, temperature or movement. All this information is critical to Mobile Operators since it allows them to offer useful services through optimized networks to end-users. Figure 1 presents possible sources of context information.

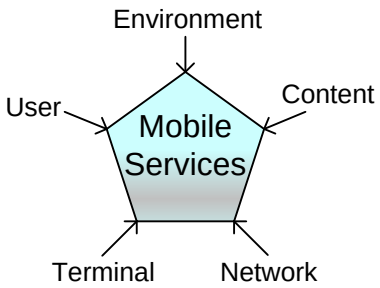


Figure 1. Possible Context Sources

B. Service Scenario

This is the raining season. A huge storm is flooding the regional roads letting the population out-of-the-way. Unhappily, there is often a lot of kills victims of these natural catastrophes and, most of them, are tourists caught napping in the storm. In order to reduce their impact, the Government is funding Mobile Operators to offer a service that streams videos indicating to each of their clients in critical situation the best way to reach a shelter having food and potable water. The name of the service is: Get safe.

Mobile Operators using their native capabilities can imprecisely check their clients location based on their attached antennas. If they are nearby a flooding area, Mobile Operators check which ones have on their terminals the Global

 ISBN: 978-1-61208-164-9

Positioning System (GPS) and try to immediately get the right users' location in order to provide the most accurate location-based helping video. Due to those users, Mobile Operators require context information and, consequently, they become context consumers of a context service that converts user geographic coordinates into user street location. The geographic coordinates are supplied to the Context Service by the context provider based on the clients GPS information. Furthermore, Mobile Operators makes even use on their clients profile in order to check their native language.

Based on users' location and using their specific language, Mobile Operators can offer understandable videos with clear indications about the way to reach a shelter saving hundreds of lives.

C. Related Work

The use of context in mobile networks is nowadays a key research topic in scientific community. Its usage in telecommunications systems allows the service personalization over optimized networks.

The FP7 (Framework Program 7) SENSEI (Integrating the Physical with the Digital World of the Network of the Future) project aimed to provide the necessary network and information management services to enable consistent and accurate context information retrieval and interaction with the physical environment [1]. The work carried out in [2] has proposed a set of interfaces enabling Applications and Services to access real world information.

PERSIST (Personal Self-Improving Smart Spaces) was a FP7 project intending to develop Personal Smart Spaces (PSS) that were able to learn and reason about users, their intentions, preferences and context [3]. The PSS was enabled to extend and enhance as the user meets other smart spaces. In [4] is presented the overall design of the main components supporting the operation of Personal Smart Spaces.

The FP7 C-CAST (Context Casting) project aimed at evolving mobile multimedia multicasting to exploit the increasing integration of mobile devices with our everyday physical world and environment [5]. Furthermore, it has designed an innovative approach allowing personalized content delivery to multiple mobile users independently of the underlying networks. In [6] and [7] was defined an architecture to support a complete context management functionality along with service components like group management and content selection. The work carried out in [8] and [9] has defined a framework to collect sensor data, distribute context information and manage efficiently context aware multiparty data distribution. In [10] and [11] were developed a set of mechanisms for autonomous context driven content creation, adaptation and media delivery.

The research presented in [12] demonstrates the usefulness of context-awareness usage to improve the interface between the user and the mobile devices. The context is derived from the fusion of multiple sensors. The article also proposes a hierarchically working model to structure the context concept.

In [13] was proposed a layered conceptual architecture where the layers were increased with interpreting and reasoning functionalities, which allows the detection and the context usage.

Another research has been carried out in [14] where a layered conceptual design framework was proposed highlighting the different elements common to most context-aware architectures.

The research done in [15] has studied the impact of context, sensors and wireless networks in the telecommunications field. Several scenarios were suggested stressing the possible synergies between the defined areas.

In [16] is devised a convergent context-aware architecture where IMS is used to convey context information and to control MBMS (Multimedia Broadcast and Multicast Service) and E-MBMS (Evolved MBMS) multimedia channels allowing Mobile Operators to easily offer innovative services over efficient networks.

The work introduced in [17] exposes the benefits of using context information for the vertical handover decision procedures. It devises a context-aware information server skilled to manage dynamic information retrieved from both the network and the terminal side entities, which leads to an improved handover process.

III. IMS AS A CONTEXT FACILITATOR

A. IMS Architecture

3GPP (3rd Generation Partnership Project) Release 5 has introduced IMS as an extension of the UMTS (Universal Mobile Telecommunication System) architecture [18]. It has added a set of new functions linked by new standardized interfaces. IMS uses the IETF (Internet Engineering Task Force) SIP (Session Initiation Protocol) in order to manage multimedia sessions [19]. It provides QoS (Quality of Service) by means of resource reservation and allows operators new charging schemes for multimedia sessions. Finally, IMS makes possible fast service deployment enabling more and better services to end customers. The IMS architecture, as defined by 3GPP in [18], is presented in Figure 2.

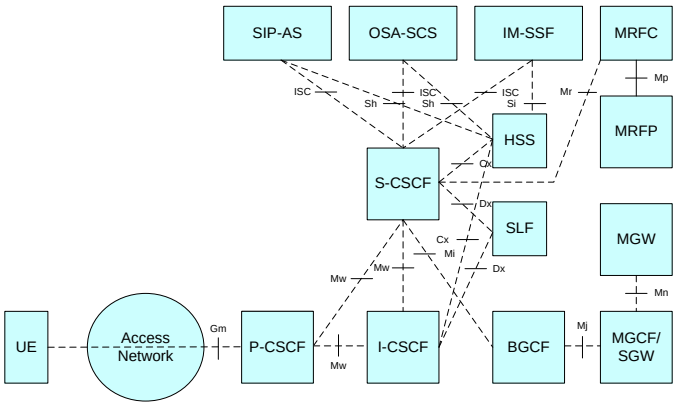


Figure 2. 3GPP Release 5 IMS Architecture

B. *Context as a Service*

Context is defined in [20] as the information employed to characterize the situation of entities. A situation is defined in [21] as the state of a context at a certain point in space at a certain point in time, identified by a name. A situation is even matched up to a snapshot taken by a camera where it captures the momentary profile of the context attributes. A service can be considered context-aware if it uses context information to adapt its behavior to optimally perform its tasks.

Context-aware services require services that provide context-aware information. Consequently, context-aware services need context services to afford users' situation information. A Context Service (CxS) can be defined as a service that makes available context information to context-aware systems [16]. CxS uses context information provided by Context Producers (CxP) and presents users' related situation to Context Consumers (CxC). So, CxSs gather context information from CxPs and make it available to CxCs.

CxS can have different complexity behaviours. It can just bypass basic sensor information, such as the user location, or it can apply demanding algorithms based on mixed context sources in order to get more complex information such as users' wishes or needs.

CxS can collect context information from different sources. The information can then be filtered to select only the valid data. Additionally, fusion processes can be applied enabling the consistent values attainment through the integration of similar data sources. Aggregation mechanisms can run over heterogeneous sources of information to obtain higher level context data. As a final point, reasoning mechanisms allow CxS to wrap up users' situation by applying logical rules over collected context information.

CxC, CxS and CxP can exchange context information by request (where is the client now?), periodically, (where is the client in each 5 minutes?) or on an event-based (let me know when the client leaves the train station). The functional interaction can be seen in Figure 3.

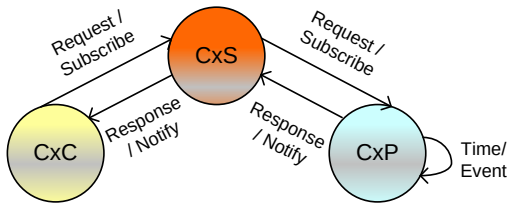


Figure 3. CxC, CxS and CxP interactions

C. *Context as an IMS Service*

One of the main IMS advantages is the easy development of multimedia services. Following the same approach, IMS can be very useful as a context service facilitator enabling awareness in mobile services. This analogy is presented in Figure 4.

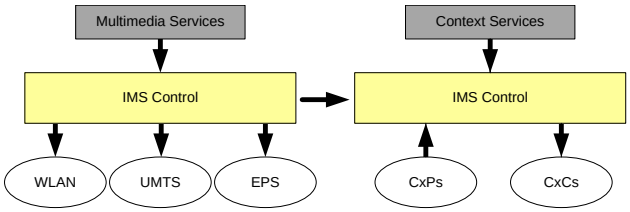


Figure 4. IMS Moving Towards a Context Enabler

This paper proposes to have Context Services playing de role of IMS specific applications that make the bridge between context producers and context consumers. It is here devised IMS as a context facilitator in which the IMS Application Servers are used to manage context instead of the standard multimedia services. The proposed framework is context agnostic being capable of controlling all types of context information. CxS shall collect context information from CxP, then they shall apply their own logic and finally they shall provide inferred context information to CxC. This process shall take place independently of the type of information exchanged between all context entities.

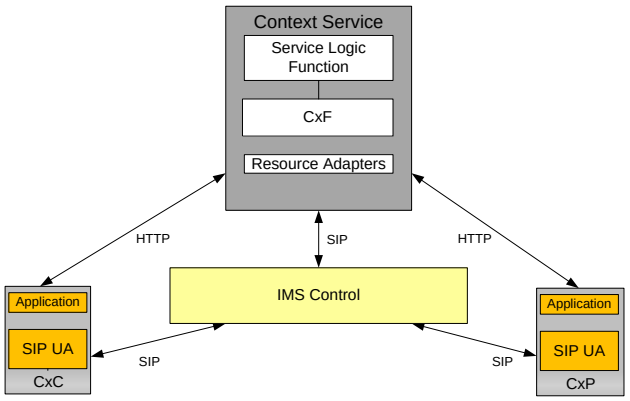


Figure 5. Context Framework

CxS can be further elaborated by splitting the application logic from the Context Function (CxF) that has functional roles allowing applications to bypass the details of communications with CxC and CxP. For instance, all the management of context dialogs established between CxP, CxS and CxC shall be transparent to applications. Also, Resources Adapters are fundamental since they hide the signaling details to the higher layers. Consequently they enable both SIP signaling for context transfer and HTTP for, for instance, service generic information presentation or subscription purposes.

CxP and CxC shall be considered as SIP UA (User Agents) that are able to register under IMS networks and to manage SIP specific signaling. They shall have their own logic allowing the context information transfer between involved entities. As stated in [18], SIP is the protocol adopted for session control in the IMS systems. Therefore, it makes sense to use it for an IMS-based context management. The context framework can be seen above in Figure 5.

IV. SCENARIO DEMONSTRATION

The use of IMS as a context enabler allows the fast context services creation. An instance of the signaling procedures for

 ISBN: 978-1-61208-164-9

the scenario suggested in Section II.B using the framework proposed in this paper can be seen in Figure 6. The signaling is detailed below.

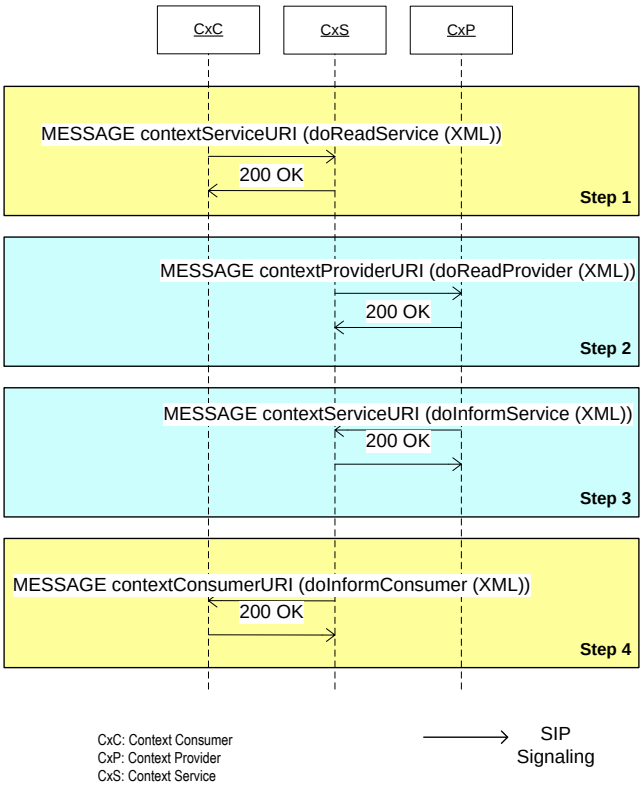


Figure 6. Signalling Flows

A. Step 1

A Mobile Operator (CxC) providing the "Get Safe" service intends to obtain immediately the exact street address of one of its clients in order to provide him the most appropriated video to let him reach the shelter. For that it sends a SIP MESSAGE towards the CxS encompassing a XML file describing the user identification, which in this case is the user SIP address. The XML file sent by the Mobile Operator is presented in Figure 7.

```
<?xml version="1.0" ?>
<!DOCTYPE hdr SYSTEM "MLP_HDR_200.DTD">
<svc_init>
    <hdr ver="2.0.0">
        <client>
            <id>ptin</id>
            <pwd>test</pwd>
            <serviceid>Get Safe</serviceid>
            <servicetype type="PASSIVE"/>
        </client>
    </hdr>
    <slir ver="2.0.0" res_type="PERSISTENT">
      <msids>
        <msid type="URI">bob@ptin.pt</msid>
      </msids>
    </slir>
</svc_init>
```

Figure 7. Request XML file sent from CxC towards CxS

B. Step 2

The CxS after validating the request coming from the CxC forwards it towards the CxP using a SIP MESSAGE where the XML file is included, but instead of requesting the street address, it will request the geographic location.

C. Step 3

The CxP checks the mobile identity and uses its own mechanisms to get the user geographic location. The information is compiled in the XML file which is sent towards the CxS inside a SIP MESSAGE, as presented in Figure 8.

```
<svc_result>
  <slia ver="2.0.0">
    <pos>
      <gsm_net_param>
        <cgi>
          <cellid>35086</cellid>
          <lac>305</lac>
        </cgi>
      </gsm_net_param>
      <msid type="URI">bob@ptin.pt</msid>
      <pd>
        <alt>87.145419</alt>
        <shape>
          <circle>
            <point>
              <ll_point>
                <lat>38.741602</lat>
                <long>-9.197874</long>
              </ll_point>
            </point>
            <rad>791.795320536137</rad>
          </circle>
        </shape>
        <time utc_off="+0000">20101215152929</time>
      </pd>
    </pos>
    <result resid="0">OK</result>
  </slia>
</svc_result>
```

Figure 8. XML file sent from CxP to CxS including user geographic location

D. Step 4

The CxS shall use its own mechanisms to translate geographic location into addresses location. After doing the conversion, CxS sends the answer back towards the CxC, which in this case is the Mobile Operator, providing the Get Safe service. A SIP MESSAGE is sent containing the XML file described in Figure 9. Having the right user location, Mobile Operators can now select the video that best helps the user running away from the flood.

```
<?xml version="1.0" ?>
<get_place_result>
     <content type=street>
        Avenida Comandante Jika
     </content>
</get_place_result>
```

Figure 9. XML file sent from CxS towards CxC including the street address

V. CONCLUSIONS AND FUTURE WORK

This paper has devised a framework where IMS works as a context enabler allowing Mobile Operators to offer useful services to their end-clients. The context information usage enables the selection of content that best matches the user

 ISBN: 978-1-61208-164-9

situation making possible to have personalized content distribution. A service scenario was here presented exposing the potentialities of the Internet of Things usage in the mobile services domain.

System performance evaluation is envisaged as future work. A particular attention will be given to the delays introduced by the context management. Moreover, the impact of the additional signaling introduced by the context information transport will be analyzed in the IMS-based context-aware system.

REFERENCES

[1] http://www.ict-sensei.org/ [Accessed 2 June 2011]

[2] SENSEI, Deliverable D3.2, "Reference Architecture", January 2009

[3] http://www.ict-persist.eu/ [Accessed 2 June 2011]

[4] PERSIST, Deliverable D3.1, "Detailed design for personal smart spaces", March 2009

[5] http://www.ict-ccast.eu/ [Accessed 2 June 2011]

[6] C-CAST, Deliverable D6, "Requirements and concepts for context casting service enablers and context management", November 2008

[7] C-CAST, Deliverable D12, "Specification context casting service enablers, context management and context brokering", June 2009

[8] C-CAST, Deliverable D7, "Requirements and concepts for context detection and context-aware multiparty transport", November 2008

[9] C-CAST, Deliverable D13, "Specification of context detection and context-aware multiparty transport", June 2009

[10] C-CAST, Deliverable D8, "Requirements and concepts for content casting", November 2008

[11] C-CAST, Deliverable D14, "Specification of content casting", June 2009

[12] Albrecht Schmidt, Michael Beigl, and Hans-W. Gellersen, "There is more to Context than Location", Computers and Graphics, Volume 23, Pages 893—901, 1998

[13] H. Ailisto, P. Alahuhta, V. Haataja, V. Kylloenen, and M. Lindholm, "Structuring context aware applications: Five-layer model & example case", Concepts & Models for Ubiquitous Computing (UbiComp 2002), Sweden, 2002

[14] M. Baldauf, S. Dustdar, and F. Rosenberg, "A Survey on Contextaware Systems", International Journal of Ad Hoc and Ubiquitous Computing, Volume 2, Issue 4, Pages 263–277, June 2007

[15] R. Aguiar and D. Gomes, "Quasi-omniscient Networks - Scenarios on Context Capturing and New Services through Wireless Sensor Networks", Wireless Personal Communications, Springer Volume 45, Pages 497-509, June 2008

[16] F. Cabral Pinto, A. Videira, N. Carapeto, and M. Dinis, "Context-aware Multimedia Distribution for Multiparty Communications", 6th International Mobile Multimedia Communications Conference (MOBIMEDIA 2010), Lisbon, Portugal, September 2010

[17] P. Neves, J. Soares, S. Sargento, H. Pires, and F. Fontes, "Context-aware media independent information server for optimized seamless handover procedures", Computer Networks: The International Journal of Computer and Telecommunications Networking, Volume 55, Issue 7, Pages 1498-1519, May 2011

[18] 3GPP TS 23.228 V5.14.0 (2005-09), IP Multimedia Subsystem (IMS); Stage 2, (Release 5)

[19] IETF RFC 3261, SIP: Session Initiation Protocol

[20] K. Dey, "Understanding and Using Context", Personal Ubiquitous Computing, Volume 5, Issue 1, Pages 4-7, February 2001

[21] A. Zimmermann, "Context Management and Personalisation", PhD Thesis, University of Aachen, 2007

 ISBN: 978-1-61208-164-9

Challenges in Building a Mobile SpeechWeb Browser

Randy J. Fortier
School of Computer Science
University of Windsor
Windsor, Canada
rfortier@uwindsor.ca

Richard A. Frost
School of Computer Science
University of Windsor
Windsor, Canada
richard@uwindsor.ca

***Abstract*—Mobile devices support speech interfaces that are inadequate for working with web applications. Providing speech-based interaction with web applications from a mobile device opens up a world of computing to users on the go. We have developed a mobile browser that allows users to easily interact with SpeechWeb applications from their mobile device in a hands-free mode. To overcome the limitations of mobile devices, a new method of parsing speech has been developed. The result is a more accessible web, allowing users to do some of the more interactive tasks, even while traveling, that traditionally required a personal computer.**

Keywords-SpeechWeb; speech web; speech recognition; voice recognition; speech applications; web applications; mobile application; natural language processing.

I. INTRODUCTION

A. An Introduction to SpeechWeb Applications

Speech applications have recently been the subject of increasing interest. This is especially true for mobile devices where traditional input mechanisms are limited and output displays are inconveniently small. On mobile phones, which are often used while driving, using a keyboard or touchscreen for input and/or a display for output are both impractical and dangerous. Speech applications are common for dialing phone numbers, obtaining driving directions, sending text messages, and checking E-Mail. However, there are still many applications, such as web applications, that do not have a speech interface on mobile devices.

The web browser grew quickly during the 1990s to become the most popular Internet application. This growth was fostered by the ease of development of web sites, through standardized mechanisms. Even non-programmers could develop a web site to share their ideas. During the following decade, web sites evolved into interactive web applications, due mainly to the fact that anyone could develop those applications on any platform using any programming language. In the current decade, applications and data are migrating to the cloud, and many of these applications are web applications.

A SpeechWeb is a hyperlinked set of web pages with speech interfaces. SpeechWebs allow interrelated pages to connect. The traditional web and a SpeechWeb could be used to represent a semantic web. Speech websites are websites with speech interfaces. Speech websites could be speech interfaces to existing web pages, or specialized speech applications with a web interface. An important consideration for building a semantic web is to remove emphasis on notation and make it easier for people to develop their own sites. A SpeechWeb should facilitate the creation of sites with speech interfaces, with a broad range of complexity, in any programming language. By harnessing the same simplicity of the world wide web in the 1990s, a SpeechWeb can experience similar growth.

As an example, consider a mobile user driving to work. Current technology allows them to be entertained, make phone calls, send text messages, and navigate, all using hands-free interaction. Through SpeechWeb applications, these users could also obtain other information, including data stored in the cloud, and conduct transactions online using their voice. For example, a user could search for and hear restaurant reviews, translate French to English, or learn geographical statistics about a new state or province they have just entered (e.g., What is the population of the state of Michigan?). Google voice search implements a SpeechWeb application specifically for search, and is widely available for mobile devices [8].

The PipeBeach project [6] tried to merge the existing standards of VoiceXML and WML. At the time, WML was becoming popular as a platform for the mobile web. Since then, WML has declined in popularity, due to the increased power of mobile devices. The idea of combining a visual markup language and a speech markup language to create a multi-modal interfaces has also been used within traditional browsers [4]. The PipeBeach project has since been discontinued. Our project attempts to create a SpeechWeb browser that uses voice as its primary input and output mechanism. The w3voice project [7] is another attempt at creating a SpeechWeb infrastructure. This project uses an architecture that results in significant data transfer, and heavy processing loads on the server tier.

B. Limitations of Speech Interfaces

Most current speech interfaces to web applications use a screen-scraping approach. They simply read the text on a web page and provide mechanisms for the user to skip paragraphs, follow links, and fill out forms. Finding answers to a question can be very difficult. Speech interfaces to many of the existing web applications are not sufficient.

 ISBN: 978-1-61208-164-9

C. Limitations of Mobile Devices

The limitations of mobile devices are, for the most part, widely known. The screens are small, and current implementations can be difficult to read in bright light. Also due to the small screen, interacting with the touch screen can be error-prone. Keyboards, if available, are inconvenient, error-prone, and not ergonomically designed.

While download rates are catching up to home Internet access speeds, users still want to limit data transfer, due to very high costs on typical mobile plans. Browsing the traditional web results in a very large amount of data flow. Mobile-optimized websites often do not limit data flow, but merely use style to re-structure the information visually.

Speech interfaces also have limitations on mobile devices. Processing, memory, and disk space are limited. Voice recognition can test these limits. Even more significant, mobile devices often do not support grammar-based recognition, but only dictation. Dictation, while more flexible, is the less accurate approach, since any word in the target language could be spoken at any moment. Grammar-based recognition can improve recognition accuracy by narrowing the possible spoken words based on what is acceptable according to the grammar, and therefore relevant to the application.

D. Accessing Speech Applications with Mobile Devices

Our efforts have been focused on making SpeechWeb applications easier to create, much as web applications and websites were easy to create in the early days of the world wide web. As a result of our research into these problems, we have produced the following:

- A working prototype of a mobile SpeechWeb application browser.
- A method for improving the recognition accuracy of dictation-based voice recognizers.
- An XML-based language for describing SpeechWeb application interfaces, called SWML.
- A framework that allows programmers and non-programmers alike to create SpeechWeb applications.

E. Outline of this Paper

This paper first provides an overview of our approach. We describe an architecture suitable for a SpeechWeb. We describe an example language, SWML, that could be used for creating SpeechWeb applications. We discuss a method for overcoming the limitations of dictation-based speech recognition, common in mobile devices. Finally, we analyze our approach, and discuss related work.

II. Our Approach

Our research has developed a SpeechWeb browser for mobile devices, using the Android platform. This application is designed to take advantage of the mobility and specialized hardware provided by these devices, and account for their limitations. This browser provides a client for interacting with SpeechWeb applications. SpeechWeb applications can be designed to take advantage of the speech interface, and minimize the effort required for the user to carry out their desired tasks.

As an example, consider a encyclopedic application. There are several such applications on the traditional web. By integrating a speech browser with an application that uses natural language processing, and semantic evaluation, the user can merely ask a question and immediately obtain the answer. Other application tasks, such as filling out registration forms and conducting banking transactions can also be accommodated through a question/answer approach. Simpler applications, analogous to the websites created by non-programmers in the 1990s, are also possible, provided that there are simple, yet flexible, tools to write them. Applications that are more complex are also possible.

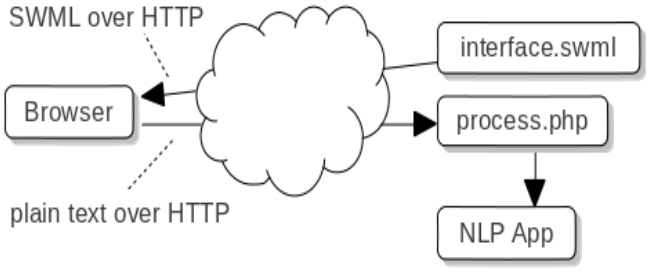


Figure 1: An overview of the SpeechWeb system

With the goal of making SpeechWeb applications easy to develop, an XML-based language for creating the speech interfaces for web applications, SpeechWeb markup language (SWML), has been created. This language allows SpeechWeb applications to be created in a similar manner to how web applications are created.

The SpeechWeb application browser also mitigates the inaccuracy typical of dictation-based voice recognition software by post-recognition filtering using a grammar. While this approach doesn't provide the accuracy of traditional grammar-based recognition, it does improve recognition accuracy significantly.

Our group has also developed a SpeechWeb application infrastructure, which should help users create their own SpeechWeb applications, even if they do not understand programming languages [1,2]. Our work into natural language processing has also produced several interesting applications, which have been integrated into the SpeechWeb.

III. Details

A. SpeechWeb Architecture

Previously, it has been proposed by our research group that SpeechWeb applications use a local recognition, remote processing (LRRP) architecture for transmitting data to a

 ISBN: 978-1-61208-164-9

speech application [1]. In this model, all voice recognition occurs on the client side. This was motivated by the fact that the raw audio data would be very large, and yet most user devices are rich clients. On mobile devices, this architecture is even more appropriate due to typical users' desire to limit data flow. Performing the voice recognition on the client-side results in less data flow, since the output of voice recognition is much more compact plain text. A comparison of these speech application architectures is given in Figure 2.

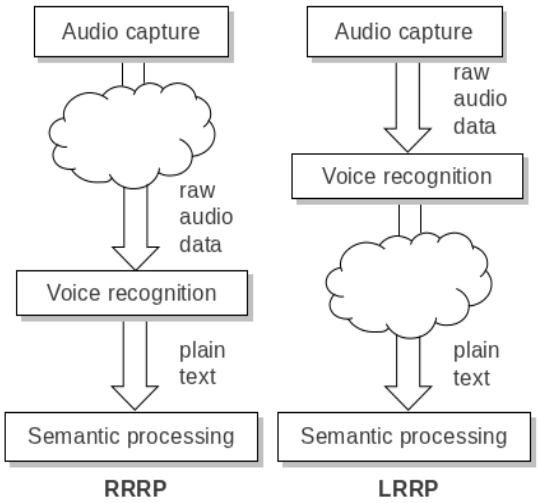


Figure 2: A comparison of RRRP and LRRP architectures

In an LRRP web application, the voice recognition occurs on the client device. This process converts the raw audio data into plain text. This plain text is transmitted to the server side of the web application, in a similar fashion to how data is sent to the server side in an web 2.0 web application via AJAX. This approach has the following significant benefits:

- Data flow is typically reduced by a far greater amount than would be possible through compression alone
- Server-side processing requirements for a server side SpeechWeb application under heavy load are significantly reduced

This approach is in contrast to that used by speech-based calling centres, which use voice recognition to allow users to navigate through a complex telephone system. These calling centres use remote recognition, remote processing (RRRP), increasing the processing requirements on the server side. For SpeechWeb applications, where data would be labeled as data by the mobile provider, the cost of transferring the large amount of audio data would also be a significant problem.

B. SpeechWeb Markup Language (SWML)

SWML is an XML-based language, not unlike HTML for traditional web applications. The purpose is to convey information to users, and provide means of collecting information from them. The XML syntax is familiar to many existing web developers. As it can also be easily parsed by software, tools can be developed to make development even easier for non-technical users.

For a file format for SpeechWeb applications, there are three primary requirements. First, there should be away of welcoming the user and giving them instructions on how to use the SpeechWeb application. Second, there needs to be a way to specify what are the valid expressions of the application's language, often expressed as a grammar. The main purpose of the grammar on the client is to improve the accuracy of speech recognition. Third, there needs to be a way to determine how the SpeechWeb application should respond to inputs. For example, when the application has collected sufficient data, it should send that data to the server for processing.

VoiceXML [3] is a similar file format to SWML. Both formats are largely interchangeable, but the main advantage of SWML over VoiceXML is its simplicity. A simple file format, similar to HTML in the 1990s, should make it easier for non-programmers to develop SpeechWeb applications. However, for more interactive applications, VoiceXML would be preferred. One advantage of VoiceXML over SWML is that VoiceXML documents can specify, using JavaScript, how to respond to inputs using client-side code. This capability allows AJAX-like interaction in a SpeechWeb application.

The existing format for SWML is a simple, proof-of-concept file format, which allows developers of a SpeechWeb application to specify:

- An initial prompt, which gives the user an idea of what they can do with the application
- The conditions that must be satisfied in order for data to be submitted to the server side of the SpeechWeb application
- The grammar rules, which describe the sort of sentences the user can say

Shown in Figure 3 is a simple SWML document, which could be used by a non-technical user for a simple question/answer SpeechWeb application. The grammar can simply describe the possible questions that are valid.

The prompt is used to introduce the user to the SpeechWeb application. It can be used simply to greet the user, or it could be used to provide instructions on how to use it. Existing prototype applications use relatively short prompts, in keeping with the minimized data flow policy.

The submit condition describes when the browser has collected the right amount and type of data to send to the application. In the example in Figure 4, a sentence is the unit of data transfer. Once the browser gathers a complete sentence, defined in the grammar itself, it sends that data onto the server to the URL specified. This server side of the application can answer the question itself, or act as a web portal to a non-web application. A submit condition can also include boolean expressions, which is useful for surveys

 ISBN: 978-1-61208-164-9

where multiple questions need to be answered before the results are submitted to the server tier.

The grammar substructure describes a complete context-free grammar. Our browser uses this grammar to improve the accuracy of its voice recognition. In addition, the browser also supports local parsing and optional client-side transformation. Terminal symbols are described using a traditional rule, containing a name (category) and its definition (rhs). The definition is often a single word, but can be phrases or sentences in more trivial applications. If there is more than one terminal rule for the same category, each of the rules are treated as *or conditions*.

For non-terminals, nearly the same format is used. The category names the non-terminal, and the rhs describes a sequence of terminals and non-terminals, separated by a space. This sequence describes a single rule for that non-terminal. If there are multiple rules for the same category, each is treated as an *or condition*. An optional transform can be added, which describes how the expression should be modified before being submitted to the server tier. This eliminates the need, in some applications, where the server tier requires the expressions to be transformed in some way, such as parenthesization. This is optional, since some applications have no need for such transformations, and other applications perform the transformations on the server tier.

```
<speechweb>
 <prompt>
   Welcome to the solar system encyclopedia.
 </prompt>

 <submit-condition url="/simple.php">
   <any-of options="question" />
 </submit-condition>

 <grammar>
   <terminal category="sentence"
          rhs="what is your name" />
   <terminal category="sentence"
          rhs="what is your favourite colour" />
   <terminal category="sentence"
          rhs="what is your favourite food" />
   <terminal category="sentence"
          rhs="how old are you" />
   <non-terminal category="question"
          rhs="sentence" />
 </grammar>
</speechweb>
```

Figure 3: A Simple SWML Document

The submit-condition can be extended to accommodate less trivial conditions. For example, if the SpeechWeb application asked a question, the user could answer the question directly, or perhaps ask a question of their own. A SpeechWeb application could be designed to conduct an oral survey with the user, requiring that some or all questions have been answered. The example in Figure 4 illustrates a simple SpeechWeb application that requires the user to make statements of a litigious nature. The submit-condition can contain arbitrarily nested and and or conditions to facilitate both of these scenarios.

```
<speechweb>
 <prompt>
   Welcome to the solar system encyclopedia.
 </prompt>

 <submit-condition url="/submit.php">
   <all-of options="s1,s2" />
 </submit-condition>

 <grammar>
   <terminal category="s1" rhs="I hereby
acknowledge that the work being submitted is my
own work" />
   <terminal category="s2" rhs="I hereby deny
that I have submitted all or part of this work
for an assignment in another course" />
 </grammar>
</speechweb>
```

Figure 4: An SWML Document with a Non-trivial Submit Condition

Shown in Figure 5 is another sample SWML document. This example illustrates a context-free grammar for a very small subset of English. In this case, there are a set of terminal and non-terminal rules describing the valid sentences of the language. In this example, the grammar is a context-free grammar, used for a natural language processing application. For simpler applications, the grammar could be as simple as a list of acceptable phrases or questions.

In the example in Figure 5, the grammar has also been augmented with rules for transforming the English sentences into sentences marked up with parse information (e.g., planet → COMMON_NOUN(planet)). In this example speech application, to which we're sending the user's utterance, some words have been defined as semantic functions, so the user's phrase has been modified to have the expected syntax. In this example, the vp (verb phrase) function returns a list of matching objects (for example, a list of objects that rotate). The np (noun phrase) function filters this list according to its noun (for example, earth might limit the results to include – at most – earth).

```
<speechweb>
 ...omitted for brevity...
 <grammar>
   <terminal category="pnoun" rhs="luna" />
   <terminal category="pnoun" rhs="earth" />
   <terminal category="tverb" rhs="orbits" />
   <terminal category="iverb" rhs="rotates" />

   <non-terminal category="np" rhs="pnoun" />
   <non-terminal category="vp" rhs="iverb" />
   <non-terminal category="vp" rhs="tverb np"
         transform="^tverb ( ^np )" />
   <non-terminal category="s"  rhs="np vp"
         transform="^np ( ^vp )" />
 </grammar>
</speechweb>
```

Figure 5: An SWML Document with Transformation Rules

C. *Improving Recognition Accuracy*

Grammar-based voice recognition is more accurate than dictation-based voice recognition for one simple reason: The grammar is used to limit the possible words that can be accepted. Consider the grammar in Figures 5 and 6. If the

 ISBN: 978-1-61208-164-9

user says "mars," it matches an np (noun phrase). If parsing an s (sentence), we next expect a vp (verb phrase). In this simple grammar, only two words are allowed next (orbits and rotates). As this is a simple example, such results should not be expected in the general case, but the reduction is still significant. With fewer choices available, the voice recognizer has a better chance of being right about what was uttered.

Unfortunately, many mobile devices do not support grammar-based voice recognition. To get past this limitation, we can apply the grammar after recognition for dictation-based voice recognizers that support the return of a list of possible phrases (often including probabilities for each phrase). Any phrases that do not follow the syntax established by the grammar are eliminated as options, and the remaining phrase with the highest probability is chosen. This process is illustrated in Figure 7.

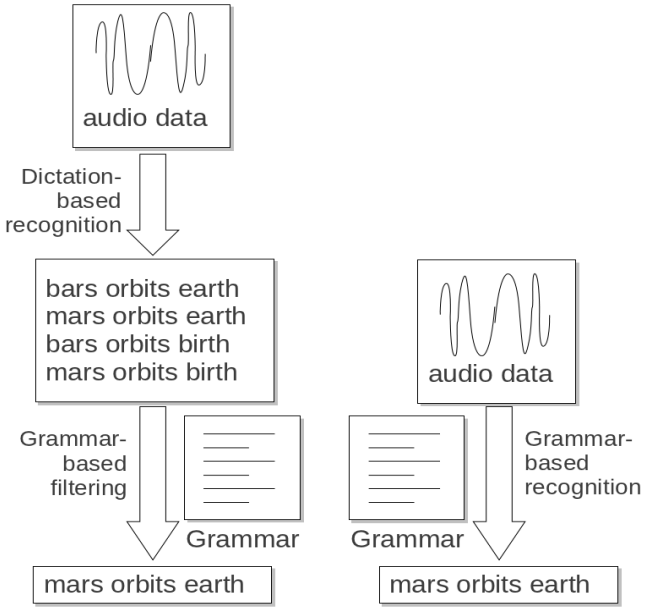


Figure 7: A Comparison of Data Flows in Grammar-based Recognition and Grammar-based Filtering

Dictation-based voice recognizers generally allow the user to utter any sequence of English words, grammatical or not. This flexibility is useful when writing E-Mails or word processing documents, but results in frequent mistakes. When a sequence of words is uttered, it does not limit the words that follow; although it may affect the probability assigned to the results in some context-capable recognizers.

The main difference between grammar-based recognition and applying grammar-based filtering to dictation-based recognition results is the granularity, to which the grammar is applied. In grammar-based recognition, each word is subjected to the constraints of the grammar before it is accepted. In grammar-based filtering, a number of phrases have already been recognized and any phrases with non-grammatical words or sequences of words are eliminated.

D. Benefits to Our Approach

The mobile SpeechWeb application browser gives mobile users the ability to use speech interaction for applications that currently do not support speech input. This browser differs from a typical speech browser, such as those used by the visually impaired, in that it is not intended to simply dictate the text in a traditional web application but provide a browser for a whole new web built upon speech interfaces.

In the cloud computing era, desktop applications and their data are migrating onto the web. Speech applications provide a convenient way to access data. Mobile applications allow people to access this data while on the go. Combining the convenience of a speech interface with the mobile devices is a natural progression. There are examples of speech applications on mobile devices. However, speech applications that provide access to locally-stored data, obtained through traditional approaches (e.g., using IMAP to download E-Mail messages), are not sufficient for most users needs. These users are increasingly using web applications in their daily life. These applications compute and store data on the web. When web applications provide access to data in the cloud, a speech interface to those web applications allows users to access to that data in a non-traditional way. The advantage is that speech interfaces allow mobile device users to access their data in a hands-free manner. For example, users could access cloud data while driving, if those applications had a speech interface.

The browser limits data flow in two significant ways. First, it uses the LRRP architecture that sends only the post-recognition plaintext across the network. Second, the SpeechWeb applications are specially designed with efficient access to narrow information in mind. Rather than download a complete web page with pages of text, and using a cumbersome speech interface to navigate through that text, the speech applications should be designed to give exactly the information the user needs. For example, a encyclopedic SpeechWeb application might answer queries about the data they contain, unlike web-based encyclopedias that return all known data on a specified topic. This strategy is more appropriate for mobile users, since – as their name implies – they are on the move and may not be able to navigate through data as easily as a user at a desktop application or traditional web application.

IV. ANALYSIS

The post-recognition parsing takes a list of possible spoken phrases, in the order of their probability and removes any phrases that do not match the grammar. The order of probability is retained. Assuming the grammar is correct, any phrase not matched by the grammar is not correct. Therefore, we would expect the recognition accuracy to be as good or better than dictation-based recognition.

To demonstrate typical improvement, a user study was conducted. In this study, users are shown 28 valid sentences, containing transitive verbs and noun phrases, from a sample English-based grammar and are asked to speak these phrases aloud. The dictation-based recognition proceeds to generate

 ISBN: 978-1-61208-164-9

up to 25 possible spoken phrases based on the user's speech input. The position of the correct phrase in this list, if any, is recorded. This same position is determined again, after the list of possible spoken phrases is filtered by post-recognition parsing. These two statistics are used, along with the percentage of first-rank results, to compare the accuracy of each approach. The results of this study have been included in Figure 8. In this table, the results for post-recognition parsing are found under the name grammar filtering.

Method	Average Rank	Percentage of First-Rank
Dictation-based	8.16	19.64%
Grammar filtering	4.96	40.54%

Figure 8: A Comparison of the Accuracy of Grammar-based Recognition and Grammar-based Filtering

As shown in Figure 8, the recognition accuracy in the user study was improved by 20.9%. Recognition in this case has very poor accuracy, as is typical for dictation-based recognition when recognizing complete sentences. As expected, the average rank of the correct sentence in the list of possible utterances has improved, since non-grammatical sentences in this list are removed in the process.

V. Related Work

A. Related Projects

The simple XML-based document format, SWML, was created for simplicity; to facilitate non-technical users creating SpeechWeb applications. VoiceXML [3], X+V [4], and Salt [5] are other file formats that contain similar data. However, these formats are complex, which can be a deterrent for non-technical users. Users can begin creating SpeechWeb applications quickly, and for simple question/answer applications the SWML document can be auto-generated. For more interactive SpeechWeb applications, such as those requiring scripted behaviour, the flexibility of VoiceXML would outweigh its complexity. None of these speech application formats have widespread support on mobile platforms.

The discontinued PipeBeach project [6] provided a speech interface to the traditional web for mobile devices. This project builds upon the VoiceXML standard, and as such is dependent upon VoiceXML application support. At the time, mobile devices with speech recognition capability were not yet widely available, and thus no VoiceXML browsers could be executed on the devices. One of the project's goals was to combine the WML and VoiceXML standards. One approach is server-side translation of WML into VoiceXML. The popularity of WML was limited, due to rapid changes in mobile device capabilities.

The w3voice project [7] is a Japanese-language SpeechWeb initiative. The project uses an RRRP architecture, sending raw audio data to the server side for processing by a third-party speech recognizer. RRRP architectures require large data transfers, since the raw audio data must be transferred to the server-tier for recognition. RRRP architectures also require significant resources on the server tier for recognition, since all clients' audio must be recognized on the same site. LRRP SpeechWeb architectures, on the other hand, send only plain text to the server tier, and use a decentralized recognition model.

Our research group has also developed a desktop SpeechWeb browser, based on X+V [9]. Support for X+V is not provided by any known open source applications for mobile devices, and thus a simple port of this browser was insufficient. The mobile SpeechWeb browser uses a custom file format, and post-recognition parsing to improve accuracy.

VI. Conclusions and Future Work

Our group has created a speech browser for mobile devices, that improves recognition where the device uses dictation-based speech recognition. The browser performs voice recognition on the device itself, to reduce data transfer requirements and server processing requirements.

This SpeechWeb application browser has been integrated into the larger SpeechWeb project, the goal of which is to create useful speech-based web applications and encourage others to do the same. We have developed speech applications for query encyclopedic databases, such as information about the solar system, as well as simple applications, such as one that tells jokes. Applications that are used to conduct speech-based surveys are being developed to service people in areas where mobile phones are common, but traditional computing devices are rare.

The SpeechWeb application infrastructure is being used as a test platform for natural language syntax and semantic processing research, allowing additional semantic evaluation constructs to be demonstrated, and providing useful SpeechWeb applications to users.

References

[1] R. A. Frost, “A Call for a Public-Domain SpeechWeb,” *Communications of the ACM*, vol. 48, iss. 11, pp. 45-49, November 2005.

[2] R. A. Frost, A. Karaki, D. Dufour, J. Greig, R. Hafiz, Y.Shi, S. Daichendt, S. Chandon, J. Barolak, and R. Fortier, “MySpeechWeb: Software to Facilitate the Construction and Deployment of Speech Applications on the Web,” *Proceedings of the 10th International ACM SIGACCESS Conference on Computers and Accessibility*, pp. 249-250, October 2008.

[3] http://www.w3.org/TR/voicexml21/ [retrieved: October 17, 2011]

[4] http://www.w3.org/TR/xhtml+voice/ [retrieved: October 17, 2011]

[5] http://msdn.microsoft.com/en-us/library/ms994629.aspx [retrieved: October 17, 2011]

[6] http://www.w3.org/2000/09/Papers/Pipebeach.html [retrieved: October 17, 2011]

[7] http://w3voice.jp [retrieved: September 1, 2011]

[8] http://www.google.com/mobile/voice-search/ [retrieved: October 17, 2011]

[9] Richard A. Frost, Xiaoli Ma, Yue Shi, “A browser for a public-domain SpeechWeb.” *Proceedings of the Sixteenth International World Wide Web Conference*, pp. 1307-1308, May 2007

 ISBN: 978-1-61208-164-9

Motivation for Collective Action in the Smart Living Business Ecosystem

Fatemeh Nikayin, Mark de Reuver
Faculty of Technology, Policy and Management
Technology University of Delft
Delft, The Netherlands
{f.a.nikayin; g.a.dereuver}@tudelft.nl

Abstract - **While smart home services have been on the agenda for over three decades, advances in mobile technologies and concepts like Internet of things are creating a new wave of interest in the market. Traditionally, smart home applications were offered in stovepipe architectures by individual organizations, leading to a plethora of service platforms. Today, smart home service providers are increasingly looking to collaborate in order to jointly develop and share common service platforms. However, collective action literature asserts that such collaboration will only take place if the motives to collaborate outweigh the hurdles. In this paper, we study which motivational drivers lead to collective action for smart living services. We do so in a survey study among 140 home installation companies that are member of the major Dutch branch organization. We find that the tendency to collaborate is mainly driven by motives related to new business opportunities, and that more strategic and solidarity motives do not play a role.**

Keywords - collective action; business ecosystem; motivation; platform; smart living

I. INTRODUCTION

Thanks to mobile technologies and concepts like the Internet of things, the vision of Smart homes is changing from simple home automation systems toward advanced smart entertainment, health support and energy management services. Obviously, what is "smart" depends on time [1]. In the 1980s, the "smartness" of smart home concepts merely involved predefined automation of appliance tasks. Since the year 2000, smartness involves much more flexible task automation adapting to the situation based on past usage data, user preferences and interaction with other devices. In addition, the (mobile) Internet make smart home applications accessible regardless of the device and location of the user [2]. Therefore, the concept of "smart home" no longer fits and we coin the notion of "Smart Living" to represent bundles of innovative ICT-enabled services that aim to add value for home tasks and routines.

Although smart homes have been on the agenda for over three decades, and despite many commercialization attempts in different sectors, smart living services typically do not make it into the mass market [3]. This might be a result of the great fragmentation and complexity in smart living service platforms. In an earlier paper, we find that smart living service platforms are often based on closed architectures and are typically sector specific [4]. Typically, functionalities are replicated in the various industry-specific service platforms, and they are not being shared or reused. Practitioners working in the field of smart home services increasingly point to this fragmentation of service platforms as one of the major hurdles, which is also clear when considering major standardization initiatives like KNX.

As such, there are opportunities for service providers to share such generic functionalities on a common service platform to be used in multiple service offerings [4]. Sharing service platforms and collaborating across industry sectors may not only reduce investment costs, but may also reduce complexity and increase flexibility for consumers. Moreover, open innovation literature stipulates that sharing platforms across company boundaries may increase service innovation [5, 6]. As such, collaboration for smart living services may lead to new business opportunities. Other motivation for open forms of collaboration may come from trends like corporate social responsibility and people-planet-profit paradigms (i.e., sustainability of natural resources), which lead to more altruistic and solidarity types of motives.

To achieve such vision of common service platform for Smart Living services, actors from distinct sectors of industry need to work collectively. However, difficulties in cooperation specifically when actors are from different sectors, may hamper collaboration in this domain. For instance, one of the critical issues is that while actors cooperate for creating a shared value, they compete over having the biggest piece of pie [7]. Accordingly, several problems may arise such as conflicts over division of costs, revenues and investments between parties as well as the division of roles and responsibilities [8]. On the other side, increasing dependency between parties may also influence the governance mechanisms and raise concerns over trust or risk of opportunistic behavior of parties [9, 10].

Such issues of cooperation have often looked from the perspective of game theory or mechanism design [11] and there are relatively less empirical studies have been done in this field. While extensive bodies of literature on collective action have discussed motivations in different context like social and political [12-14], less attention has been paid to motivations in high-tech industry, especially to the smart living domain. Indeed fostering innovation in high-tech

 ISBN: 978-1-61208-164-9

industry like Smart Living is mainly dependent on the collaboration between several actors to integrate their knowledge and resources. Such open innovation happens only when different parties are motivated enough to work together. Therefore, to mobilize such cooperation, decision makers in high-tech industry need to know how those external independent parties may become interested in an innovative cooperation [15].

This paper aims to improve understanding on what drives or blocks collaboration in the field of smart living. More specifically, we explore which types of motivational sources exist in the field of smart living and analyze how those sources of motivation in turn affect the tendency to collaborate for smart living services. To do so, we develop and analyze the results of a survey among 140 installation companies that are active in the field of smart living services. While doing so, we compare different types of smart living services, i.e., energy types of services and entertainment and security services.

This paper is organized as follows: Section II presents theoretical background. Section III provides the method. In Section IV, we present the results, and finally, in Section V, we discuss the results and make recommendation for future study.

II. THEORETICAL BACKGROUND AND RESEARCH MODEL

Technology-wise, a service platform is an evolving system in the form of hardware architecture, an operating system or a software framework. A typical service platform usually contains several components that are required by the services running on that platform, and which those services would otherwise need to include themselves [16]. The network of service providers and platform providers that are working together around a service platform to stimulate innovation around it can be viewed as a 'business ecosystem' [17]. One of the important characteristics of business ecosystems is the interconnectedness between actors which make it necessary for them to cooperate for a shared fate [18]. As such, it is in the interest of most members of a business ecosystem to work collectively to develop and expand an existing market [19]. However, there may be several hurdles that hinder actors to join a business ecosystem and cooperate around an innovation. Examples of such obstacles are handling conflicts, differing motivations and conflicting strategic interests.

The cooperation within members of a business ecosystem can be viewed through the of lens collective action theory. Collective action theory is often applied to explain phenomena in which heterogeneous actors collaborate in order to reach a common goal [20, 21], especially when there are sources of conflicts in achieving 'common goal' through individual action [22]. In collective action literature, motivation is considered as an enabler for cooperation. On a general level, motivation can be viewed as an impetus or inspiration that move an individual towards something [23]. Such inspiration in collective action is typically toward pursuit of a common goal.

The classical dilemma of collective action is that "rational, self-interested individuals will not act to achieve their common or group interests" and they tend to free-ride on contributions of others [24]. Such issue of free-riding may hinder many actors from entering into a cooperation and lead to 'start-up dilemma' [25]. To solve the free-riding problem and to motivate actors for cooperation, Olson [24] argued the essence of 'selective incentives'. 'Selective incentives' can be viewed as those private benefits that are provided for those individuals who have contributed for provision of collective good [26]. Thus, those actors with high interests in 'selective incentives' are more likely to move in a cooperation [27].

The two terms 'motivations' and 'incentives' have been used in the literature interchangeably [14, 27]. However, in this research we distinguish 'motivations' from 'incentives' in a way that motivations are intrinsic or extrinsic impetus toward achieving common, while incentives are those benefits that are provided within a group to stimulate cooperation. In this paper, we mainly focus on the motivations and how they play roles in starting up collaboration.

Based on the previous discussion, we propose the following hypothesis:

H1. Stronger motivations to be involved in smart living projects increase the collective orientation in smart living projects

There are several interpretation of the notion of *Collective orientation* in the literature [28]. Following Driskel and Salas [29], we view collective orientation as an individual's tendency to work collectively rather than alone.

Several streams of literature have studied motivation for collective action in different contexts and proposed different categories of motivations [12, 13, 30]. While some studies suggests that cooperation between individuals may be induced by financial motives [14], others identify other types of motives like *normative, occupational, lobbying, material, social* and *information-motive* [27] that play roles in enabling cooperation. In the psychology literature, motivations are generally categorized into two types of intrinsic and extrinsic motivations. According to [23] intrinsic motivations can be defined as "doing the activity for its inherent satisfaction rather than for some separable consequence." An intrinsic motivated person, perform an activity because of the fun, challenges or the good feeling that the activity entail. In compared to intrinsic motivation, [23] define extrinsic motivation as an action that is induced by instrumental value and is toward achieving 'separable outcome'.

Similar to humans, companies may also encourage in specific activities on the bases of their intrinsic or extrinsic motivations. Put this in the context of Smart Living domain, there might be several types of intrinsic and/or extrinsic motivations for companies to cooperate over a common service platform. For instance, one possible intrinsic motivation could be to make life easier and more convenient for people. However, in competitive business world, extrinsic factors tend to be more critical. As such, in this paper we focus on exploring the extrinsic motives that lay behind cooperation in the smart living domain.

 ISBN: 978-1-61208-164-9

One of the obvious forms of extrinsic motivation is the business value from cost reduction or more income. For instance, platform providers or service providers may invest in a common service platform if they expect that they can reduce their costs and have more return on their investments. Business value can also be achieved by gaining access to specific information (e.g., customers, market), innovative technology, and/or new market opportunities [15]. We refer to these types of motives as 'new business' motives:

H1a. Stronger motivations to be involved in smart living projects for generating new business opportunities increase the tendency to act collectively in smart living projects

Another less tangible types of extrinsic motivation in the smart living domain is networking and building up relationship. Typically, companies may engage in cooperation to enlarge their networks, extend business opportunities and access more partners and projects. These types of motives are mainly cooperation-oriented and their values last longer than business values [15]. We generally refer to them as 'solidarity motives':

H1b. Stronger motivations to be involved in smart living projects for solidarity reasons increase the tendency to act collectively in smart living projects

Beside business and solidarity motives, companies may participate in a cooperative activity to achieve more high-level strategic objectives related to their status and reputation within a market [31]. For instance, being the first one in accessing or using a new technology is important to build up status and strategic position in the market. In this paper, we refer to these motives as 'strategic motives':

H1c. Stronger motivations to be involved in smart living projects for strategic reasons increase the tendency to act collectively in smart living projects

We assume that collaboration between actors is needed to get smart living projects towards the implementation and commercialization stage. Underlying assumptions are that sharing of service platforms that provide generic modular functions, like identifications, authorization and managed data storage, or business functions, like support, management and maintenance, will make it easier to develop new services for smart living service providers [32]. In addition to that, sharing of risks, investment funds and knowledge may benefit the smart living projects. To test these conjectures, we will test the following hypothesis:

H2. The stronger the tendency to act collectively in smart living projects, the more likely that the actor is involved in smart living projects

Obviously, there may be a direct effect between motivations to be involved in smart living projects and the extent to which actors are involved in these projects, without the collective orientation mediating that effect. To test for such direct effect of motivation on smart living involvement, we will also test the following hypothesis.

H3. The stronger the motivations to be involved in smart living projects, the more likely that the actor is involved in smart living projects

Figure 1 visualizes the conceptual model for this paper.

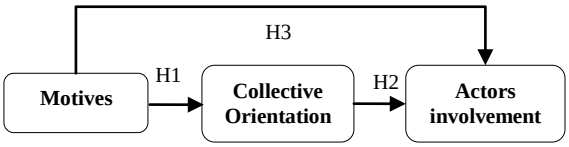


Figure 1. Research Model.

III. METHOD

A. Sampling

We conducted a survey among members of a Dutch branch organization that are providing technical and installation solutions in the area of Smart Living. The survey was conducted through an online questionnaire in January 2011. While the branch organization has in total 5300 members subscribed, 1796 of them were invited to participate that are already involved in domotics or other ICT-enabled solutions. An invitation e-mail was sent through the branch organization to which 144 members responded (response rate 8%). 66% of them participated after receiving the first email and 34% after receiving a reminder. To check non-response bias, we compared the means of the first group of respondents to the latter group of respondents and found no significant differences in overall data [33]. Of the 144 responses, 133 were valid for analysis.

Regarding the background of respondents, 70% of respondents were the owners of the company, 22% were involved in operational management and 8% were project managers. The majority of the participants were involved in strategic and policy management (79%). Similar to the population of installation companies, most participants work at SMEs, as only 14% have more than 100 employees.

B. Measures and Items in the questionnarie

The items to measure motivations were largely adapted from [27, 34]. Respondents were asked to indicate to what extent the statements are important for them, using Likert seven-point scales ranged from *not important* to *very important*. We conduct Exploratory Factor Analysis, using principal axis factors with Oblimin Rotation method and Kaiser Normalization on all the measurement scales, see Table 1. The three extracted factors were consistent with our expected motive types, i.e., *new business motives*, *strategic motives* and *solidarity motive.*

The likelihood of engaging in collective action is conceptualized as *'collective orientation'* and the items were mainly adapted from [35, 36]. Table 2 shows acceptable factor loadings (i.e., > .5) and construct reliability (i.e., > .7).

With regard to the involved actors and their sectors, the respondents were asked to indicate to what extent they are involved in different field of smart living projects. (7-point

Likert scales ranging from *not involved at all* to *is our core business)*. Typically, actors in the smart living domain are involved in offering different types of activities. For instance, many companies are involved in offering energy management solutions as well as security and safety services; several companies are providing domotics and home automation solutions. Recently, healthcare solutions like telecare or telemedicine are also gaining momentum. Accordingly, we include measurements items for four categories of activities in energy, security, health, entertainment and communication fields. These categories for the smart living domain have been confirmed by an expert in the domain.

Table 1. Extracted factors for motivation for cooperation (Oblimin Rotation- KMO =.94, Bartlett's test= 2063.55, p < .001).

Our organization should participate in smart living projects to: (7-point scale: Highly unimportant – Highly important)	**New business Motive α= .95**	**Solidarity Motive α= .92**	**Strategic Motive α= .92**	**Communalities**
To improve our reputation			.89	.89
To improve our position			.77	.74
To emphasize our mission and objectives			.63	.76
To experiment with new technologies	,67			.77
To create new market	.81			.74
To improve our market position	.75			.83
To increase our income	.88			.80
To gain access to more customers	.80			.76
To have market opportunities	.97			.90
To increase cooperation with other companies		.55		.71
To develop joint smart living projects		.84		.82
To raise funds for smart living projects		.86		.77
To strengthen the relationship with other organizations		.49		.80
To help other organizations in smart living projects		.56		.63

Table 2. Extracted factor for positive attitute towards cooperation (Oblimin Rotation- KMO=.74, Bartlett's test= 187.29, p < .001).

To what extent do you find collaboration necessary for smart living services? (7-point scale: Highly disagree – Highly agree)	**Collective orientation α= .79**	**Communalities**
Collaboration leads to problems (R)	.59	.35
The reasons to collaborate are scarce in smart living services (R)	.90	.81
Collaboration leads to better services	.64	.41
It's belter to deliver smart living services alone rather than with others (R)	.71	.50

(R) Reversely coded

For actors' involvement, despite our expectation of the four categories of smart living services, only two factors were extracted based on a cut-off eigenvalue of 1.00, see Table 3. The first factor includes activities in the field of energy and the second factor covers other fields like, security, health, entertainment and communication. We refer to the first factor as *'energy field'* and the second factor as *'other fields'*.

Table 3. Extracted factors for fields of involvement (Oblimin Rotation- KMO=.83, Bartlett's test= 658.43, p < .001).

To what extent is your company currently involved in smart living projects in the field of: (7-point scale: We are totally not involved – Is our core business)	**Energy fields α= .83**	**Other fields α= .88**	**Communalities**
Energy supply smart grids and smart meters	.57		.53
Systems for heating and ventilations management	.81		.63
Applications in the health		.78	.63
Applications that improve elderly independent living		.83	.66
Integrated entertainment and Info. communication services		.84	.65
Smart security services		.75	.59
Smart anywhere any time working services		.62	.53
Smart climate systems	.85		.68
Intelligent water management system	.68		.48

IV. RESULTS

A. Correlations

After extracting factors, in order to test our research model, we correlate the extracted factors. (See Table. 4 for correlation between the factors)

Table 4. Correlation between extracted factros (* p < .05, ** p < .01).

	New business Motive	**Solidarity Motive**	**Strategic Motive**	**Collective Orientation**	**Other fields**
New business motive	1				
Solidarity Motive	.70**	1			
Strategic Motive	.80**	.72**	1		
Collective orientation	.19*	.12	.11	1	
Other fields	.24**	.30**	.34**	-.033	1
Energy field	.11	.22*	.22**	-.058	.57**

With regard to our first hypothesis, not surprisingly, there is a positive correlation between *new business motives* and *collective orientation*. However, there is not a significant relation between the two other types of motives (i.e., *solidarity* and *strategic motives*) and collective orientation.

This implies that material values play more important role in motivating actors to work together in this domain. Still, considering the strong correlations between the different dimensions of motives, there may indeed be a direct or second-order effect of the other motivation types and the collective orientation.

H1 – Partly supported (H1a: Supported; H2b: Rejected; H1c : Rejected)

In our second hypothesis, we assume that collective orientation leads to actor's involvement in smart living project. However, it appeared that actors' involvement is not related to their collective orientations. Put simply, even if actors are involved in the smart living projects, it does not mean that they have positive attitude toward cooperation.

H2 – Rejected

We also correlate the factors of motives to the factors of actors' involvement to control for the direct effects between them. Apparently, the actors in all fields are generally motivated for cooperation. However, their motivation is mainly self-centered and toward strategic positioning in the market, i.e., towards improving the reputations, status or emphasizing their own objectives.

H3 – Supported

V. DISCUSSION AND RECOMMENDATIONS FOR FUTURE STUDIES

The results indicate that the primary motivation for cooperation in the smart living domain is to create *new business* opportunities. This result is opposed to our initial lessons in our ongoing, qualitative case studies where many practitioners refer to solidarity and cooperative-oriented motives to be important in this domain. Possibly, there is little cooperation going on in this field at this moment and mostly actors are still perusing their goal in isolation.

Despite our expectations, there is not any relation between *'collective orientation'* and the actors' current involvement in this domain. In other words, collaboration between actors is not related to the extent to which they are involved in smart living projects. Apparently, collaboration is not a prerequisite at this moment for being involved in smart living projects. Alternatively, this may imply that there might be several issues in collaboration that hinder even interested actors to move in this domain.

With regard to the motivations, there is a stronger relation between motivation and involvement in *other fields* rather than the *energy field.* This is quite in line with our observations in the domain where many service providers in energy sectors are offering isolated smart metering services [4]. However, in both groups, strategic motives are stronger than solidarity and business motives. This might be explained by the fact that still there is not a dominant actor in the smart living domain and actors from distinct sectors seek to effectively position themselves in this growing market. As such, companies are tapping into each other business and trying to prove themselves in this industry. For instance, telecom companies are considering to provide energy services to households through their fiber infrastructure [37]. This indicates that having a strategic position, status and reputation in the smart living domain is in the interest of all the involved actors.

Despite the competition for dominance in this domain, literature discusses the importance of inter-organization cooperation for the growth of smart living industry [38]. The current trends of proprietary service platforms with differentiated standards leave no space for cooperation, while the promise of a shared service platform highlights the growing importance of cooperation between parties around the platform, i.e., through open API or open standards, to stimulate innovation in the smart living business ecosystem. However, the challenge is how to set up such collaboration, considering several underlying issues that may hamper actors to move in cooperation.

In this paper, we aimed to answer the questions about motivations that lay behind the actors' cooperation in the smart living domain, though, the questions about 'selective incentives' and their importance in persuading actors for cooperation remained unanswered. Furthermore, demotivation issues, like conflicting strategic interests, lack of trust and disagreement over division of costs and benefits, that may hinder cooperation are missing in this study. We also didn't study the effects of actors' performance or the level of interdependency among them. We suggest that further research include the effects of those issues in their studies.

As in any cross-sectional survey study, a limitation is that we measured the independent, mediating and dependent construct at the same moment in time. As such, we cannot test one of the conditions of causality, that is, time difference.

Regarding the population, this study just includes the installation companies and no service providers, network providers, or IT vendors. This makes our results stronger in terms of internal validity, though, less strong regarding external validity. As such, one subject to be explored in the future studies is whether the motivations differ when other actors are included in the population.

In this study, we mainly use exploratory data analysis techniques to explore the measurement scales and the causal model. In subsequent analysis, we will use more confirmatory and stringent techniques to test the results. We will use confirmatory factor analysis to test the measurement model. Moreover, we will use structural regression analysis in SEM to more stringently test the mediation effect that the collective orientation may have on the relation between motivations and involvement in smart living projects.

Possibly, given the strong correlations between the different types of motivations, there may be a multilevel structure inherent in the measurement model. A second-order construct *Motivations* that influences the three underlying

 ISBN: 978-1-61208-164-9

dimensions explored in this paper may better explain the other theoretical constructs. Similarly, the strong correlation between the two types of smart living projects may be explained by a higher-order factor. We will test for such higher-order factors in our subsequent research steps using structural equation modeling.

Reference

1. Weiser, M., *Open house*. Review, the web magazine of the Interactive Telecommunications Program of New York University, ITP Review, 1996. **2**.
2. Rohracher, H., *Smart Homes and Energy Efficiency Constructive Technology Assessment of ICT Use in Sustainable Buildings*, in *ACEEE*. 2002. pp. 241-252.
3. Peine, A., *Understanding the dynamics of technological configurations: A conceptual framework and the case of Smart Homes.* Technological Forecasting and Social Change, 2009. **76**(3): pp. 396-409.
4. Nikayin, F., D. Skournetou, and M. De Reuver, *Establishing a Common Service Platform for Smart Living: Challenges and a Research Agenda.* Toward Useful Services for Elderly and People with Disabilities, 2011: pp. 251-255.
5. Eisenmann, T.R., et al., *Opening platforms: How, when and why?* 2008: Harvard Business School.
6. Chesbrough, H.W., *Open innovation: The new imperative for creating and profiting from technology*. 2003: Harvard Business Press.
7. Brandenburger, A.M. and B.J. Nalebuff, *Co-opetition: A revolutionary mindset that combines competition and cooperation: The game theory strategy that's changing the game of business*. 1997: HarperCollinsBusiness.
8. de Reuver, M., H. Bouwman, and T. Haaker, *Mobile business models: organizational and financial design issues that matter*. Electronic Markets, 2009. **19**(1): pp. 3-13.
9. De Reuver, M., *Governing mobile service innovation in co-evolving value networks*, in *Faculty of Technology, management and policy*. 2009, Technology University of Delft: Delft. pp. 159.
10. Williamson, O.E., *Corporate finance and corporate governance.* The journal of finance, 1988. **43**(3): pp. 567-591.
11. Locher, T., et al., *Free riding in BitTorrent is cheap.* IRVINE IS BURNING, 2006. **300**: pp. 85.
12. Elster, J., *The cement of society: A study of social order*. 1989: Cambridge Univ Pr.
13. Schlozman, K.L., S. Verba, and H.E. Brady, *Participation's Not a Paradox: The View from American Activists.* British Journal of Political Science, 1995. **25**(01): pp. 1-36.
14. Gillinson, S., *Why cooperate? A multi-disciplinary study of collective action.* Overseas Development Institute, 2004.
15. Boudreau, K.J. and K.R. Lakhani, *How to manage outside innovation.* mIt Sloan management review, 2009. **50**(4): pp. 69-76.
16. Evans, D., A. Hagiu, and R. Schmalensee, *Invisible engines: how software platforms drive innovation and transform industries*. 2006: The MIT Press.
17. Moore, J., *Predators and prey: a new ecology of competition.* Harvard Business Review, 1993. **71**: pp. 75-75.
18. Iansiti, M. and R. Levien, *Strategy as ecology.* Harvard Business Review, 2004. **82**(3): pp. 68-81.
19. Moore, J.F., *Business ecosystems and the view from the firm.* Antitrust Bull., 2006. **51**: pp. 31.
20. Oliver, P., G. Marwell, and R. Teixeira, *A theory of the critical mass. I. Interdependence, group heterogeneity, and the production of collective action.* American Journal of Sociology, 1985. **91**(3): pp. 522-556.
21. Poteete, A. and E. Ostrom, *Heterogeneity, group size and collective action: The role of institutions in forest management.* Development and change, 2004. **35**(3): pp. 435-461.
22. Keohane, R.O., *Cooperation and International Regimes*, in *After Hegemony*. 1984, Princeton University Press: New Jersey.
23. Ryan, R.M. and E.L. Deci, *Intrinsic and Extrinsic Motivations: Classic Definitions and New Directions* 1.* Contemporary educational psychology, 2000. **25**(1): pp. 54-67.
24. Olson, M., *The Logic of Collective Action: Public Goods and the Theory of Groups*. Revised edition ed. 1971, Massachusett: Harward University Press. 186.
25. Bina, M. and G.M. Giaglis, *Unwired collective action: Motivations of wireless community participants.* 2006.
26. Oliver, P., *Rewards and punishments as selective incentives for collective action: theoretical investigations.* American Journal of Sociology, 1980. **85**(6): pp. 1356-1375.
27. Knoke, D., *Incentives in collective action organizations.* American Sociological Review, 1988. **53**(3): pp. 311-329.
28. Alavi, S.B. and J. McCormick, *A new approach to studying collective orientation in team contexts.*
29. Driskell, J.E. and E. Salas, *Collective behavior and team performance.* Human Factors: The Journal of the Human Factors and Ergonomics Society, 1992. **34**(3): pp. 277-288.
30. King, D.C. and J.L. Walker, *The provision of benefits by interest groups in the United States.* The Journal of Politics, 1992. **54**(02): pp. 394-426.
31. Lopes, H., A.C. Santos, and N. Teles, *The motives for cooperation in work organizations.* Journal of Institutional Economics, 2009. **5**(03): pp. 315-338.
32. Feng, W., *Remote service provision for connected homes* 2010, De Montfort University. pp. 163.
33. Armstrong, J.S. and T.S. Overton, *Estimating nonresponse bias in mail surveys.* Journal of marketing research, 1977. **14**(3): pp. 396-402.
34. Sako, M., *Suppliers' associations in the Japanese automobile industry: collective action for technololgy diffusion.* Cambridge Journal of Economics, 1996. **20**(6): pp. 651.
35. Kelly, C. and S. Breinlinger, *Identity and injustice: Exploring women's participation in collective action.* Journal of Community & Applied Social Psychology, 1995. **5**(1): pp. 41-57.
36. Frege, C., *Union membership in post-socialist East Germany: who participates in collective action?* British Journal of Industrial Relations, 1996. **34**(3): pp. 387-413.
37. Jeff St., J., *The Telco Home Energy Invasion*. 2009, GreenTechMedia.
38. Novais, P., et al., *Inter-organization cooperation for ambient assisted living*. Journal of Ambient Intelligence and Smart Environments, 2010. **2**(2): pp. 179-195.

 ISBN: 978-1-61208-164-9

Extending Friend-to-Friend Computing to Mobile Environments

Sven Kirsimäe, Ulrich Norbisrath, Georg Singer, Satish Narayana Srirama, Artjom Lind
Institute of Computer Science, University of Tartu
J. Liivi 2, Tartu, Estonia
Sven.Kirsimae@ut.ee, Ulrich.Norbisrath@ut.ee, Georg.Singer@ut.ee, Satish.Srirama@ut.ee, Artjom.Lind@ut.ee

***Abstract*—Friend-to-Friend (F2F) computing is a popular peer to peer computing framework, bootstrapped by instant messaging. Friend-to-Friend (F2F) Computing is a simple distributed computing concept where participants are each others friends, allowing computational tasks to be shared with each other as easily as friendship. The widespread availability of applications and services on mobile phones is one of the major recent developments of the current software industry. Due to the also emerging market for cloud computing services we mainly find centralized structures. Friend-to-Friend computing and other Peer-to-Peer (P2P) computing solutions provide a decentralized alternative. However these are not very common in mobile environments. This paper investigates how to extend the Friend-to-Friend computing framework to mobile environments. We describe our process and point out possible pitfalls and achievements. For this investigation, we ported our private cloud environment Friend-to-Friend Computing to Android and Symbian. We demonstrate a mobile gaming application using Friend-to-Friend Mobile.**

***Keywords*—Distributed Computing; Peer-to-Peer; Social Networks; Android; Symbian; Mobile Software; Mobile Networking and Management**

I. Introduction

In this paper, we will present a case study on how we ported our F2F Computing framework to Android and the Symbian S60 mobile platform. Based on our observations we will outline the process of making P2P computing frameworks mobile aware.

Along with the increased availability of software as a service (SaaS) on PCs, services are also moving into the mobile market. Moreover, a smart phone nowadays has become a commodity device with millions subscriptions worldwide. It is not just a mere voice only device anymore, but offers many alternative communication technologies. The importance of running applications and accessing services becomes the key demand from the users. Also having access to these in a convenient manner plays an important role. Users do not care about installation process, installation locations–on the phone or in the cloud–, only about the provided functionality. They appreciate the possibility of having a set of self selected applications on their devices. The importance of this application and service support becomes evident with the success of the iPhone [1] and Android platforms as well as with the struggling of Nokia [2] to provide a competing architecture. In addition, Nokia's investment in Maemo and Meego, opensourcing and withdrawing Symbian and now the switch to Windows 7 at the same time as well as Google's success with Android prove the same point. Their strategy is to offer other ways to distribute applications and services as an alternative to Apple's very popular and successful but proprietary platform.

The approach of providing computational resources from smart phones for various collaborative tasks is conceptually similar to providing services on them. This was studied at the mobile web service provisioning project [3], where Mobile Hosts were developed, that provide basic services from smart phones. Mobile Hosts enable seamless integration of user-specific services to the enterprise by following web service standards, also on the radio link and via resource constrained smart phones [4].

Friend-to-Friend (F2F) Computing is a simple distributed computing concept where participants are friends or acquaintances of each other, allowing computational tasks to be shared with each other as easily as friendship. It will be explained in more details in Section II. A network of friends is the base for F2F Computing. A similar concept can be seen among mobile device users. People using these devices are often socially connected. There are multiple services for mobiles available nowadays, which support this concept. Examples for such services are Facebook, Twitter, Flickr, blogging, youtubing, or multiplayer games.

In this paper we describe the simply installable extension of F2F Computing called F2F Mobile running on the Android and Symbian mobile operating systems. We will show the achieved transparency between mobile and static systems from a developer's point of view. We will show some of the difficulties in extending P2P computing to the mobile environment, which have to be addressed by other mobile platform developers. We will also present an application demonstrated on top of our F2F Mobile.

The rest of the paper is organized as follows. Section 2 will give a brief outline of Friend-to-Friend (F2F) Computing. Section 3 introduces the F2F Mobile concept, its implementation details, the demonstrated application, and a small reference of criteria for mobile platform selection and implementation issues. Section 4 shows related work for F2F and mobile environments. Section 5 concludes the paper with future research directions.

II. Friend-to-Friend Computing

Friend-to-Friend (F2F) Computing was initially motivated by the complexity of setup, usage, and administration of Grid

 ISBN: 978-1-61208-164-9

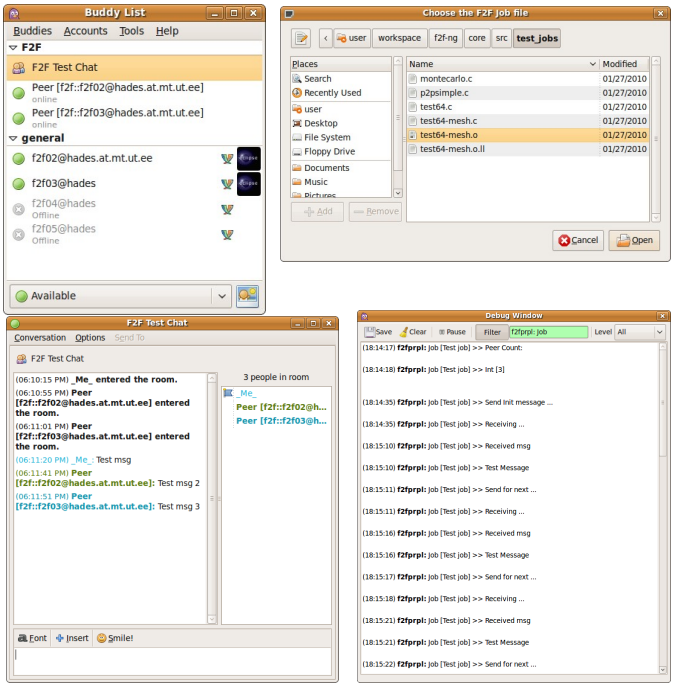


Figure 1. F2F Pidgin plugin screenshots: F2F Group friends, F2F Chat, application selection and debug window

networks. Encouraged by the fact that Skype made voiceoverIP (VoIP) usable for everyone and these deficiencies of current Grids, we combined the best parts of both ideas into a system today known as Friend-to-Friend Computing. F2F Computing is an open source project for spontaneously running distributed applications using the resources provided by computers of friends. The initial setup of the virtual parallel or distributed computer environment makes use of Instant Messaging (IM) software as the triggering means. As a result, people can share computation tasks and other resources in an easy and intuitive manner through their social connections.

Creating an F2F Computing network consists of specifying a set of IM contacts and starting distributed applications or services. An important aspect of F2F Computing is its ability to be plugged into various different instant messengers and its independence of and interoperability between different instant messaging protocols. F2F Computing tries to choose from a set of network protocols the fastest one to connect each peer with each others. It also uses Network Address Translator (NAT) traversal techniques to accomplish efficient connections in case of address translating network devices. Figure 1 shows a collection of screenshots of dialogs of the F2F Pidgin plugin. It shows the list of friends added to a F2F Group, an F2F Chat, the application selection, and a debug window of a running F2F application.

The new architecture of F2F Computing (see Figure 2) allows different clients for setting up F2F networks and running applications or services on them (F2F Adapters). Most commonly used are plugins for IM (we have plugins for Pidgin and SIP Communicator – now Jitsi [5]). To run an F2F network the plugins need to be installed on the devices of all participants. One of the group members has to initiate the F2F Group by adding initial participants. After authorization by these, applications or services can be started as jobs. There are also command-line clients for starting initiators without user interaction.

Once the F2F network is established, the clients can exchange files between their friends in their contact lists, use collaboration tools like a whiteboard, play games, or accelerate computational intensive tasks like rendering or research simulations. Computational applications that have been carried out on F2F Computing include Monte Carlo computations, distributed matrix multiplication making it possible to solve large distributed systems of linear equations, and rendering tasks on Blender [6].

The first version of F2F Computing [7] was written in Java and was realized as a plugin of the multi protocol instant messenger SIP Communicator (Jitsi). F2F Computing was later rewritten to have a lower footprint. The core is now implemented in C allowing it to be ported even to restricted platforms like mobile devices. On a standard x86 system, the core itself has now only a size of 34k. We also re-implemented the application layer. This means, we implemented the instant messaging adapter as a plugin for Pidgin and as a Python command line client. For the execution adapter, which executes the tasks, we initially used Python. Python is here only used as a prototype for the access to the actual core. Currently, we are adapting Low Level Virtual Machine (LLVM) [8] and Java Virtual Machine (JVM) as execution adapter to allow other languages (like C, C++, and FORTRAN) for the executed applications and services. Python is only used here to access the F2F API.

The F2F Computing framework needs fast communication between participating peers (friends). The research was triggered by the need for fastest available communication for distributed desktop computing (cycle scavenging) applications. However, the fact that also other IM applications like VoIP, Video over IP and file transfer, are in real need for direct and fast communication availability, F2F Computing has grown into other application domains as well. We have successfully used F2F Computing for the aforementioned computational applications, for teaching, a cross-IM-brand whiteboard application, two computer games and a file transfer application.

Figure 2 shows the F2F Computing architecture. The framework consists of four layers: F2F enabled applications, application, adapter, core, and communication. In the adapter layer we provide the abstraction from the different communication providers of the communication layer in one uniform interface for F2F Computing applications, which can be run via the Computing adapter as byte-code compiled from various languages. This abstraction provides access to the concepts service, peer, and group.

One of the important requirements of F2F is reliable connectivity between peers and speed of communication. Therefore the framework has various connectivity possibilities like the TCP communication provider, reliable UDP communication provider, or the IM communication provider. In the future, the communication layer can be easily extended by implementing the specific communication providers (for example, *Bluetooth*, *Infiniband*). The framework chooses always the fastest way. If

 ISBN: 978-1-61208-164-9

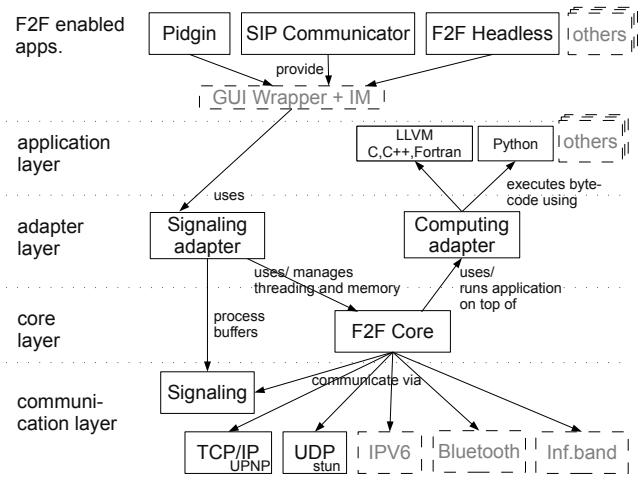


Figure 2. The F2F Computing architecture

peers are in the same local network direct TCP should be possible. If it is not, the framework attempts to establish UDP connection using NAT traversal techniques implemented in the UDP communication provider. In the worst case, if peer-to-peer connectivity is impossible, the messages are sent through instant messages of the respective instant messenger.

III. F2F Mobile

F2F Computing is based on a simple and transparent social concept: being friends or acquainted with others. It also offers an easy way to share resources on demand between a network of such friends. The mobile phone is one of the most important appliances for supporting social interaction. The demand for services and applications ubiquitously supporting this social interaction is growing [9]. However, due to the strongly fragmented nature of mobile technology, current service and application development is complex and focused on single services or products. Our hypothesis is that F2F Mobile will introduce a transparent environment for social applications and services. F2F Computing allows a spontaneous creation of social networks and the deployment of applications and services within this network. Including mobile devices in such a network is an obvious step.

Consider the use case where a group of people want to share some content in private and independent of third party services. Having F2F Computing embedded into mobile devices allows them to spontaneously create a private network of friends in order to share content or run an application. F2F Computing relies on a third party service only while bootstrapping the P2P network. After direct connections were established the communication is happening only between the peers in the private group.

Before designing F2F Mobile, we considered two major factors – the market penetration of mobile platforms on the one hand and the simplicity of third party development of mobile platforms on the other hand. At the time of carrying out this research, Nokia's Symbian, Google's Android, and Apple's iPhone OS had approximately the same share of applications on smart phones. However in short time Android and iPhone OS have significantly gained importance and currently Nokia is far below 50 percents market share. Today we would suggest to target the Android platform first as it comes with less restrictions especially in a legal sense than Apple's iPhone OS. It is not allowed to run a virtual machine, like for example Python, without violating the EULA. As Android was not that popular and not that feature-rich (there was no possibility to run native code) at the time we started our research, we selected Nokia's Symbian S60 for our F2F Mobile prototype. It provided a proved C development environment, Python support, and accessibility to the devices. However, the next F2F Mobile port will be Android based, the development process will be addressed later in this section.

F2F Computing is currently mainly based on C and several language interfaces. Symbian S60 supports by default a development in C and C++. A Software Development Kit (SDK) [10] and a development environment based on eclipse [11] are provided for free. A well maintained Python port [12] exists. However, the build of Symbian binaries is supported well only on Windows platforms, and this was a biggest disadvantage, as our research lab is Linux-based. In comparison Android SDK is supported on all three platforms (Windows, Linux and Mac OS).

Mobile devices are usually very restricted concerning their hard- and software. Applications are event-driven rather than multithreaded. In Symbian multithreading is possible and is used inside the Operating System, but it is generally avoided in applications, because it potentially creates several kilobytes of overhead per thread. Therefore, we avoided threads in the Symbian port. Because of memory limitations, applications are restricted in comparison to standard PC systems in their memory allocation strategy. The heap might be only several Megabytes and the stack be even smaller. Allocating memory dynamically can easily lead to termination of the application, system crashes, or kernel panics. Therefore, as a first step, the core of F2F Computing was rewritten completely in ANSI C without threading and with static memory management. Symbian does not provide the standard C libraries by default. We used Open C/C++ [11] to provide the missing C libraries. For the Python adapter and Python execution environment, we had to add more missing C libraries and the corresponding Python interfaces. To allow simple installation of the F2F Mobile the imported libraries were packaged additionally into the F2F Mobile package, also including the F2F Core, and the Python F2F Adapter as application layer (Figure 2).

Another limitation is that the S60 SDK emulator only emulates but not replicates the mobile device. In some cases it even represents fewer constraints (memory, access rights) compared with the real mobile device. Developing, testing and debugging with the SDK means that the application needs to be re-tested on the real devices in the real environment to make sure it is behaving as expected.

For the installation, there have to be four packages installed (in this order): Python for S60 3rd Edition 1.4.5, corresponding Python shell, Open C/C++, and the F2F Mobile package. The actual receiving client Python script comes as a separate file. It will be started from the shell. After logging in to

 ISBN: 978-1-61208-164-9

Figure 3. Screenshots of demonstration application on F2F Mobile.

F2F Mobile core, it will wait for the application–the F2F task– to arrive. The application itself can be sent either from another mobile device (with a server script) or from an instant messenger with the F2F plugin installed on a full PC. The application arrives to the mobile client and executes on all the arriving peers and automatically makes the spontaneous F2F network infrastructure available for all participants. As one sample application we implemented a small hangman style game taking approximately 200 lines of code. The application runs both on mobile and PC platform. At the moment the GUI is still coded separately for PC and mobile platform. In this game one player thinks of a word and all other players have to guess it in turns, either guessing a new letter or the whole word. If somebody guesses the word, this player can think of the next word. Figure 3 shows some screenshots of the start up and running the game on two different mobile devices. The first two screenshots were taken on an Nokia E70, the second two in the S60 Emulator. The setup we used here were a PC with the emulator, an E70, an E61, and a PC with Pidgin and the F2F Plugin. We submitted the application from the PC. The first two screenshots show the configuration and application deployment phase, the last two screenshots are taken while the game is running.

Table I and Table II summarize our experiences, when porting F2F Computing to Symbian S60. Table I summarizes possible problems porting a platform to a mobile environment. Table II shall be a reference for the technical realization of such a port. It depicts our problems and selections we had to make to achieve the port.

We did some experiments with Maemo and could prove

criterion	comment
platforms to consider	Symbian, iPhone, Android, Blackberry RIM, Windows Mobile, Maemo [13], Meego
market penetration, for personal target market	• Take into account current penetration and make platform decision accordingly • Development/change of the penetration over time (growing, shrinking?)
maturity of platform	How established is the platform? Has there been done already a lot of development on it? Is there an active community of developers? iPhone OS and Android are for example still young players, but have a very big and fast growing community.
distribution channels	How can applications be uploaded to a phone, is it only possible via an application store, do other possibilities exist? Is the kind of material, which can be distributed, restricted?
non functional requirements	• Are used programming languages supported? • Does an SDK exist? Is it free or not? • Is a development environment provided? Is it free or not?
legal issues	• License fees, are all interfaces available, which are needed? • Consider legal technical restrictions (for example no virtual machine allowed on iPhone) • Application signing issues, can only signed applications be run?, how difficult/expensive is signing? • Do you use open source libraries (GPLv3), that you will not be allowed to use in a potential restricted environment? • Are there usable open source (OS) components? These can avoid a lot of problems and simplify development very much.

Table I
SELECTION CRITERIA FOR A TARGET MOBILE PLATFORM.

that we can compile the F2F Computing with Python adapter directly. This is not surprising as Maemo is basically a far less stripped down Linux than Android.

Addressing the criteria, which are summarized in the tables, we present our initial experience with the F2F Computing Android port. Portability and performance are only two of the many advantages of Friend-to-Friend Computing. Its modular architecture allows to build different sets of installations in order to fit specific platforms. In the Android port we reuse the F2F Core component with only minor changes to the code. These changes are mainly in regard to providing the Java interfaces. The graphical user interface (front-end) can then be implemented in Java using Android specific widgets. Having a platform specific front-end helps utilizing the corresponding platform related features like touch screen, drag and drop, scaling, or 3D rendering. Utilizing these features is essential to provide user-friendly GUI.

The performance is granted by the core component written in C. During implementation of the Symbian S60 port, we faced multiple issues based on limits of S60 platform (Open

 ISBN: 978-1-61208-164-9

criterion	comment
threading	Try to avoid threading, some platforms might have problems.
multi-tasking	There might be different levels of multi-tasking available. iPhone: no multi tasking for applications, Maemo full pre-emptive.
memory management	Low footprint, avoid dynamic memory allocation.
test on real device	The Symbian emulator has different restrictions.
languages	Choose a good balance between languages you use, try to have an abstracting core or library, try to use a wrapper.
library dependencies	Check the dependencies and try if referenced libraries are either available or compilable. We had problems with hashlib and Pyexpat (XML).
datatype sizes	Make sure to wrap datatypes, so they can be used in a mobile environment. Usually the size has to be fixed in number of used bits.
ensure simplicity of installation	See, how the packaging mechanism for you target platform works. Is it possible to pack all in one package?

Table II
CRITERIA TO ACCOUNT FOR IN THE IMPLEMENTATION PHASE.

C/C++ library, small memory and avoiding the threads usage). According to Android documentation there is a Bionic library [14] (custom libc) and Native Development Kit (NDK) [15] starting from Android 1.5. While taking a closer look at the Bionic library we found incomplete support of the POSIX threads and C++ exceptions. The pthread_cancel() is not supported and pthread_once() is limited as we use not these functions neither C++ in F2F Core component implementation it will be well portable to Android platform featured with Bionic library. We add the Android tool-chain to our build scripts. This is not an issue due the Python based SConstructor (SCons) [16] builder that we use. There is an exhausting manual how to provide a new tool-chain for SCons [17], there is also project [18] using combination of SCons and Android NDK. Considering legal issues, there is no special licenses needed for Android development and the development tools are free to download for Linux, Windows and Mac.

Android uses Java classes (Widgets) to provide an API for writing user interfaces. In order to wire the Java based user interface and the native C-code of F2F Core and the corresponding C routines are exposed to Java using Java Native Interface (JNI). The entire process can be completely automated with Simplified Interface Wrapper Generator (SWIG) [19]. The same solution was applied to expose F2F Core routines to the Python back- and front-end in the F2F mobile prototype for Symbian S60.

The computing engines are essential to remote-execute the code in F2F. For Symbian there was Python support, for Android there are two ways: either we build standard Python from sources [20] using NDK or we use the scripting-layer (SL4A) [21]. SL4A allows to access Android Native API using various scripting languages including Python. In addition it is possible to run scripts from native code using Intent Builders of the Android SDK. This means that we are able to run the remote Python Scripts on Android platform. Having the Python support is essential to have compatibility with existing Symbian S60 port and the Desktop builds of F2F Computing. As Android is Java based there is not much effort needed to provide Java computing engine for F2F Android port. The F2F Core component is loaded into JVM (Dalvik) at runtime, corresponding Java routines are accessible using standard JNI methods (FindClass, GetMethodID, CallVoidMethod, CallIntMethod).

Another project we carried out on Android was porting LLVM with the purpose to execute the native code remotely on Android platform. The LLVM port to F2F was successful on the Desktop platforms, therefore the next step to keep compatibility is to introduce it on the mobile platforms. As LLVM supports multiple languages (C/C++/Fortran/Java/Python) porting it once to Android allows us not care about running remote Python or Java code.

Android standard and native development kits are supported by major platforms (Windows, Linux and Mac OS X). In addition there is an Android development plugin for Eclipse.

These experiments and observations with Android show that the modular architecture of F2F Computing is well suited to be ported to Android.

IV. RELATED WORK

There exist several groups trying to realize concepts similar to F2F. We took a look at the akogrimo [22], [23] project. It claims to achieve the step “from Cluster Grids toward Mobile Collaborative Business Grids”. It outlines a similar vision to our F2F Mobile on several whitepapers. Case studies, market analysis, and real deployed networks are missing. Therefore, the actual differences in software and implications from the mobile fragmentation are not addressed.

F2F Computing focuses mainly on the mobile client side and on the option to spontaneously setup private cloud environments. Conceptually, it does not distinguish the type of device that is participating in the F2F Computing network. Therefore, F2F Mobile works similar to the mobile clients for public clouds, yet helps in sharing resources and CPU cycles of individual mobiles. Thus F2F Mobile looks similar to P2P Computing, but we still distinguish the general perception of P2P and F2F Computing. For example, Boinc [24] is regarded as a P2P Computing application. It allows harvesting the CPU cycles via a central mediator. Even if it is called a P2P Computing solution, it only facilitates P2P in a collaborative sense but still uses a star topology and therefore becomes an example for a client server architecture. In that sense F2F is more related to P2P than these traditional applications that are thought to be P2P.

Regarding moving P2P based systems to the mobiles we have studied other projects, ex: LightPeers [25]. LightPeers provides a good proposal of the lightweight P2P platform with a well defined architecture. They also propose P2P protocols that are minimalistic and easy to implement. They studied other existing P2P protocol libraries (JXTA, JXME, Proem) and identified the problems with adapting these technologies for the mobile environment, thus were proposing their own

 ISBN: 978-1-61208-164-9

set of architecture and protocols. However, there is not much information provided about the implementation as well as the way to distribute and parallelize applications that run on top of the LightPeers platform.

V. Conclusion and Future Work

With this paper, we showed how the F2F Computing platform can be extended to mobile devices. This drift supports establishing mobile private clouds with significantly less effort. The paper listed several tricks and difficulties and choices taken to avoid them in building F2F Mobile. The paper also showed some applications that were implemented, proving the technical feasibility of the concept.

Not only proving the concept of F2F Mobile, the study also provides several guidelines in building such systems. Especially with the tables provided in Section 3, we highlighted the points in selecting the destination mobile platforms and architectural choices to be taken care of in building application and services. This study will generally be useful for any community that is working and building private clouds and distributed computing platforms that have in foresight shifting their platforms to mobiles.

While the first prototype of F2F Mobile is ready, it provides a lot of scope for further research. Our efforts are especially directed to the possibility of writing code in any language and be able to run it on mobiles. For this, we are developing support of other execution adapters like LLVM [8] in addition to Python. This will allow us to support multiple languages for development and in terms of hardware architecture heterogeneous execution environments. A further important goal is to create a release candidate for Android. If Apple changes some of its software restrictions an iPhone port would also be possible. Furthermore, we are also interested in building more applications and services for F2F Computing, especially in collaborative, mobile gaming, and m-learning domains. Another huge issue is a standardization of some GUI elements offered transparently in the F2F environment. Due to the severely different screens and use patterns on the client devices such an abstraction will be a challenging research area.

Because of the strong fragmentation in terms of hardware, software, and operators, there still remain many problems with transparent on demand software deployment across multiple static and mobile participants in today's networks.

VI. Acknowledgement

This paper was supported by the European Social Fund through the Estonian Doctoral School in Information and Communication Technology.

References

[1] 148Apps.biz | apple iTunes app store metrics, statistics and numbers for iPhone apps. Available from: http://148apps.biz/app-store-metrics/ [cited July 19, 2011].

[2] Nokia's application store faces apple dominance. *Time*, May 2009. Available from: http://www.time.com/time/business/article/0,8599,1900901,00.html [cited July 19, 2011].

[3] S. N Srirama, M. Jarke, and W. Prinz. Mobile web service provisioning. In *Proceedings of the Advanced Int'l Conference on Telecommunications and Int'l Conference on Internet and Web Applications and Services*, page 120, 2006.

[4] S. N Srirama and M. Jarke. Mobile hosts in enterprise service integration. *International Journal of Web Engineering and Technology*, 5(2):187–213, 2009.

[5] Jitsi (sip communicator). Available from: http://www.jitsi.org/ [cited July 19, 2011].

[6] Blender. Available from: http://www.blender.org/ [cited July 19, 2011].

[7] U.Norbisrath, K.Kraaner, E.Vainikko, and O.Batrashev. Friend-to-Friend computing - instant messaging based spontaneous desktop grid. In *Internet and Web Applications and Services, International Conference on*, volume 0, pages 245–256, Los Alamitos, CA, USA, 2008. IEEE Computer Society.

[8] C. Lattner and V. Adve. LLVM: a compilation framework for lifelong program analysis & transformation. In *Proceedings of the international symposium on Code generation and optimization: feedback-directed and runtime optimization*, 2004.

[9] J.Steele. comScore: mobile internet becoming a daily activity for many. Available from: http://www.comscore.com/Press_Events/Press_Releases/2009/3/Daily_Mobile_Internet_Usage_Grows [cited July 19, 2011].

[10] R.Edwards and L.Barker. *Developing Series 60 Applications: A Guide for Symbian OS C++ Developers*. Pearson Higher Education, 2004. Available from: http://portal.acm.org/citation.cfm?id=983988.

[11] Nokia developer - qt development frameworks. Available from: http://www.developer.nokia.com/Develop/Qt/Qt_technology.xhtml [cited July 19, 2011].

[12] J.Laurila, M.Marchetti, and E.martt. Python for s60. Available from: https://garage.maemo.org/frs/?group_id=854 [cited July 19, 2011].

[13] Home of the maemo community. Available from: http://maemo.org/ [cited July 19, 2011].

[14] Bionic c library overview. Available from: http://www.netmite.com/android/mydroid/1.5/bionic/libc/docs/OVERVIEW.TXT [cited July 19, 2011].

[15] Android NDK android developers. Available from: http://developer.android.com/sdk/ndk/index.html [cited July 19, 2011].

[16] SCons: a software construction tool. Available from: http://www.scons.org/ [cited July 19, 2011].

[17] SCons new target platform and new toolchain. Available from: http://buildman.net/sbuild_man/porting_eng/new_platform.htm#Step_2 [cited July 19, 2011].

[18] AllJoyn - proximity-based peer-to-peer technology. Available from: https://www.alljoyn.org/ [cited July 19, 2011].

[19] Simplified wrapper and interface generator. Available from: http://www.swig.org/ [cited July 19, 2011].

[20] python-for-android - Py4A - google project hosting. Available from: http://code.google.com/p/python-for-android/ [cited July 19, 2011].

[21] android-scripting. Available from: http://code.google.com/p/android-scripting/ [cited August 11, 2010].

[22] S. Wesner, T. Dimitrakos, K. Jeffrey, and H. Performance. Akogrimo-the grid goes mobile. *ERCIM news no59 2004*.

[23] C. Loos. E-health with mobile grids: The akogrimo heart monitoring and emergency scenario. *EU Akogrimo project Whitepaper*, 2006.

[24] D.P. Anderson. BOINC: A system for public-resource computing and storage. In *proceedings of the 5th IEEE/ACM International Workshop on Grid Computing*, pages 4–10. IEEE Computer Society, 2004.

[25] B.Guldbjerg Christensen. Lightpeers: A lightweight mobile p2p platform. In *Pervasive Computing and Communications Workshops, 2007. PerCom Workshops '07. Fifth Annual IEEE International Conference on*, pages 132 –136, 2007.

 ISBN: 978-1-61208-164-9

Use of Emerging Mobile Technologies in Portfolio Development

Ejaz Ahmed, Rupert Ward
School of Computing and Engineering
University of Huddersfield, UK
{e.ahmed, r.r.ward}@hud.ac.uk

Stephen White
School of Human and Health Sciences
University of Huddersfield, UK
stephen.white@hud.ac.uk

Abdul Jabbar
The Business School
University of Huddersfield, UK
a.jabbar@hud.ac.uk

***Abstract* - In the UK, implementing personal development planning (PDP) is an obligatory requirement across all Higher Education awards. This has led to a number of institutions requiring students to produce electronic portfolios to meet this requirement. However, far too little attention has been paid to utilising the powerful functionalities and high levels of connectivity of emerging mobile technology. This social study seeks to discover a potential role of emerging mobile technology in portfolio development and its effects on students' reflective capacity and engagement with PDP. To raise students' engagement with PDP, a mobile application (HUD iPDP) for Apple mobile devices was developed with fifty-one undergraduate students participating in this study. The data collected was both qualitative and quantitative. Results revealed a high level of interest among students and the potential for mobile technology to enhance the process of PDP.**

Keywords - Reflection; PDP; e-Portfolio; Mobile learning.

I. INTRODUCTION

With the rapid development of mobile technology, its suitability to learning activities is growing. It is providing students with access to truly mobile computers that fit in their hands and can go in their pockets. Mobile devices today are more powerful in functionality and connectivity than the desktop computers we used to have in the late 1990s. Students are attracted to new mobile phones because they are small, interactive and provide connectivity. Their ubiquity provides a valuable opportunity for educators to embed learning more effectively by enabling students to reflect at any point on their studies and development.

Personal Development Planning (PDP) is considered a significant pedagogical tool in higher education. It enhances the capacity for learners to reflect, plan and take responsibility for the primary objectives of PDP [1]. The traditional paper-based portfolio format has existed in HE in the past; however, the recent trend has been towards electronic e-Portfolio. The terms 'e-Portfolio', 'Progress File' and 'PDP' are often mentioned interchangeably in the literature [2]. JISC projects discovered that there have been tangible benefits in the use of e-Portfolios in relation to efficiency and enhancement in quality of PDP [3]. Most e-Portfolios are dynamic web applications using databases which enhance the quality of evidence, reflection, skills development and students' motivation. Emerging mobile technologies are equipped with hardware and software powerful enough to provide functionality and a high level of connectivity to easily augment existing e-Portfolios. As we use mobile devices in portfolio development, this can be described as m-portfolio (mobile Portfolio).

The aim of this paper is to discuss the outcome of a social study conducted to investigate the potential use of mobile technology in portfolio development. The paper evaluates the students' experience with PDP using smartphones, and with a bespoke mobile application to support PDP, which was developed and tested. The paper consists of four parts. First, it reviews the existing literature relevant to PDP, e-portfolio and role of mobile technology within this. Following this, the research method and procedures used in the study are presented. Next, results are discussed and summarised. Finally the paper concludes with a discussion on the implications, limitations and directions for further research.

II. CONCEPTUAL FRAMEWORK

The current policy on Personal Development Planning (PDP) emerged from the Dearing Report [4] which recommended that UK Higher Education Institutions (HEIs) should formulate a progress file, PDP, to enable students to 'monitor, build and reflect on their personal development' [4]. The Dearing Report advocated HEIs provide a mechanism for PDP but left the actual implementation to the discretion of individual institutions. The Quality Assurance Agency for Higher Education [5], who oversee its use, define PDP as 'a structured and supported process undertaken by an individual to reflect upon their own learning, performance and / or achievement and to plan for their personal, educational and career development' [5]. The concept of personal development itself had existed in many institutions [6] long before Dearing's recommendations, with the idea of a 'reflective practitioner' [7] already popular in nursing and teaching professions for example.

Reflection is a key element of the process of PDP and acts as a vehicle for turning 'experience into learning' [8] by combining different thoughts and ideas together. This personal experience in combination with formal learning results in 'deep' learning [9]. The QAA guidelines for PDP state that reflection is 'a process that involves self-reflection, the creation of personal records, planning and monitoring progress towards the achievement of personal objectives [5]. Boyd & Fales [10] define reflection as 'a process of internally examining and exploring and issue of concern triggered by an experience, which creates and clarifies meaning in terms of self and results in a changed conceptual perspective'. It has been suggested that reflection process based on personal experience at regular instances enables

 ISBN: 978-1-61208-164-9

students to clarify for themselves the process of development.

A considerable amount of literature has been published suggesting benefits from implementing e-portfolios [9, 11, 12]. Using a paper-based portfolio (PBPs) has been an approach practiced by some disciplines in HE such as nursing, teacher training, art and finance. However, electronic portfolios are becoming more commonplace as technology advances. Electronic portfolios or web-based-portfolios (WBPs) are preferred over paper-based portfolios (PBPs) because they enhance students' motivation and are more user-friendly [13]. Madden [14] described an e-portfolio as 'an archive of material, relating to an individual, held in a digital format'. Many projects funded by JISC [3] discovered that e-Portfolios enhance the quality of evidence, reflection and the skills development process. An electronic portfolio saves time in information retrieval, supports reflection, raises presentation and improves students' motivation for PDP. It gives students an opportunity to customise the PDP and increase their ability to share and transfer information more conveniently. Research shows that time spent on PDP increases significantly with the use of web-based portfolios as compared to paper-based portfolios [13].

Advances in mobile technology are changing the pedagogical possibilities of 'Mobile Learning'. Research suggests mobile technology can enhance various features of teaching and learning such as reducing the time for tedious work, engaging students in learning activities, facilitating group collaborative learning, empowering the teacher to monitor students' learning progress and recording teaching and learning processes as portfolios [15]. The positive implications of e-Portfolios and pedagogical possibilities of new mobile technologies can be used to enhance the process of PDP by using it in portfolio development.

III. METHODOLOGY

The target population for this study consisted of first year undergraduate students in the School of Computing and Engineering at the University of Huddersfield. A sample of 74 randomly chosen students was divided into three groups; group A comprised 27 students with Apple mobile devices (iPhone, iPod Touch and iPad), group B consisted of 27 students with non-Apple smartphones; whilst group C contained a control group of 20 students. The control group was not introduced to the study until their views were collected in the form of questionnaires, interviews and a focus group session. From each group, eight participating students were randomly chosen for interviews and eight for focus group sessions. This selection was made from the students who completed the online survey. The length of this study was approximately eight weeks, which started from the first week of the students' academic year in university.

In order to evaluate the students' perception of using mobile technology to enhance PDP, a mobile application for Apple mobile devices (iPhone, iPod Touch and iPad) was developed (Figure 1). The selection of Apple mobile devices for this study was made due to their high level of functionality, reliability, usability and design. They also were more popular amongst the student population at the time of the app development. The aim of the development was to develop an attractive tool that would enrich teaching and learning by providing students with an engaging means of creating, adding and accessing PDP contents on a mobile device. The application was introduced to the users of the Apple mobile devices during the first week of their academic year. No training was given to the HUD iPDP users; however, a user guide was made available to them via the Blackboard VLE, used at the University of Huddersfield.

A questionnaire was generated and pre-tested using a convenience sample of 10 second year IT students using the method described by Cooper and Schindler [16] called collaborative participant pretesting. Data for the main study was collected using an online controlled questionnaire during week 8 of the students' academic year. Incentives in the form of books were provided to participants in acknowledgement of participation in this study and to compensate for the time taken, but they were not promised such incentives before the experiment. One week after the initial call for completion of the controlled online survey, a reminder email was sent to the participants who had not completed the survey. Interviews and focus group sessions were arranged during week 8 and 9.

In the questionnaire, 29 multiple choice questions including demographic questions were set. Most open ended questions from the online questionnaire were also included in the list of discussion topics for the focus group sessions. A few questions were also further explored during the one to one interviews. The following four key questions were asked in this study:

- Which method would you prefer to complete your PDP?
- Do you think the mobile devices can raise your motivation by providing access to your PDP anywhere and at any time?
- Regardless of the mobile device you are using at present, what features would you like to have and what services would you like to access via a mobile device?

Figure 1. HUD iPDP application

 ISBN: 978-1-61208-164-9

- What are your major concerns about using mobile devices in portfolio development?

All responses to the questionnaire, received from three groups, were classified separately to analyse the differences. Students in group A (users of Apple mobile devices) were also asked the following additional questions related to the HUD iPDP application:

- Do you think the HUD iPDP app has helped you in updating the contents of your PDP?
- Was it easy to collect the contents in the form of text, audio, image and video for your portfolio?
- Which features of the HUD iPDP did you find useful?
- Which features of the HUD iPDP did you not like?
- Please provide any further suggestions to improve the app

The last three questions were open questions to collect qualitative information about the developed application.

IV. RESULTS AND DATA ANALYSIS

The response rate from the online questionnaire, interviews and focus group sessions was good. Fifty one students responded to the questionnaire representing approximately 20% of the entire cohort and 69% of the sample group of 74 students. The majority (88%) of the respondents were between the ages of 18 and 25. Table I shows cross-tabulation between groups and their response rates in detail.

The results of the questionnaire show that most students are eager to use an online portfolio system. Those with smartphones had the greatest tendency towards the use of mobiles in the PDP process, perhaps because of their exposure to the app. A multiple choice question was asked from all participants to know their preferred method to complete PDP. Overall, a large majority chose an online portfolio system (67%) followed by an offline electronic portfolio system (53%). However, on analysing the individual results from each group it revealed that the preferred method to work with PDP for the students in group A was online using a PC or laptop (83%) followed by using mobile devices (50%). Table II provides us more detail on students' preferred method to complete the PDP and a visual representation can be seen in Figure 2.

The results summarised in Table II were further explored in interviews and focus group sessions, which revealed that low scores for using mobile devices to organise PDP were

TABLE I. RESPONSE RATES

	Survey	Interview	Focus Group
APPLE	18 (67%)	8 (100%)	7 (88%)
NON-APPLE	17 (63%)	6 (75%)	5 (63%)
CONTROL GROUP	16 (80%)	8 (100%)	8 (100%)
	51 (69%)	22 (92%)	20 (83%)

TABLE II. PREFERRED METHOD TO COMPLETE PDP

	Paper-based	On a PC/laptop offline	Online using a PC/laptop	On an internet enabled mobile device
APPLE	6%	33%	83%	50%
NON-APPLE	6%	47%	71%	35%
CONTROL GROUP	13%	81%	44%	25%
TOTAL	8%	53%	67%	37%

as a result of a lack of synchronisation functionality in the HUD iPDP for the group A, and because of the lack of availability of any suitable application for the students in group B, who were using other mobile devices. Group A was using the HUD iPDP application, which was only helping students in data collection. No online platform was available to students to sync data automatically. Applications were not able to communicate with the Blackboard portfolio system due to a number of security issues. Moreover, it was not compulsory for the students to use Blackboard but they were allowed to create their own online portfolio or use any open source portfolio system available online.

All the students who participated in the focus group sessions and attended interviews suggested that they would have used the HUD iPDP application if more synchronisation functionality had been made available to them. Although two students in group B indicated that they used their mobiles in portfolio development, the large majority expressed disappointment with the unavailability of an appropriate application. Low tendency for using mobile devices in portfolio development among group C was because the idea of using a mobile portfolio was new to them. By comparing the results, it can be seen quite clearly that the students in group A are in favour of using mobiles in portfolio development compared to groups B and C (Figure 2) which is positive indication given the above mentioned grounds.

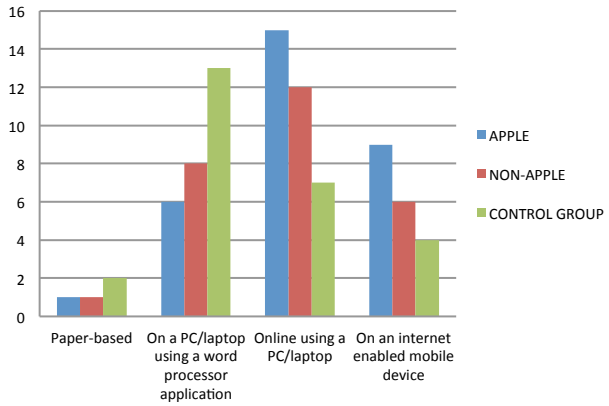


Figure 2. Preferred Method to Complete PDP

 ISBN: 978-1-61208-164-9

The second key question asked was about students' views on the potential of mobile devices in raising motivation for PDP. Four possible answers were given to choose from. Table III shows cross-tabulation survey results from each group, also shown graphically in Figure 3. It is quite clear from the Figure 3 that students are enthusiastic about the use of mobile technology and support its use for portfolio organisers.

The next key question was about the desired feature students would like to have in a mobile application. The data collected suggests that students were keen to see a number of other features to support their studies such as access to Blackboard, learning resources, lecture notes, class timetable, assignment deadlines, feedback on assignments, library catalogue and many more. The overwhelming emphasis, in students' feedback, was that they valued the affordances of mobile technology and were enthusiastic about using it in their university experience.

The last key question asked was about the major concerns on using mobile devices in portfolio development. Syncing data, interactivity, content quality, speed, reliability and security were the options, which were rated (on a scale of 1-5, 1 is lowest and 5 is highest). Students in all groups considered them equally important.

Additional questions were asked from group A (users of the HUD iPDP app). Eighteen survey responses were gathered of which 15 (83%) participants still had the HUD iPDP installed on their mobile devices. Out of 15 students, 14 (93%) used the HUD iPDP application to collect the content for their PDP. Most students found the various features easy to use and were satisfied with the application in general. However, a few students also pointed out in the focus group session and interviews that lack of training in PDP and unavailability of an online version of application with data sync functionality caused low level of engagement with the application. Out of 14 students, 8 (57%) believed that HUD iPDP app helped them in content collection, however, 6 students (43%) did not find it useful. Open questions in the survey, interviews and focus group session revealed that the primary reason for less interest among the students was the data transfer issue from mobile devices to e-portfolio.

A number of issues were identified from the interviews and focus group session with all three groups. This study clearly discovered a demand from students for a coherent multi-functional application with synchronicity and availability of an appropriate application to support the PDP process in different devices. Students were enthusiastic about availability of learning resources on their mobile devices.

TABLE III. VIEWS ON POTENTIAL OF MOBILE DEVICES TO RAISE MOTIVATION FOR PDP

	Yes	Probably Yes	Probably No	No
APPLE	56%	44%	0%	0%
NON-APPLE	18%	59%	24%	0%
CONTROL GROUP	25%	69%	0%	6%
TOTAL	33%	57%	8%	2%

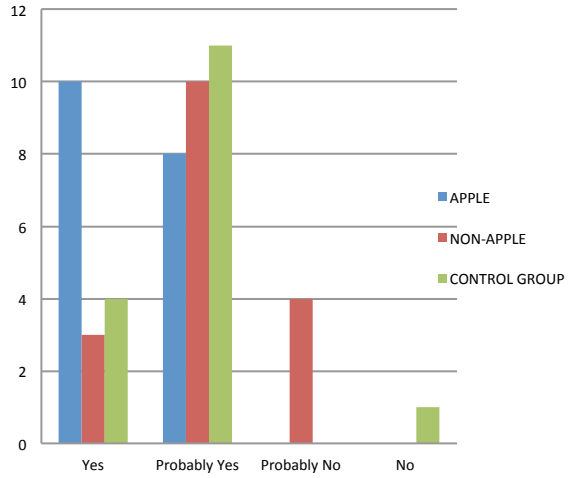


Figure 3. Views on potential of mobile devices to raise motivation for PDP

V. CONCLUSION AND FUTURE WORK

The research objective was to uncover the potential of mobile technology in portfolio development. A mobile application was developed for PDP to provide an interface for Apple mobile devices and tested to analyse its potential impact on portfolio development. It has been revealed from the students' responses that there is huge interest in the use of mobile technology in this domain. Using mobile devices enhances student motivation, quality of e-portfolio and can improve ease of reflection. However, research on m-portfolios is still in its infancy and needs extensive pedagogical research. More research is required for example to discover if support of m-portfolio for e-portfolio makes a difference in students' engagements with their studies across a range of portfolio approaches. As mobile technology is becoming ever more accessible to students, the knowledge base in this domain needs expanding to understand its true value. Future work would aim to develop guidelines for the use of m-portfolio applications.

A new phase, an m-portfolio project, has started. In addition to making changes in the HUD iPDP application, a bespoke web application will be developed for e-portfolio. This will assist in resolving the synchronicity issue faced with Blackboard e-portfolio system. Another pilot study will be carried out during the next academic year in order to conduct thorough functionality, usability as well as pedagogical evaluations. Feedback via questionnaires, focus groups and interviews will be collected and analysed. One of the main aspects of research will focus on whether using mobile devices in portfolio development raises students' engagement with PDP and enhances its content quality.

REFERENCES

[1] I. Barbu, D. Popescu, L. Zadeh, J. Kacprozyk, N. Mastorakis, A. Kuri-Morales, P. Borne, and L. Kazovsky, "Personal

development - connecting element between human resource and career development," *Recent Advances in Business Administration,* 4th WSEAS International Conference on Business Administration (ICBA '10), pp. 228-232, 2010.

[2] J. Rowe, "Evaluating a new e-PDP tool and its relationship with personal tutoring," *Journal of Learning Development in Higher Education, Special Edition: November 2010,* pp. 1-18, 2010.

[3] JISC infoNet, (2009). *Why Use e-Portfolios*. Retrieved from: <http://www.jiscinfonet.ac.uk/infokits/e-portfolios/why-use> on 12 October 2011

[4] NCIHE, "Higher Education in the Learning Society (Dearing Report): Report of the National Committee of Inquiry into Higher Education Academy," HMSO, London, 1997.

[5] Quality Assurance Agency for Higher Education, (2001). *Guidelines for HE Progress Files*. Retrieved from: <http://www.qaa.ac.uk/Publications/InformationAndGuidance/Documents/progfile2001.pdf> on 12 October 2011

[6] A. Slight and S. Bloxham, "Embedding personal development planning into the social sciences," *LATISS: Learning and Teaching in the Social Sciences,* vol. 2, pp. 191-206, 2005.

[7] D. A. Schön, *Educating the reflective practitioner*: Jossey-Bass, San Francisco, 1987.

[8] D. Boud, *Reflection: Turning experience into learning*: Routledge, 1985.

[9] N. J. Entwistle and P. Ramsden, *Understanding Student Learning*: Croom Helm, London, 1983.

[10] E. M. Boyd and A. W. Fales, "Reflective learning: Key to learning from experience," *Journal of Humanistic Psychology,* vol. 23, pp. 99-117, 1983.

[11] G. Roberts, W. Aalderink, J. Cook, M. Feijen, J. Harvey, S. Lee, and V. Wade, "Reflective learning, future thinking: digital repositories, e-portfolios, informal learning and ubiquitous computing," ALT/SURF/ILTA1 Spring Conference Research Seminar, Trinity College, Dublin, pp. 1-13, 2005.

[12] P. Butler, "A review of the literature on portfolios and electronic portfolios," *Technical report, eCDF ePortfolio Project,* 2006.

[13] E. W. Driessen, A. M. M. Muijtjens, J. Van Tartwijk, and C. P. M. Van Der Vleuten, "Web or paper based portfolios: is there a difference?," *Medical education,* vol. 41, pp. 1067-1073, 2007.

[14] T. Madden, *Supporting Student e-Portfolios*: Higher Education Academy Physical Sciences Centre. Retrieved from:<http://www.heacademy.ac.uk/assets/ps/documents/practice_guides/eportfolios_JISC.pdf> on 12 October 2011

[15] J. C. Yang and Y. L. Lin, "Development and Evaluation of an Interactive Mobile Learning Environment with Shared Display Groupware," *Educational Technology and Society*, vol. 13(1) pp. 195-207, 2010.

[16] D. R. Cooper, P. S. Schindler, *Business research methods.* New Delhi, Tata McGraw-Hill, 2004.

 ISBN: 978-1-61208-164-9

A Dynamic Approach for User Privacy Management in Location-based Mobile Services

Amr Ali-Eldin
Ordina ICT B.V.
Management & Consultancy
Ringwade 1, 3439 LM, Nieuwegein, the Netherlands
Tel.: (+31)30 663 7315
Fax: (+31)30 663 7496
e-mail: amr.ali-eldin@ordina.nl

***Abstract*— Recently, we have noticed the wide spread of GPS enabled mobile phones, which enable mobile applications to track users locations and start pushing customized advertisements to them. For a small benefit a user might get from these Ads, a user might be willing to share his or her location without even knowing the impact this might have on his or her privacy. In this paper, we propose a dynamic approach for evaluating those coming requests for users' locations based on users pre-described privacy preferences by providing users with what we call a *Privacy Threat Level* (*PTL*) indicator. We have developed a simulation console and presented a scenario showing how this approach can work in practice.**

Keywords-Privacy; User preferences; Information Collectors; Smart Phones; Service Providers; Pervasive Computing; Location-based services (LBS)

I. INTRODUCTION

Recent development in pervasive computing have paved the way for the deployment of pervasive and ubiquitous services [1]. We have also seen how the introduction of the latest technology of smart phones like iPhone 4, Blackberry, Android, iPads etc. has led to a complete set of location-based services (LBS) capabilities like road navigators for example. LBS collect and use users location to provide new or improved services [2]. Despite the benefits these services can bring to users and stakeholders, they pose a threat to user privacy. We have also noticed that most smart phones now come with a built-in GPS capability, which makes it possible for mobile applications to get users location and start pushing advertisements and services. According to a recent survey by the Mobile Marketing Association (MMA) [3], about two thirds of iPhone owners now use location-based services at least once a week mostly to locate nearby points of interests, shops and services. Location information may be collected rather unobtrusively or passively and used by service providers without users' notice or informed consent and that represents a real threat to user privacy.

Consider the case, a system engineer Jo works for a system developing international company, which supports different oil refining sites located in the sea. Jo has a smart phone with an application installed that is called BeThere. BeThere provides Jo with the logistic services to help him with his work activities and guarantee his safety. If Jo wants to leave one site to go to another, he plans his trip via BeThere. A helicopter comes to pick him up from the place where he is. BeThere also has business partners near each location: tourist guides, hotels and restaurants. BeThere keeps a profile of Jo, which got Jo's personal information such as name, identity etc., payment information, location information, and calendar information.

BeThere business partners or simply third parties will also like to have some of Jo's private information for their services provisioning or promotions even though they are unknown to Jo. This can mean that Jo will not know that they collect private information. Furthermore, Jo will not be able to know, which party collects what information from BeThere even when he is triggered by their push services or Ads. Tourist guides will like to gather Jo's personal, and location information to provide customized guiding. They will collect payment information as well. Hotels will like to gather Jo's identity information and payment information to recommend accommodation in each location. Restaurants will like to gather context information: location, eating preference, and schedules to provide suitable meals (see Fig. 1).

Although Jo's first privacy preferences will be that no third parties can have access to his information, Jo will be interested to use specific services depending on his situation. It is not that he gets push services that he will not be interested in. Sometimes there might be an interesting pop-up with a nice offer, which one cannot refuse. For example, when Jo enters a restaurant, he wouldn't mind it if the restaurant sends him an offer of what they can offer of drinks with a special price. Most of the times we see that LBS services are based on opt-in subscription from the customer. But what we also can see is that these types of services are pushed to customers in a dynamic way. Jo is not against that but would like the process used to get these services to be reliable, simple, flexible and safe. Jo, as many others, is very concerned about having control of his privacy at anytime and everywhere specially with the spread of such push services.

Accordingly Jo, as a customer of BeThere, will like to be able to express his privacy preferences when using BeThere's services. Jo will initially allow travel agencies and tourist guides to have access to his information while entertainment providers will be blocked. Additionally Jo would like to be informed when there is a privacy breach and

 ISBN: 978-1-61208-164-9

to intervene. Last and not least, he wants to be able to change his preferences at any time, which makes the process of managing his privacy preferences complex.

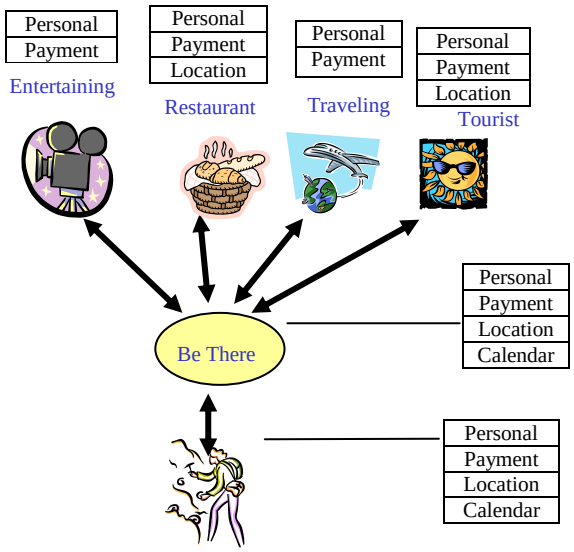


Figure 1. Jo and *BeThere*

The rest of this paper is organized as follows. In Section II, we discuss existing approaches and related work. Then in Section III, we discuss the most relevant principles of the P3P platform. Next, we introduce our proposed Privacy Threat Level *(PTL)* approach in Section IV. Thereupon in Section V, we present a prototype example based on what the concepts where prototyped. We finalize by presenting our conclusions in Section VI.

II. EXISTING APPROACHES AND THEIR LIMITATIONS

Privacy threats emerge as a result of the linkage between user identifying information and his or her context-related data. Therefore, most literature has focused on the separation between both types of information when dealing with privacy issues: whether to control users identities, by controlling identity capturing through the use of anonymity solutions as in [4-8], or by access control mechanisms like [9-12], distributing and encrypting of data packets in [13] and physical security through limiting data access within a specified area in [14].

Most of these approaches presented so far focus on conventional data management techniques, which are static [5, 9-11, 13, 15-17]. In other words, they are not aware of user context. Knowing user's context, it may be possible to recover his or her identity even if his or her real identity itself is not communicated. For example, if an anonymous user (on a chat site) tells someone that he was working for a company X from year 2000 till year 2006. Then, his identity is now limited to employees of company X till year 2006. Knowing employees who left company X in year 2006 and knowing the user's current location and or personal interests can help reveal that person's real identity. Therefore, we like to argue that not only user identity information but other information with different degrees of confidentiality should be protected as well, which in turns represent the context of that user.

Controlling the full collection of user contexts may represent the most realistic approach in pervasive environments towards user privacy protection as we have seen in Jo's case. This can be achieved by either reducing the accuracy of the collected data as in [18] or by enforcing user decisions of whether to allow user context to be collected by a certain party. In order to do so, information collectors' ways of dealing with the user contextual information need to be communicated to the user to be able to make a decision. Besides that users should be able to describe their preferences when it comes to their private information. In Jo's example, when he receives a pop-up pushing some nice meal or drink asking for his location, Jo would like to control who else can get this information. One of the leading efforts in this approach, the platform of privacy preferences (P3P) [19] has defined a way of describing information collectors / service providers data practices that constitute a P3P privacy policy. Each practice possesses a descriptive value that is defined by APPEL, the P3P Preference Exchange Language 1.0 [20], which was proposed as the language for expressing user preferences. We think that a P3P based description of user preferences is considered insufficient for describing a dynamic data enriched environment such as the pervasive and mobile environment because it is focused on Internet applications and may not have support to dynamic situations as we will see later in this paper.

One of the well-known P3P based privacy preferences description approaches is 'AT&T privacy bird' [21]. AT&T privacy birds help Internet users to stay informed about how information they provide to Web sites can be used. An AT&T Privacy Bird automatically searches for privacy policies at every website a users visits and asks users for their privacy strictness levels. They can also customize their preferences themselves by importing an XML pre-defined preferences list. To the best of the author's knowledge, the AT&T privacy bird is not designed to deal with mobile and pervasive environments.

Based on the above-given review of previous research results, we argue that there is a need for the development of a flexible approach for privacy that can deal with the dynamics as present in pervasive and mobile computing environments. By preference, such models should be consonant with existing successful, de facto standard platforms for privacy preferences. In this paper, we adopt the P3P as reference model and add some enhancements to suit with dynamic environments.

III. THE PLATFORM OF PRIVACY PREFERENCES (P3P)

P3P [19] has defined a number of data practices that together constitute a P3P privacy policy. A Privacy Policy is a collection of both vocabulary and data elements that describe the data practices of particular website (or section of a web site). A Privacy Policy includes a sequence of statement elements that may have the following sub elements:

 ISBN: 978-1-61208-164-9

- *Purpose:* A purpose is represented in the P3P syntax as a *PURPOSE* element. Each *PURPOSE* element can contain one or more sub elements that describe a site's reasons for collecting the information. The P3P vocabulary defines twelve kinds of purposes.
- *Recipient:* The recipient defines the party with, which the collected data will be shared. *Recipient* is represented in the P3P syntax as a *RECIPIENT* element, which can contain one or more sub elements that describe kinds of recipients. The P3P vocabulary defines six types of recipients.
- *Retention:* Retention defines the duration for, which the collected information will be kept. Retention is represented in the P3P syntax as a *RETENTION* element, which can contain one or more sub elements that describe kinds of retentions. The P3P vocabulary defines five types of retentions.
- *Consent Behaviour:* The consent is defined in P3P to be of three kinds; *request, limited* and *block*. A request consent means complete agreement from the user, and a block consent means no agreement at all. A limited consent, however, assumes consent with blocking identification information from transmission.

Given these data elements, a typical P3P model for users' privacy preferences (in terms of consent decisions to be made) is based on the following rule description:

{<*purpose*>, <*recipient*>, <*retention*>} ➔ *user consent behaviour*

A. P3P Limitations

As we have discussed above, privacy preferences are used to describe users' allowed data practices, i.e., they define what users allow the service providers or information collectors to do with their information. A user may specify privacy preferences written in APPEL [20]. The process of writing users preferences using APPEL rules that function properly is cumbersome due to some limitations and shortcomings in the APPEL language design principles [22]. One of these shortcomings is that people cannot specify what is acceptable rather than specifying what is unacceptable. It is not easy to write an APPEL statement that defines request consent for a specific behaviour of a service provider. Agrawal, Kieren et al. [22] argue that even exact connectives (or-exact, and-exact) will result in incorrect behaviour when being used to avoid this problem. Therefore, they proposed Xpref based on Xpath [23] to replace APPEL specifications. Though the approach of correcting APPEL seems a fundamental one [22], we assume that replacing APPEL is a process that will take a long time and needs a lot of effort as well. In this paper we will adopt another approach here by making use of the so-termed PTL Approach.

IV. THE PTL APPROACH

In this section, we present a way of dealing with dynamics through asserting a PTL value to data practices combinations. At each moment in time, we get an updated user consent decision that corresponds to the dynamics caused by the change of user situation or location.

A. Calculating aggregated PTL Values

Information collectors such as service providers usually present one single list of practices, expecting users to either accept it or reject it as a whole. However in practice, different combinations of data practices as offered in a service provider's privacy policy can have different impacts on user privacy concerns. This impact may vary from one user to another and from one context to another. In our approach, the difference in impact on privacy is expressed in the form of a numeric value, which is a weighting value reflecting the threat a particular request for information poses to a person's privacy. The PTL is calculated by dynamically evaluating the service provider data practices.

The PTL always has a value between 0 and 1: the higher the value, the higher the underlying threat to privacy. For example, if the threat value of requests with telemarketing purpose is set to 0.8 and that of contact to 0.5, this means that collecting user data for telemarketing is considered more invasive than for contact. What the user specifies in his or her preferences using the PTL approach is how he or she thinks a telemarketing purpose is threatening to his privacy concern. Average users are unlikely to understanding P3P vocabularies, and as a result there is a possibility that the values they define do not accurately reflect what they want. This means that a way has to be found to carry out the weighting process in a user-friendly way, taking into account the changing domain specifications.

As argued above, each request may pose a threat to privacy depending on how the requested information will be dealt with. In other words, depending on what we call allowed data practices compared to asked ones. However, combinations of practices can have different impacts on privacy. For example, the purpose of 'individual analysis' can have a lower PTL value if combined with a recipient of 'ours' rather than that of 'unrelated third parties'. Here, 'ours' may represent the set of family people or group of close friends and 'unrelated third parties' may denote the set of non-business partners. Fig. 2 shows an example of how practices' combinations can affect the various PTLs: the tailoring purpose PTL equals 0.6 and if combined with other data practices, the overall combination has different aggregated PTL values. For example, a "tailoring, ours" combination has the lowest combined PTL in violating privacy compared to the "tailoring, third parties" combination.

For the sake of simplicity, in this paper we assign PTL values per combination of practices rather than per single practice. The final aggregated PTL of a certain engagement of a service provider is thus composed based on the aggregation of all PTLs per practices' combinations.

 ISBN: 978-1-61208-164-9

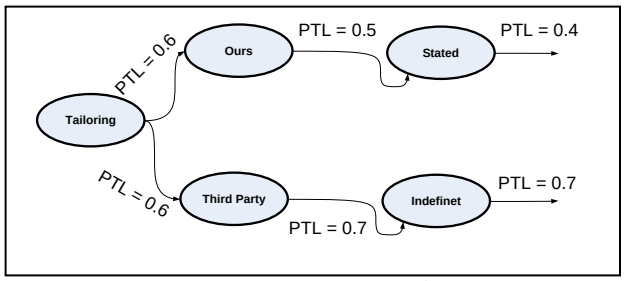


Figure 2. Aggregating weight combinations of data practices yielding different PTL values.

B. Privacy Rules Design

In the previous subsection, we argued that combinations of data practices influence PTL values of given services providers' data practices. In this section, we elaborate further on this issue. Preferences description refers to the way preferences will be presented to the end users. In this paper, our proposed model is based on the P3P specifications' one. Service providers' data practices are also expressed using P3P vocabularies [19]. Our proposed model uses PTL values as a representation of how users think their privacy can be violated. We can define the PTL model in the following way:

{*<purpose>*, *<recipient>*, *<retention>*, *<situation>*} ➔ *PTL*

The underlying rationale for choosing this rule description is to make it more user-friendly. This way one user can indicate a PTL value instead of having to worry about making a consent decision herself. Before a user decides whether to give consent or not, he or she has first to think whether this rule is threatening his or her privacy by assigning a PTL value. He or She can afterwards decide based on the aggregated PTL values whether to give consent or not. If we look again at the previous example adding the situation to the model has impacted the overall PTL (see Fig. 3).

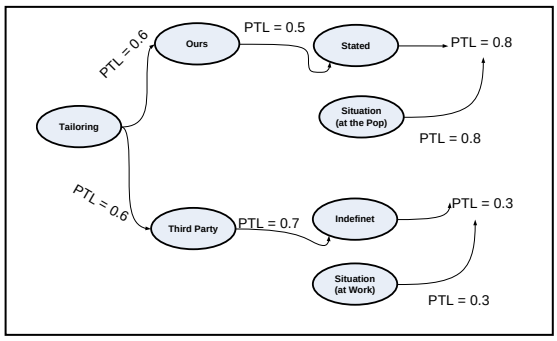


Figure 3. Aggregating the Situation Impact on PTL

We assume that a privacy preference consists of one or more privacy preferences rules (statements). Each rule consists of a specific data practices combination, associated consent behaviour, and a PTL value. The statements are connected via connectives as described below. Our proposed privacy preference rule, or just referred to as privacy rule in the rest of the paper, can be described as follows:

{<Purpose (n)>, <Recipient (n)>, <Retention (n)>, <consent behaviour>, <PTL (n)>}, <Rule Connective>

C. Rule Connectives

Rule connectives are logical operators that define the influence of each privacy statement on the others in order to evaluate the overall behaviour. We adopt the P3P defined connectives; AND, OR, NON-AND, NON-OR, AND-EXACT and OR-EXACT. The connectives govern the way preferences statements are compared to those in the privacy policy as follows:

1) AND

A rule will fire only and only if all contained statements are found in a privacy policy and matched.

2) OR

Any match of the contained statements is enough for firing**.**

3) NON-AND

Any of the contained statements should not match (logical complement of AND).

4) NON-OR

None of the contained statements should match (logical complement of OR).

5) AND-EXACT

All contained statements should match in the privacy policy of the collector for acceptance and no other statements (not matched) should exist in the privacy policy.

6) OR-EXACT

A match of any of the contained statements is enough to fire both and no other statements should exist in the privacy policy.

D. Privacy evaluation mechanism

As mentioned above, a privacy rule is a statement specifying a PTL value associated with a certain data practices combinations and associated situation. Furthermore, situation also influences PTL values. The next step is to specify the evaluation mechanism needed to automate the process of assessing PTL values. A weighting analyzer will be needed to develop an output that consists of {consent, PTL} for example, {request, Low}, which means a consent type of request with a PTL value of Low. The consent here refers to the output coming from APPEL evaluation. The weighting analyzer will look for practices combinations in the policy and accumulate the overall PTL value. Within the weighting analyzer, evaluation takes on the following pattern: first find available combination matches and then accumulate weights according to matching combinations.

 ISBN: 978-1-61208-164-9

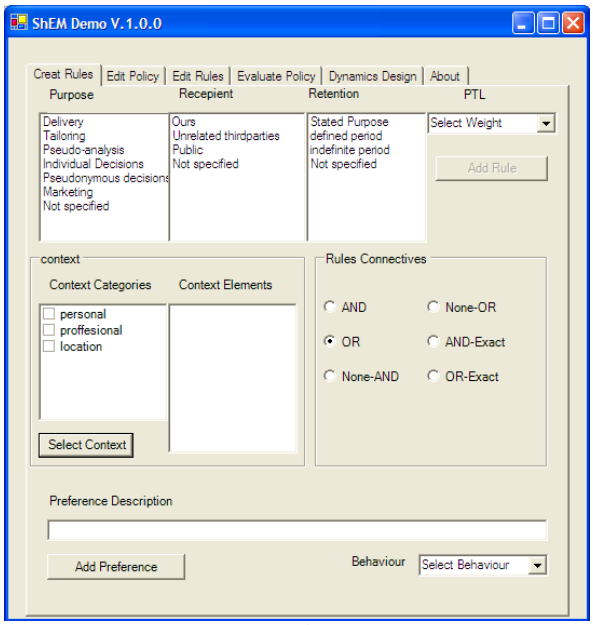


Figure 4. Assign Preferences Form

V. A PROTOTYPE EXAMPLE

We have built a prototype console using visual basic.Net in order to simulate the evaluation process of the proposed privacy rules using the PTL approach. Fig. 4 shows how users would have to assign preferences. In this prototype, we provide two ways of expressing user preferences; a static and a dynamic way. In the static way, preferences consist of allowed practices combinations associated with a PTL value. While in the dynamic way, we associated PTL values with the user situation. In this console, and for the sake of simplicity, we used a PTL range from 1 to 10. The user could assign data practices values per context group as well. We defined three context groups; personal, professional and location. Each preference could have multiple data practices combinations. Each preference is assigned a behaviour value. The user can define the rule connective among the six defined connectives in the P3P specifications. Fig. 5 shows how the form of assigning dynamic preferences can look like.

Let us get back to Jo's case as was shown in Fig. 1. For the sake of simplicity, we assume the following:

- We take an example of only location type of requests.
- *PTL* values are accumulated using an OR logic meaning that we take the highest value among the matching rules connected with the OR connector.
- Low *PTL* values means $PTL <= 4$, Medium *PTL* values means $4 < PTL <= 7$, High *PTL* means $PTL > 7$.
- Jo's situation is classified to three situations: {"in my Room"; "in the Hotel's Bar / restaurant"; "at the client"}. Associated PTL values are {8,4,2}.

Jo classifies his allowed data practices combinations (privacy preferences / rules) as shown in Table I while Information collectors asked data practices for location information are shown in Table II.

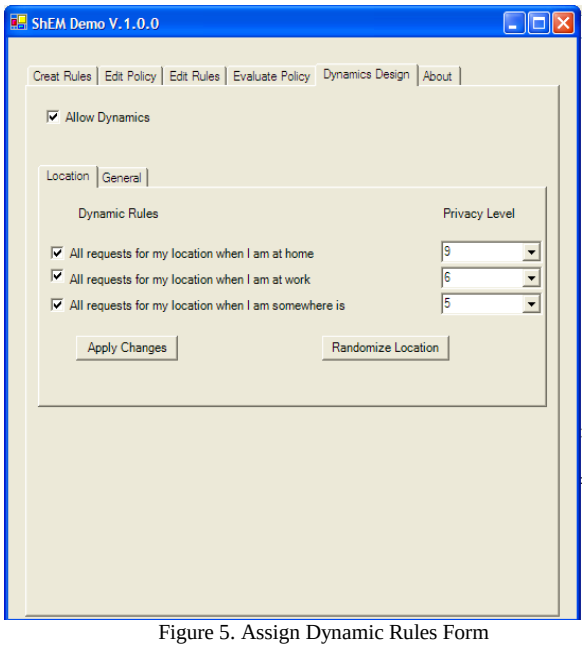


Figure 5. Assign Dynamic Rules Form

TABLE I. JO'S PRIVACY PREFERENCES

Purpose	Recipient	Retention	Consent	PTL
Not specified	Not specified	Stated Purpose	Request	3
Not specified	Not specified	indefinite period	Block	9

TABLE II. INFORMATION COLLECTORS ASKED DATA PRACTICES

Information Collector	Purpose	Recipient	Retention
Entertaining	Marketing	Not specified	indefinite period
	Delivery	Ours	Stated Purpose
Traveling	Not specified	Not specified	indefinite period
Restaurent	Marketing	Not specified	Stated Purpose
Tourist	Marketing	Not specified	Stated Purpose
BeThere	Not Specified	Ours	Stated Purpose

In case *BeThere* requests location of Jo, then this is what happens:

- The weighting analyzer collects BeThere asked data practices. These are: {Not Specified, Ours and Stated Purpose}
- The weighting analyzer checks Jo's allowed data practices and associated *PTL*s. According to APPEL evaluation, the output consent behaviour becomes "Request" while the accumulated *PTL* value is "3". Then the weighting analyzer checks for Jo's current location that is asked by BeThere, let's assume that Jo is at the the restaurant, this has

 ISBN: 978-1-61208-164-9

a *PTL* value of "4". This leads to a final *PTL* of "4". The recommended output becomes then {"Request", "Low"}.

- In case Jo location changes, for example to "in my room", *PTL* becomes "8", having said that, then the final *PTL* becomes "High". Though APPEL output doesn not change and remains "Request", Jo should reject this transaction at that time because his situation being at his room is considered highly private by him.

The rest of the information collector's evaluation takes place similarly. The evaluation output is displayed in Tables III & IV. Table III shows the output behaviour without taking Jo's situation into account while Table IV shows the evaluation taking Jo's situation into consideration. For example, collection of Jo's location is rejected for indefinite retentions although the other allowed data practice of Entertaining has an output of {"Request" & "3"}, the final static evaluation for Entertaining would be: {"Block", "9"}. When taking dynamics into account, Jo's evaluation does not change much because the static evaluation scored already privacy threat when dealing with Entertaining service provider.

TABLE III. STATIC EVALUATION OF JO'S PRIVACY

Inf. Collectors	Purpose	Recipient	Retention	Fired Conse-nt	*PTL*
Entertaining	Marketing	Not specified	indefinite period	Block	9
	Delivery	Ours	Stated Purpose	Reques t	3
Traveling	Not specified	Not specified	indefinite period	Block	9
Restaurent	Marketing	Not specified	Stated Purpose	Reques t	3
Tourist	Marketing	Not specified	Stated Purpose	Reques t	3
BeThere	Not Specified	Ours	Stated Purpose	Reques t	3

TABLE IV. DYNAMIC EVALUATION OF JO'S PRIVACY

	In my room		In the hotel's bar / restaurant		At the client	
	APPEL Output	*PTL*	*APPEL Output*	*PTL*	*APPEL output*	*PTL*
Entertaining	Block	9	Block	9	Block	9
Traveling	Block	9	Block	9	Block	9
Restaurent	Request	8	Request	4	Request	3
Tourist	Request	8	Request	4	Request	3
BeThere	Request	8	Request	4	Request	3

From the above tables we notice that for some cases the static evaluation scored already a threat when the service provider is intending to keep Jo's details for indefinite period of time as in case of *Entertaining* and *Travelling* ones (see Table III). Therefore the dynamic evaluation will not be expected to differ. In other cases, the static evaluation can allow the transaction while the dynamic one detects the threat. For example in the case of "BeThere" and when the situation changes to "in my room" as shown in Table IV the final evaluation has shown a high *PTL*, which means we should update the static *PTL* value. In this case, Jo should get a message warning him from continuing this operation.

VI. CONCLUSIONS AND DIRECTIONS

In this paper, we proposed a privacy control approach for location-based services (LBS), which takes into account the dynamics of such environment into the design of users' privacy rules. To do so we proposed a privacy threat level indicator (PTL) to be inserted in rules description. PTL refers to the amount of threat expected to user privacy when using a data practices combination. We also porposed the way to evaluate privacy rules using both APPEL and PTL approach. We have developed a simulation console that shows how dynamic preferences are described and evaluated in practice. We also presented a scenario showing how this approach can work in practice.

Average users are unlikely to understanding P3P vocabularies, and as a result there is a possibility that the values they define do not accurately reflect what they want. This means that a way has to be found to carry out the weighting process in a user-friendly way, taking into account the changing domain specifications. To do so, some empirical studies should be carried out to understand how to come up with sensible PTL values. Another possible next step is to implement and integrate the console with an LBS pilot or with an operational LBS on any of the new devices platforms such as iPhone or iPads and let real users try it and record their experience with our approach.

ACKNOWLEDGMENT

The author would like to thank Dr. Jan van den Berg very much for his help with reviewing parts of this paper. Furthermore, the author aknowledges the support he got from Ordina to publish and present this work in the press.

REFERENCES

1. S. Kalasapur, M. Kumar, and B. Shirazi, "Evaluating Service Oriented Architectures (SOA) in Pervasive Computing," Fourth IEEE International Conference on Pervasive Computing and Communications (PerCom'06), 2006, pp. 275-285.
2. M. Ackerman, T. Darrell, and D.J. Weitzner, "Privacy in context," HCI, 2001, vol. 16, issue 2, pp. 167-179.
3. F. Lardinois, "Two-Thirds of iPhone Users Now Use Location-Based Services at Least Once a Week," http://www.readwriteweb.com/archives/location_services_used_by_two_thirds_of_iphone_users.php, April 22, 2010, Last Access Date: July 15th, 2011.
4. D. Riboni, L. Pareschi, and C. Bettini, "Shadow attacks on users' anonymity in pervasive computing environments," Pervasive and Mobile Computing, 2008, vol. 4, issue 6, pp. 819-835.

 ISBN: 978-1-61208-164-9

5. D. Chaum, "Security without Identification Card Computers to make Big Brother Obsolete," Communications of ACM, 1985, vol. 28, issue 10, pp. 1034-1044.
6. J. Camenisch, and E.V. Herreweghen, "Design and Implementation of Idemix Anonymous Credential System," Proc. the 9th ACM conference on Computer and communications security, New York, USA, 2002, pp. 21-30.
7. A. Lysyanskayal, R.L. Rivest, A. Sahai, and S. Wolf, "Pseudonym Systems," Sixth Annual Workshop on Selected Areas in Cryptography (SAC '99), 1999, Springer-Verlag LNCS, pp. 184-199.
8. A. Coen-Porisini, P. Colombo, and S. Sicari, "Dealing with anonymity in wireless sensor networks," the ACM Symposium on Applied Computing (SAC10), 2010, Sierre, Switzerland, pp. 2216-2223.
9. R.S. Sandhu, and P. Samarati, "Access control: principle and practice," Communications Magazine, IEEE, 1994, vol. 32, issue 9, pp. 40 - 48.
10. B. Schneier, "Cryptographic design vulnerabilities," Computer, 1998, vol.31, issue 9, pp. 29 - 33.
11. R. Agrawal and J. Kiernan, "Watermarking relational databases," the 28th VLDB Conference, 2002, Hong Kong, China, pp. 155-166.
12. B. Carminati, E. Ferrari, and A. Perego, "Enforcing access control in Web-based social networks," ACM Transactions on Information and System Security (TISSEC), 2009, vol. 13, issue 1, article no.6.
13. C. Clifton, M. Kantarcioglu, J. Vaidya, X. Lin, and M. Y. Zhu, "Tools for Privacy Preserving Distributed Data Mining," ACM SIGKDD Explorations, 2002, vol. 4, issue 2, pp. 28 - 34.
14. M. Langheinrich, "Privacy by Design- Principles of Privacy-Aware Ubiquitous Systems," Third International Conference on Ubiquitous Computing (Ubicomp2001), 2001, Springer-Verlag LNCS, pp. 273-291.
15. J.R. Rao and P. Rohatgi, "Can Pseudonymity Really Guarantee Privacy," the 9th USENIX Security Symposium, 2000, Colorado, USA, pp. 85-96.
16. M. Reiter and S. Stubblebine, "Authentication metric analysis and design," ACM Transactions on Information and System Security, 1999, vol. 2, issue 2, pp. 138-158.
17. P. Zimmermann, "PGP User's Guide," 1994, Cambridge, USA, MIT Press, Volume I and II, Distributed with the PGP software.
18. L. Pareschi, D. Riboni, A. Agostini, and C. Bettini, "Composition and Generalization of Context Data for Privacy Preservation," the Sixth Annual IEEE International Conference on Pervasive Computing and Communications (PERCOM '08), 2008, Washington DC, USA, IEEE Computer Society, pp. 429-433.
19. L. Cranor, B. Dobbs, S. Egelman, G. Hogben, J. Humphrey, M. Langheinrich, M. Marchiori, M. Presler-Marshall, J. Reagle, M. Schunter, David A., and R. Wenning, "The Platform for Privacy Preferences 1.1 (P3P1.1)," Specification W3C Working Group Note, http://www.w3.org/TR/P3P11/, 13 Nov. 2006, Last Access Date: July 27th, 2011.
20. L. Cranor, M. Langheinrich, and M. Marchiori, "A P3P Preference Exchange Language 1.0 (APPEL1.0)," W3C Working Draft, http://www.w3.org/TR/P3P-preferences/, 15 April 2002, Last Access Date: July 27th, 2011.
21. L.F. Cranor, P. Guduru, and M. Arjula, "User Interfaces for Privacy Agents," ACM Transactions on Human Computer Interactions, 2006, vol. 13, issue 2, pp. 135-178.
22. R. Agrawal, J. Kiernan, R. Srikant, and Y. Xu, "XPref: a Preference Language for P3P," Computer Networks, 2005, vol. 48, issue 5, pp. 809-827.
23. J. Clark and S. DeRose, "XML Path Language (XPath) Version 1.0," W3C Recommendation, http://www.w3.org/TR/xpath/, 16 November 1999, Last Access Date: July 15th, 2011.

 ISBN: 978-1-61208-164-9

Performance Evaluation of Distributed Application Virtualization Services Using the UMTS Mobility Model

Chung-Ping Hung* and Paul S. Min†
Department of Electrical and Systems Engineering, Washington University in St. Louis
One Brookings Drive, St. Louis, MO 63130, USA
Email: *chung23@wustl.edu †psm@wustl.edu

***Abstract*—In this paper, we first introduce how virtualization technologies can mitigate mobile application software publishing problems due to platform diversity and fragmentation. In our previous work, we proposed a distributed server arrangement and the corresponding hand-off protocol to provide better user experience for application virtualization on mobile devices and evaluated the performance using the modified UMTS outdoor to indoor pedestrian mobility model. We continue our previous work on evaluating performance of the proposed service architecture using the UMTS rural vehicular mobility model with similar modifications. In this paper, combined with our previous work, we complete the establishment of quantitative relations between the performance improvement or impact and the infrastructure related parameters in the typical mobility model.**

***Index Terms*—telecommunication and wireless networks, computer networks, information technology, UMTS mobility model.**

I. Introduction and Related Work

Advanced wireless communication and low-power semiconductor technologies have enabled mobile computing widely available to consumers and dramatically expanded our imagination on personal computing. Mobile computing devices, however, are hardly considered as technical extension of general purpose computers. Mobile computing devices are limited in computational resources, communication-oriented, and highly compact. Therefore, mobile computing devices are "advanced interactive embedded real-time systems" rather than "reduced PCs".

The nature of mobile computing devices inevitably makes the software engineering on them tightly managed by platform vendors. Although centrally managed software publishing helps software developers securing their revenue, this paradigm also enables platform vendors to take much more control on SDKs, design guidelines, and whether a 3rd party software product can be shipped or not. Furthermore, various mobile operating systems are still competing with each other and none is expected to dominate the market in 2 or 3 years. Software developers have to deal with multiple dictating bureaucracies, instead of one, to make their products available to most customers and maximize their revenue. Therefore, developing cross-platform software for mobile computing devices is a costly work using conventional frameworks.

Fortunately, virtualization technologies can work around the difficulties of deploying cross-platform mobile application software. Roughly speaking, application virtualization technologies can be categorized into two major paradigms: one is creating a compatible runtime platform, i.e., virtual machine, on each client's device and publishing well managed code packages running on top of it [1][2], and the other is executing application software on a well managed server while each client's device only handles user inputs and server outputs [3][4][5]. We generally refer to the later paradigm as the *browser-based approach* since web browsers provide a very ideal framework for it.

Despite being technically feasible, deploying a virtual machine running on top of a mobile operating system to execute downloaded common codes circumvents the official software publication platform and generally is considered a violation of the "Non-Compete" policy [7][8] by marketplace operators.[1] Therefore, the browser-based approach becomes the remaining legitimate way to provide application virtualization services on mobile computing devices for general 3rd party developers without special privilege, unless marketplace operators enforce the "Non-Compete" policy against interactive web contents.

The conventional web-based application virtualization relies on a centralized server to provide the service through established Internet infrastructure. Although this configuration can be built with low cost, the long response latency could significantly reduce the user experience since every input must travel through a series of routers and bridges to the colocation center and the corresponding update has to traverse backward through the nodes. Each node along the route induces processing delay, queuing delay, and transmission delay, and each link comprises the route induces propagation delay. Generally speaking, network delay is highly related to the geographical distance between two end points given similar network infrastructure technologies.

[1] VMware's Mobile Virtualization Platform (MVP) [6], which implements this paradigm, is not available in Android Marketplace. To install MVP on an Android phone requires sideloading, and only Android platform leaves this loophole to install apps outside the marketplace, which is at the mercy of Google and wireless service providers. In fact, some wireless providers do block sideloading on some Android phones. Furthermore, among the major mobile device players, only Android is supported by MVP. Therefore, even VMware starts their own app store for MVP, it doesn't help cross-platform software deployment anyway.

 ISBN: 978-1-61208-164-9

In our previous work [11], we have proposed an alternative configuration to address this issue that geographically partitions the service area into multiple smaller service areas and deploys a smaller server for each one to provide the service locally. The proposed configurations should significantly reduce propagation delay in most case due to the shorter average communication distance. The proposed configuration, however, has to handle hand-off cases, i.e., mobile stations in use moving from one service area to another. Therefore, we also proposed a hand-off protocol offering seamless user experience.

Handling hand-offs induces longer response latency and thus the overall performance depends on the hand-off behavior model. Therefore, we used an empirical and simplified approach to evaluate the performance as a result of infrastructure arrangement and application software's properties in our previous work [11]. We also used one of the UMTS mobility models to empirically establish the correlations between the performance and the size of each local service area and the capabilities of the network infrastructure in [12]. In this paper, we complete the performance simulation of the proposed architecture with the UMTS rural vehicular mobility model. In the UMTS rural vehicular mobility model, base stations (BSs) are sparsely but optimally placed, mobile stations (MSs) move faster and more freely, and the hand-off behavior among base stations is different as well. Consequently, the simulation program in the UMTS rural vehicular model is significantly different from the one presented in our previous work though sharing the same concept. The simulation algorithm and results based on the UMTS rural vehicular mobility model are represented in this paper.

There are several papers proposed to optimize service migration though for different applications. Bienkowski et al. proposed competitive analysis for service migration in optimizing the server allocation in VNets in [9]. Arora et al. proposed some strategies for flexible server allocation in [10] following the previous work [9]. Although these works were not specifically for mobile application virtualization, they provide a precious insight on the performance evaluation for dynamic service allocation considering both user experience and operational cost. However, the analytical approach used in these works is topological and does not focus on the user mobility and interaction models. In our approach, we simulate the user mobility geographically based on the UMTS mobility models which provides an alternative performance preview in resource migration. Furthermore, the authors of [9] and [10] allow services being temporarily interrupted during migrations, which is not feasible for application virtualization services. In the proposed configuration, application services are available to users with reduced performance during hand-offs.

The rest of this paper is organized as follows. In Section II, we describe the proposed configuration aim to improve the user experience of application virtualization services. In Section III, we propose a VM-level hand-off protocol to handle the additional information exchange brought by the proposed server configuration. We specify our experiment design, settings, and cost metrics in Section IV. Then the simulation results given different parameter adjustments are presented in Section V. Finally, we conclude our work and outline some future work we expect to do in Section VI.

II. Proposed Configuration

Running application software on a remote server while creating an illusion that the client has full control of the software in hand is conceptually similar to the usage model of time-sharing mainframe computers in the 1960s [13]. Although the communication bandwidth between terminals and mainframe servers at that time was very low by modern standards, it did not affect the user experience thanks to the text-only display and short traverse distance. However, in recent application virtualization technologies which follow the same concept, such as Virtual Desktop Infrastructure (VDI) proposed by VMware [14], much more versatile and bloated content must be exchanged over much longer distances between clients and servers than their predecessors.

An infrastructure ready to offer mobile users application virtualization services includes base stations covering the whole service area, a core network connecting base stations and servers together, and a server hosting the services. A command sent by a mobile station has to travel over the wireless channel to the BS, go through the backhaul network to the server, and then make some changes on the server. Should any update corresponding to the command be sent to the MS, the information has to travel all the way backward. In order to reduce the network delay generated by long transmission distances among the backhaul network, we geographically deploy multiple servers among a wide area to serve their nearby MSs in the proposed configuration, instead of setting up one centralized server serving all MSs.

In the proposed configuration, each server connects to several nearby BSs to form a *local service group* (LSG). The area covered by the BSs of the same LSG is defined as the *local service area* (LSA). Every BS should belong to one LSG in order to provide the service all over the wireless network's coverage area. When a user demands a virtual application program, the server of the LSG, based on VDI [14] paradigm, starts a virtual machine (VM) dedicated to the user and launches the application software on top of it. The MS only handles inputs and outputs that interact with the VM at the server.

As long as the MS stays in the same LSA, the user can enjoy using application software with low response latency. If the MS moves from the original LSA to a nearby one, a hand-off at the VM level, which transfers the runtime environment to the server of the next LSG, is triggered. Therefore, a protocol to deal with the hand-off condition is required.

III. Hand-off Protocol

The purpose of the proposed hand-off protocol is to transfer minimum information required to recreate the runtime environment on a remote server, i.e., the *snapshot*, without

 ISBN: 978-1-61208-164-9

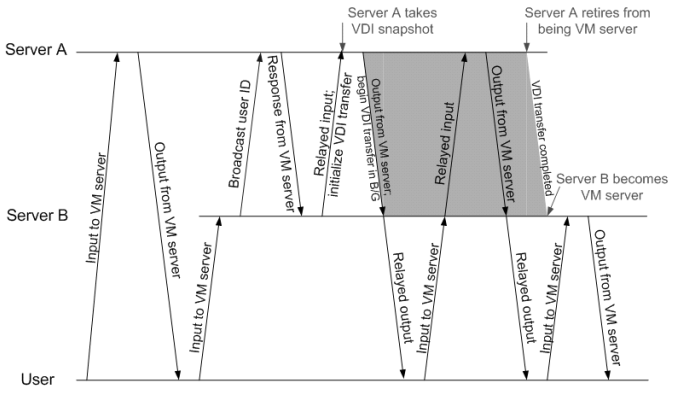


Fig. 1. Protocol timeline for an MS moving from Server A to Server B.

interrupting the service. To provide a seamless user experience during transmitting the snapshot, the next server has to record all inputs from the MS, relay all inputs to the previous server, and relay all output from the previous server to the MS, until the runtime environment resumes locally. The proposed hand-off protocol is described as below:

1) When an MS moves from Server A's to Server B's LSA and sends an input command, Server B notices a newcomer within its LSA.
2) Server B broadcasts the newcomer's identification to all geographically nearby servers.
3) Server A, which hosts the MS's runtime environment, i.e., its VM server, responds Server B's inquiry. Now Server B knows the newcomer's VM server is Server A.
4) Server B records and relays the user's input commands to Server A, signals Server A to transfer the runtime environment, and relays display updates from Server A to the newcomer.
5) Once Server A is signaled to transfer the runtime environment, it takes a snapshot.
6) Besides continually responding to the input commands relayed from Server B as the MS is still in its LSA, Server A also sends the snapshot to Server B in the background.
7) Once Server B receives the complete snapshot and recreates the runtime environment from the snapshot and base data, it internally feeds the input queue, which was recorded during the transition period, to the runtime environment. Therefore, the runtime environment state on Server B is synchronous with that on Server A after the snapshot was transferred.
8) Server A completely stops serving the MS, the MS's VM server is now Server B instead.

The timeline of the proposed hand-off protocol is illustrated in Fig. 1.

If the MS turned around and reentered Server A's LSA before the hand-off was completed, Server A can preempt the snapshot transmission and resume serving the MS as if the hand-off never happened. Since Server B relays all inputs to Server A while the MS is absent from Server A's LSA, aborting the hand-off procedure would not generate any noticeable glitch. This hand-off abortion mechanism can prevent unnecessary data transmission from moving VM servers back and forth if an MS were moving around the edge of an LSA.

On the other hand, if the MS moved to Server C's LSA before the hand-off was completed, Server C initializes another hand-off procedure with Server B. In addition to the snapshot, Server B has to transfer the input record before Server C joins the hand-off chain. We allow pipelining transmission to reduce hand-off periods and shorten subsequent hand-off chains in this scenario.

IV. Performance Evaluation

The proposed service architecture is designed to reduce interaction latency and thus provide more responsive user experience on remote controlled application virtualization services. However, due to the involvement of the hand-off protocol, the performance of the proposed service architecture depends highly on the probability of hand-offs, the geographical deployment of the BSs, and the configuration and capability of the backhaul network. The former two factors can be modeled by the test environments of existing communication systems, such as the well-published UMTS benchmarks [15]. On the other hand, the configuration and capability of the backhaul network can only be assumed based on reasonable technical and cost considerations.

A. UMTS Vehicular Mobility Model

The UMTS document [15] provides three different test environments, which are the Indoor Office, the Outdoor to Indoor and Pedestrian, and the Vehicular ones, for technology selection and evaluation. We simulated the hand-off behavior and evaluated performance of different infrastructure settings using the Outdoor to Indoor and Pedestrian mobility model in our previous work [12]. In this paper, we complete the performance evaluation by simulating the application virtualization services on the Vehicular test environment specified in the UMTS document.

As shown in Fig. 2, the UMTS rural vehicular test environment is a plain with no physical obstacle. Each MS's speed is fixed at 120 km/h. Each MS's moving direction is allowed to change up to 45° left or right every 20 meters with 20% chance. All MSs are initially uniformly distributed on the plain.

The BSs in the UMTS rural vehicular test environment are located at the dark grey dots in Fig. 2. Each BS has three directional antennae to serve tri-sectored cells. Each cell is assumed to be a hexagon and seamlessly tiles with each other. Each cell's radius R is either 2000 meters (for services up to 144kbit/s) or 500 meters (for services above 144kbit/s). Therefore, the minimum distance between two BSs can be 6 km or 1.5 km, respectively.

The original UMTS mobility model generates discontinuities on the boundaries of the test area. We consequently add some special traffic rules, known as *portals*, to eliminate the boundary discontinuities and allow the interaction among

 ISBN: 978-1-61208-164-9

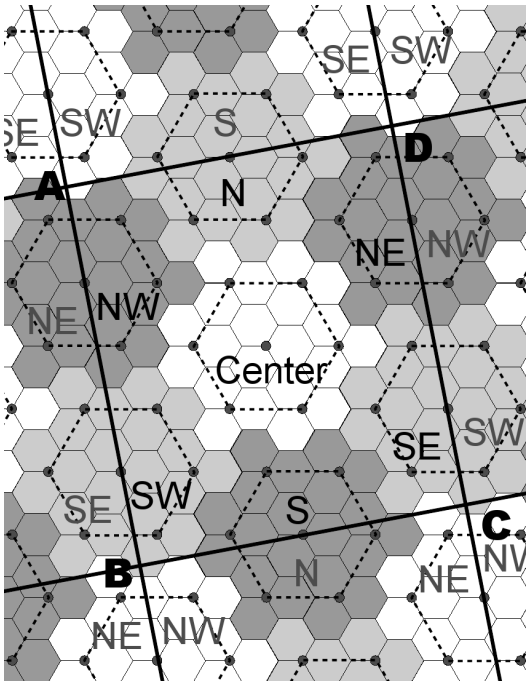


Fig. 2. The UMTS rural vehicular test environment with LSA arrangement.

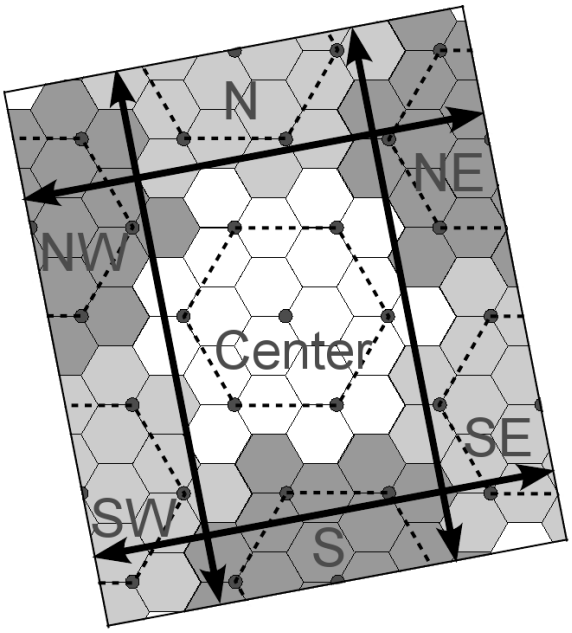


Fig. 3. Möbius County map with teleporting directions.

LSAs to be simulated and observed for an indefinite period of time. The characteristics of the portals will be detailed in the next section.

B. Möbius County

What interests us is the geographical relation between the service facilities and the MSs' moving space. As the method we conducted in our previous work [12], the first step is to define a sample area which can represent all the geographical characteristics of service infrastructure we need. We first group the BSs in Fig. 2 to form approximately hexagon-shaped LSAs which are optimized in both coverage and average transmission distance by deploying servers at the centers. As the urban counterpart, i.e., Möbius City, in our previous work [12], the sample area should include one complete LSA in the center and six neighboring halves. Given R and N, the number of the BS intervals per LSA's edge, if we align the origin to the server of an LSG, we define the Parallelogram ABCD surrounded by four straight lines, which are:

1) $\sqrt{3}x - 3(2N+1)y = -6\sqrt{3}R(3N^2+3N+1)$ on the north,
2) $\sqrt{3}x - 3(2N+1)y = 6\sqrt{3}R(3N^2+3N+1)$ on the south,
3) $\sqrt{3}(2N+1)x + y = -3\sqrt{3}R(3N^2+3N+1)$ on the west,
4) and $\sqrt{3}(2N+1)x + y = 3\sqrt{3}R(3N^2+3N+1)$ on the east.

as the sample area of our best interest. We can, therefore, crop out Parallelogram ABCD in Fig. 2 as our test area, where we call *Möbius County* as shown in Fig. 3, to represent every identical piece comprises the indefinite large test area.

Like Möbius City [12], assigning four logical LSGs in Möbius County is sufficient to figure out when, where, and how frequently an MS moves from one LSA to another. However, to apply the hand-off aborting mechanism, which was disabled in [12], we need to distinguish whether an MS is coming back to the LSA it just left or entering the LSA on the opposite side of the one it just crossed. Therefore, we have to assign an additional unique identification for each LSG.

The portals around Möbius County are also similar to those around Möbius City. Whenever an MS is about escaping from Möbius County, the portal teleports it to a proper location at the opposite side so that it reenters Möbius County. Therefore Möbius County can emulate a limitless test area. Since there is no street structure to align in Möbius County, the rules of the portals are much more simple and straightforward than of Möbius City:

1) For MSs about crossing the north boundary, teleport them to $\left(3R, -3\sqrt{3}R(2N+1)\right)$ from their current locations.
2) For MSs about crossing the south boundary, teleport them to $\left(-3R, 3\sqrt{3}R(2N+1)\right)$ from their current locations.
3) For MSs about crossing the west boundary, teleport them to $\left(-\frac{9R(2N+1)}{2}, -\frac{3\sqrt{3}R}{2}\right)$ from their current locations.
4) For MSs about crossing the east boundary, teleport them to $\left(\frac{9R(2N+1)}{2}, \frac{3\sqrt{3}R}{2}\right)$ from their current locations.

The teleport directions are shown in Fig. 3 as well.

The purpose of the portals is to eliminate all discontinuities except the MS's coordinates when it is moving out of the boundary: it keeps the same direction and speed, it associates with the same logical LSG, and preserves the geographical parameters relative to the service group's facilities. Thus,

 ISBN: 978-1-61208-164-9

everything interests us is equivalent as the MS moving into an adjacent parallelogram area in a limitless test area.

C. Configuration of Backhaul Network

We assume a mesh-styled backhaul network as we did in [12]. Therefore, each BS only has direct links to its six neighboring BSs. In the mesh-styled backhaul network, network latency between a BS and the server depends on the number of nodes along the shortest path, the total length of the path, and the relay latency per node. The former two factors are related to the coordinates of the BS and the server, while the last one is varied to simulate different nodal transmission capabilities.

D. Performance Metric

We define the response time as the average time interval between when a user sends an input and gets an expected output update. The proposed server configuration is meant to improve the response time by reducing traverse delay along the communication route from each MS to the server which is hosting the service. Factors other than the traverse delay, such as computational capabilities provided by servers, would affect the user experience and the quality of our service. Most of them, however, either affect different configurations equally, or can be overcome with reasonable cost. The traverse delay is defined as:

$$T_{tv} = 2 \cdot \left\{ \frac{L_r}{V_r} + \frac{L_l}{V_l} + N_{rt} \cdot T_{rt} + N_{rl} \cdot T_{rl} \right\} \qquad (1)$$

where L_r is the distance of radio transmission, which is the distance between the MS and the BS covering it, V_r is the propagation speed of radio, which is the speed of light, L_l is the total length of wireline transmission in the mesh network, V_l is the propagation speed in wireline, which is approximately two thirds of V_r, N_{rt} is the number of nodes along the transmission path in the mesh network, T_{rt} is the average waiting time per node in the mesh network, which includes nodal processing delay, queuing delay, and transmission delay, N_{rl} is the number of servers which are receiving the snapshot and relaying data to/from the VM server, and T_{rl} is the processing and relay time per server in the hand-off chain. Obviously, all parameters, except V_r and V_l, depend on an MS's geographical location and hand-off state.

E. Hand-off Duration

Whenever a VM-level hand-off occurs, we set up an anticipated hand-off end time by adding hand-off duration to the current time. The hand-off duration is calculated by the following equation:

$$T_{ho} = T_x + \frac{L_s}{V_l} + N_s \cdot T_{rt} \qquad (2)$$

where T_x is the total time to deliver every bit of a snapshot to media, which is the summation of queuing delay, processing delay, and transmission delay of the snapshot, which is proportional to the size of the snapshot, L_s is the total transmission distance between the current and the next VM servers, and N_s is the number of nodes between two neighboring servers, which always equals to $2N + 1$ in this case.

F. Update Time Points and Cost Charging

Although we only calculate costs at position update points, updates actually take place when a hand-off is completed in addition to when an MS reaches an update position. At each update time point, T_{tv} and transaction counts are updated concurrently.

Whenever a position update comes at T_{now}, all hand-off end times registered in queue prior to T_{now} are update time points as well. The corresponding costs have to be calculated in retrospect according to the algorithm described below:

1) Define T_n as the n_{th} earliest hand-off end time in queue, L_{sn} as the total transmission distance between servers corresponding to the n_{th} earliest hand-off in queue, L_r, L_l, N_{rt}, and N_{rl} are the current cost parameters calculated by the MS's current position and hand-off status, and T_{last} as the previous update time.
2) If $T_{now} > T_0$, insert an update time point at T_0, calculate the transaction counts by the Poisson process given user input rate λ and time duration $(T_0 - T_{last})$, set $T_{last} = T_0$, subtract N_{rl} by one, subtract N_{rt} by $\{2N + 1\}$, subtract L_l by L_{s0}, update T_{tv} according to the new parameters, and remove T_0 and corresponding L_{s0} from the queues.
3) Redo step 2 until $T_{now} < T_0$ or the queue is emptied.
4) Calculate the transaction counts by the Poisson process given λ and time duration $(T_{now} - T_{last})$, update T_{tv} according to the new parameters, and set new $T_{last} = T_{now}$.

As specified in the UMTS rural vehicular mobility model, we update the MSs' positions every 20 meters. Since a hand-off may occur at the same time, we have to handle the extra cost brought by it as well. When a new hand-off occurs with a position update at current time T_{now} while the previous update time is T_{last}, and every hand-off end time earlier than T_{now} is already treated with the above algorithm, we use the following algorithm to update the cost parameters:

1) Register the new hand-off end time and the corresponding L_s in the queue.
2) Increment N_{rl} by one.
3) N_{rt} is recalculated by the MS's current position and added by $\{N_{rl} \cdot (2N + 1)\}$.
4) Let L_l equals to the summation of all L_s's in queue.
5) T_{tv} is then updated accordingly.
6) The transaction counts are calculated by the Poisson process given λ and time duration $(T_{now} - T_{last})$, and then set new $T_{last} = T_{now}$ for the next update.

Since the variation of the geographical parameters is negligible along the 20 meters (or less) long path, every transaction in an update interval is charged with identical T_{tv} to reduce the computational complexity. Note that T_{tv} updated at time T is applied to the transactions which occur *after* T, while the transaction counts calculated at T are placed in the time interval ended at T.

 ISBN: 978-1-61208-164-9

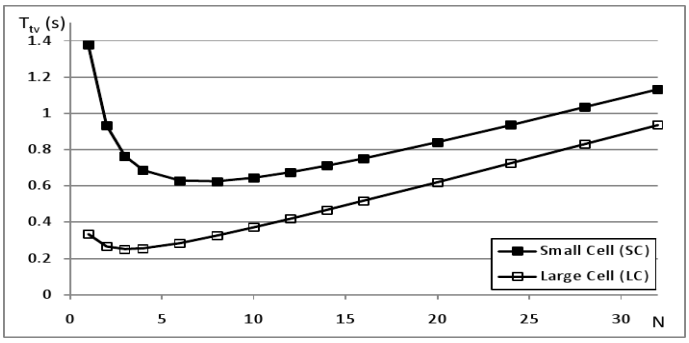


Fig. 4. Simulation results of different N of both cell configurations given $T_{rl} = 0.5s$ $T_{rt} = 20ms$, and $T_x = 600s$, $\lambda = 1.0$.

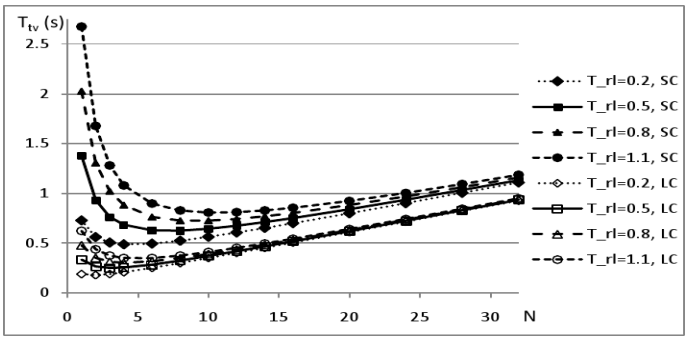


Fig. 5. Simulated T_{tv}'s of both cell configurations given $T_{rl} = 0.2s$, $0.5s$, $0.8s$, $1.1s$ and $T_{rt} = 20ms$, $T_x = 600s$, $\lambda = 1.0$.

G. Traverse Time Accounting

The average T_{tv} per transaction is calculated at the end of 100,000 independent simulations, each lasting 86400 seconds. The simulation results of variable N, T_{rt}, T_{rl}, T_x, and λ for both $R = 2000m$ or $500m$, are presented in the following section.

V. Simulation Results

We first simulate how the size of LSAs affects T_{tv} given *nominal* parameters, which are $T_{rt} = 20ms$, $T_{rl} = 500ms$, $T_x = 600s$, and $\lambda = 1.0$. The simulation results of both R settings are shown in Fig. 4.

As we can see in Fig. 4, both T_{tv}'s bear a strong resemblance in shape to the counterpart in [12] despite the significantly different mobility models. T_{tv}'s are high in small LSA configurations due to the higher hand-off occurrence rate. As N increases, T_{tv}'s first descend, level for several N's, and then linearly ascend. The descending for low N's is due to the reduction of hand-off occurrences. The smooth ascending for higher N's is caused by the higher average number of the nodes along the backhaul route and the longer average transmission distance while the hand-off occurrence rate is too low to matter. The flat bottom in between is the result of the two effects competing with each other.

Note although we compare two cell configurations, $R = 2000m$ and $R = 500m$, in the same figure, each LSA of the former one is in fact 4 times larger than of the latter one. Therefore, each MS encounters much fewer hand-offs in the large cell configuration than in the small cell one. We can also observe slightly steeper ascending for higher N's in the large cell configuration than in the small cell one due to the higher propagation delay brought by the longer wireline and wireless transmission distances.

We can conclude that in this case, setting $N = 4$ for the large cell configuration, and $N = 8$ for the small cell one, are optimal in reducing average T_{tv} and keeping the total number of the servers low, which also means lower deployment and maintenance cost.

Since the above quantitative conclusion is only applicable in this set of parameters, we adjust each parameter in the nominal set and compare the results to see how it affects T_{tv}'s as functions of N in the following subsections.

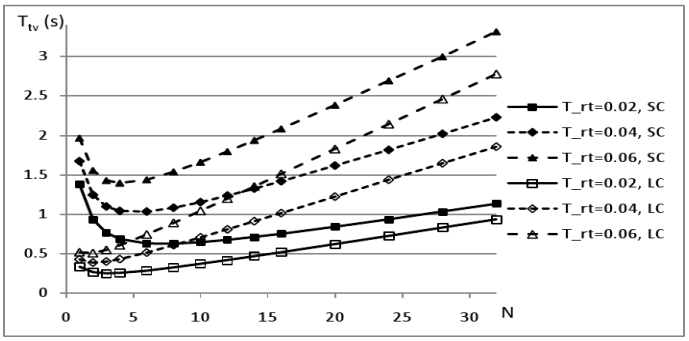


Fig. 6. Simulated T_{tv}'s of both cell configurations given $T_{rt} = 20ms$, $40ms$, $60ms$ and $T_{rl} = 500ms$, $T_x = 600s$, $\lambda = 1.0$.

A. Effect of T_{rl}

T_{rl} only participates in hand-off conditions. In this simulation, we set T_{rl} to $200ms$, $800ms$, and $1100ms$, and see how it affects both T_{tv}'s. Both simulated T_{tv}'s in large and small cell configurations as functions of N and T_{rl} given $T_{rt} = 20ms$, $T_x = 600s$, $\lambda = 1.0$ are shown in Fig. 5.

As we can see in Fig. 5, higher T_{rl} significantly increases T_{tv}'s in small LSA configurations due to the higher occurrence rate of hand-offs. As N increases, T_{tv}'s in each cell configuration given different T_{rl}'s have a tendency to converge together since the hand-off occurrence rate is dramatically reduced and thus renders the effect of T_{rl} insignificant. In the large cell configuration, T_{tv}'s converge more significantly and earlier due to the extremely low hand-off occurrence rate.

B. Effect of T_{rt}

Higher T_{rt} amplifies the influence of transmission distance. The simulated T_{tv}'s in both cell configurations as functions of N and T_{rt} given $T_{rl} = 0.5s$, $T_x = 600s$, $\lambda = 1.0$ are shown in Fig. 6.

Fig. 6 shows the comparison of T_{tv}'s of both cell configurations as functions of N given $T_{rt} = 20ms$, $40ms$, and $60ms$. Besides the resemblance in shape to the counterpart in [12], we can also notice that T_{rt} is a more decisive factor for the large cell configuration's performance due to the low hand-off

 ISBN: 978-1-61208-164-9

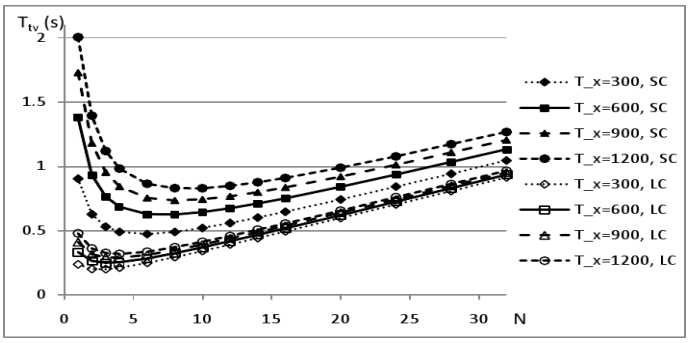


Fig. 7. Simulated T_{tv} of both cell configurations given $T_x = 300s$, $600s$, $900s$, $1200s$ and $T_{rt} = 20ms$, $T_{rl} = 0.5s$, $\lambda = 1.0$.

occurrence rate and the long average communication distance in each LSA. Even $N = 1$ can be preferable if T_{rt} is greater than 60ms in the large cell configuration.

C. Effect of T_x

T_x only affects the cost brought by hand-offs. A higher T_x may mean a larger snapshot file, a longer hand-off initialization time, or a longer queuing delay. How T_x affects T_{tv} is represented in Fig. 7.

Similar to the counterpart in [12], T_{tv}'s of each cell configuration as functions of N given different T_x's are virtually parallel for high N to each other and show very little tendency to converge as N increases. However, slightly higher optimal N brought by higher T_x in both configurations is still observable.

D. Effect of λ

Although we have shown that user input rate λ was not a relevant parameter in [12], we still simulate T_{tv}'s as functions of N given different user input rates λ in Möbius County. We again confirm that the property doesn't change in the UMTS rural vehicular mobility model.

However, we should keep in mind that the user experience depends more on the interactivity of the application software than on the absolute response latency.

VI. Conclusion And Future Work

In this paper, we complete the performance evaluation of the proposed application virtualization configuration and the corresponding hand-off protocol by applying the UMTS rural vehicular mobility model. In addition to the model Möbius City which we proposed in our previous work [12], we propose Möbius County using a similar concept to enable MSs to move in the test environment based on the UMTS rural vehicular mobility model indefinitely without dealing with any boundary condition. We simulate the network delay as a result of MSs' movements and the occurrences of VM-level hand-offs in Möbius County given variable sizes of LSAs, server relay latencies, nodal relay costs, and snapshot transmission durations.

Möbius County, combined with Möbius City, can provide a performance preview of network infrastructures aimed at improving mobile application virtualization services in large scale unknown environments. We can also design a benchmark framework specific for distributed application virtualization services based on the proposed simulation environments.

In this paper, we employ deterministic infrastructure delay parameters and a simple usage model to evaluate the performance. We will introduce more sophisticated usage and infrastructure delay model to facilitate more precise mobile application virtualization service simulations. Furthermore, besides the benefit of lowering the average response latency, the distributed application virtualization service configuration and the hand-off protocol can also be applied to load balancing and fault tolerance for better resource management and service robustness. We will investigate these potential applications in the future as well.

References

[1] Sunwook Kim et al., "On-demand software streaming system for embedded system", *WiCOM 2006 International Conference on Wireless Communications, Networking and Mobile Computing*, 22-24 Sept. 2006, pp. 1-4.

[2] EMA Report: "AppStream: Transforming On-premise Software for SaaS Delivery - without reengineering"

[3] Joeng Kim; Baratto, R.A.; Nieh, J., "An application streaming service for mobile handheld devices", *SCC'06 IEEE International Conference on Services Computing*, Sept. 2006, pp. 323-326.

[4] VMware Inc., "VMware ThinApp: Agentless Application Virtualization Overview".

[5] Ana Fernandez Vilas et al., "Providing web services over DVB-H: mobile web services", *IEEE Transactions on Consumer Electronics*, Vol. 53, No, 2, May 2007, pp. 644-652.

[6] VMware Inc., "VMware MVP (Mobile Virtualization Platform)", http://www.vmware.com/products/mobile/overview.html, Retreived 7 Aug. 2011.

[7] Apple Inc., "App Store Review Guidelines for iOS Apps", 2.7 and 2.8, http://developer.apple.com/appstore/guidelines.html, Retrieved 9 Sep. 2010.

[8] Google Inc., "Android Market Developer Distribution Agreement", 4.5, http://www.android.com/us/developer-distribution-agreement.html, Retrieved 22 Feb. 2011.

[9] Marcin Bienkowski et al., "Competitive Analysis for Service Migration in VNets", in *Proc. 2nd ACM SIGCOMM Workshop on Virtualized Infrastructure Systems and Architectures*, 2010, pp. 17-24.

[10] Dushyant Arora et al., "On the benefit of virtualization: strategies for flexible server allocation", *Hot-ICE'11 Proceedings of the 11th USENIX conference on Hot topics in management of internet, cloud, and enterprise networks and services*, 2011.

[11] Chung-Ping Hung and Paul S. Min, "Infrastructure arrangement for application virtualization services", *I2TS 2010 The 9th International Information and Telecommunication Technologies Symposium*, 13-15 Dec. 2010, Vol. 1, pp. 78-85.

[12] Chung-Ping Hung and Paul S. Min, "Service area optimization for application virtualization using UMTS mobility model", *ICOMP 2011 International Conference on Internet Computing*, 18-21 July 2011, pp. 128-134.

[13] L. P. Deutch and B. W. Lampson, "SDS 930 Time-sharing System Preliminary Reference Manual", Doc. 30.10.10, Project Genie, Univ. Cal. at Berkeley, April 1965.

[14] VMware Inc., "Virtual Desktop Infrastructure".

[15] ETSI. "Universal Mobile Telecommunications System (UMTS); selection procedures for the choice of radio transmission technologies of the UMTS (UMTS 30.03, version 3.2.0)". Technical report, European Telecommincation Standards Institute, Apr. 1998.

 ISBN: 978-1-61208-164-9

A Framework for Data Roving in Ubiquitous Computing Infrastructure

Richard E. Gunstone and David Newell
Computing and Informatics, School of Design, Engineering and Computing
Bournemouth University, Poole, Dorset, United Kingdom
rgunstone@bournemouth.ac.uk, dnewell@bournemouth.ac.uk

***Abstract*—In future Ubiquitous Computing Infrastructure, a growing data demand from users particularly of mobile devices will create further pressures on infrastructure to deliver information at the right time. We introduce the concept of "data roving", where smart infrastructure predicts the likely information needs and replicates data to parts of infrastructure in the vicinity of the user concerned. We propose this concept to reduce the impact of congestion, reduce pressure on finite infrastructure resources, and improve data access performance for the mobile user. Unknown quantities at this stage include the efficiency of anticipating information needs of users, and whether this offers a performance advantage over more straightforward demand-based caching. The main contribution of this paper is a description of how a network infrastructure could be used in predicting information needs and to support replication of data to enhance performance.**

***Index Terms*—mobility; middleware for mobile environments; ubiquitous computing; mobile content delivery networks**

I. INTRODUCTION

The concept of *ubiquitous computing* is acknowledged as beginning with the work of Mark Weiser at Xerox [1]. Weiser can be considered something of a visionary, describing a future paradigm of computing that eschewed then-accepted approaches in favour of a highly distributed, interactive and pervasive world. Similarly, we view ubiquitous computing as a move toward an environment where technology diffuses into the background and where software systems are used that adapt to user needs autonomously. A prevalent example of ubiquitous computing to date has been the emergence of *converged devices* such as smartphones and tablet-class computers, such devices being indicative of a widespread trend toward more ubiquitous architectures.

Contemporary converged devices and mobile devices (MDs) in particular have placed unprecedented demand on communications infrastructure. Current-generation MDs are pushing the data bearing capacities of even recent innovations to 3G cellular networks, necessitating widespread deployment of data service upgrades. Growth areas such as mobile video, and increasing device capabilities and consumer need, are likely to expand these demands further (and are spurring on the development of new technologies such as Cognitive Radio [2]). Increasing demands on finite radio spectrum for wireless networking indicate the introduction of smart techniques into the management of data across networks is of benefit.

We introduce *data roving* (DR) as the 'anticipation of data consumption needs in advance, by either a client or infrastructure, and the reproduction of data locally to the client to satisfy data consumption needs in sufficient time and to required quality'. We consider this to be an evolution of *data staging* (DS), which has its origins in business intelligence, data warehousing, and dynamic infrastructure provisioning. DS in these contexts refers to the provision of a duplicate or replica of an information source suitable for processing, that is then moved back into an storage library after processing is completed. (Similarly, where the replica is not moved back into a storage library, DS is equivalent to *caching*, for instance the service offered by Content Delivery Network (CDN) nodes or HTTP caches.) DR has relevance for mobile users particularly in the development of middleware for mobile environments and mobile content delivery networks.

The concept of DR, much like DS, offers benefits in terms of the reduced latency of access for the user. Rather than the mobile user performing I/O operations across a potentially large network of varying performance levels, I/O can be performed with a nearby replica of the information required. This is achieved by providing information locally to the client, which makes use of a computer system local to the MD with sufficient capacity to vend data to mobile users (in many respects similar to a CDN). Employing DR may also reduce the exposure of the mobile user to congestion effects on busy network links between the user and the source, particularly where the hop count from the client to the DR replica is shorter. Performance may also be higher if network links to the replica have equal or greater reliability than the non-DR equivalent.

The DR concept becomes important when evaluated against a backdrop of increasingly high data consumption demand by consumers, and the exchange of increasingly large file sizes especially for multimedia. The use of MDs with limited local storage also creates further pressure for data to be frequently in transit over a fixed or wireless network link. Therefore, the ability to intelligently move data around is becoming increasingly important. While concepts such as Quality of Service (QoS) provide alternative options, these are typically focused on real-time traffic management according to the type of service required.

In section 2, the paper the DR architecture is described, including a high-level process and a discussion of how information can be selected for DR. This is followed by conclusions in the final section, including consideration of related topics.

 ISBN: 978-1-61208-164-9

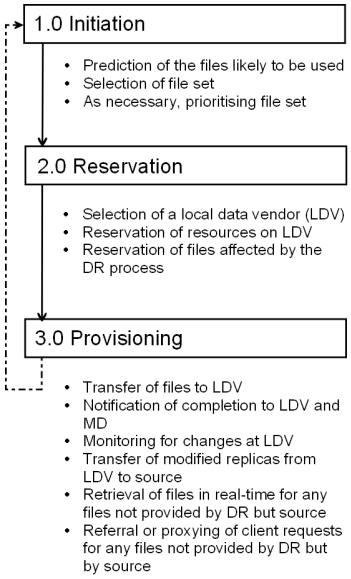


Figure 1. The three steps involved in *Data Roving*: initiating the process, reserving resources within the UCI, and provisioning of the replica(s) to a nearby LDV that also includes committing changes back to the original information store.

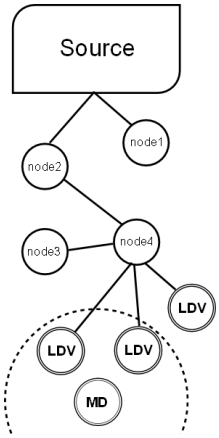


Figure 2. The infrastructural aspects to DR: the most optimal node is designated as the LDV, to provide data to MD without recourse to the source unless the predicted data needed does not match actual needs for MD. More than one LDV may be selected depending on actual and anticipated mobility modelling of MD and other relevant performance factors.

II. Data Roving Architecture

Figure 1 shows the three steps in the DR process, namely *initiation*, *reservation* and *provisioning*, and Figure 2 shows a sketch of an example Ubiquitous Computing Infrastructure (UCI) supporting DR. We define at this point several terms: the Mobile Device (MD) is a converged device that has data provision managed via DR; a Local Data Vendor (LDV) is a node in the UCI that is presently or anticipated to be in proximity (physically, topologically, or both) to the MD, of which there may be more than one at any time; and the "source" is the principal and authoritative manager of the data for the user (and MD).

The initiation phase (1) is concerned with the management of the user data in preparation for distribution to LDVs. The reservation phase (2) is concerned with the selection of LDVs and the reservation of resources both at the source and the LDV. The provisioning phase (3) is concerned with the transfer and synchronisation of files between the source and the LDV. The provisioning phase may also refer or proxy requests from the client for any files that are not available at the LDV but are provided by the source.

In the initiation phase (1), a prediction should made as to what data (files) the mobile user is likely to require. Key elements of this phase include the selection of a set of files that the mobile user is likely to require, based upon a prediction function. As necessary, this stage should include prioritisation of the files into a sorted list based upon a prioritisation rule. Potential algorithms for selecting files for DR include: (1) Selection of all files (in the case of small data stores); (2) The most recently accessed files by the user; (3) The most recently created files by the user; (4) Inspection of recently modified file content and derivation of queries to select further content; (5) A more elaborate file selection based on logical query/policy (e.g., for selecting media by attributes such as genre, artist, year, etc.) developed by the user.

At this point in the DR process, knowledge of a candidate LDV may not be available, and thus a prioritisation of the files after prediction may be suitable for later use. A variety of approaches can be employed for predicting the files a user is likely to need, and in a sense this can be parametrised based on policy settings (including for instance user preferences).

In the reservation phase (2), the process moves onto the selection of the LDV(s). This can identify more than one node in the UCI that can be used to host data nearby to the MD. The selection of candidate LDVs should integrate as many relevant factors as possible; probable factors include the present geographical location of the MD relative to the candidate nodes, the proximity of the MD to the candidate nodes based on number of network hops, quality of link types, applicable policies at the LDV, available LDV resources, etc. This may also involve a historical analysis of the characteristics of interest, of both the LDV and the MD, in an effort to predict a set of LDVs that could be of use to the MD in the future (and not just the present). This could be achieved by integrating the outputs of a mobility modelling process on the UCI and MD. The LDV chosen should be within reach or likely reach of a wireless (or fixed) links from the mobile user, and therefore be in a position to vend data to the mobile user as necessary.

The reservation phase should also reserve resources (storage space, etc.) on the LDV and update the source to maintain awareness of the present state of DR for files identified in phase I. Moreover, appropriate resources should be reserved on the LDV to cater for contention and availability.

After the reservation phase is completed, the process moves into the provisioning phase (3). This part of DR is concerned with the provisioning of data from the source to the MD, via the LDV. Files identified in phase I should in this phase be transferred to candidate LDVs. Taking into consideration the available storage capacities of, and link qualities to, the LDV, a suitable quantity of files are then transferred. Once files are transferred, both the LDV, source and MD should receive a

 ISBN: 978-1-61208-164-9

notification of recent changes.

The provisioning phase is also responsible for monitoring for changes to the files transferred to the LDV, and these changes should be propagated back to the source at the earliest opportunity. Finally, the provisioning phase should provide a function for referral of clients to the source-managed files if these are not available on a local LDV and/or provide a proxy service to retrieve these files on behalf of an MD. The mobile device should be notified by the UCI that a nearby LDV is available for data access needs. The provisioning phase also encompasses the access from the MD to the data provided by the LDV.

Research related to phase 3 includes the work of Flynn et al. [3], which explored the use of data staging (not to be confused with the more generic use of data staging used in the introduction) through the use of the Coda filesystem. They address issues that are beyond the scope of this work at this stage, though the issues are of relevance, and in particular investigate the issue of using untrusted nodes in a network as a vehicle for DS, an aspect that requires further investigation in this framework.

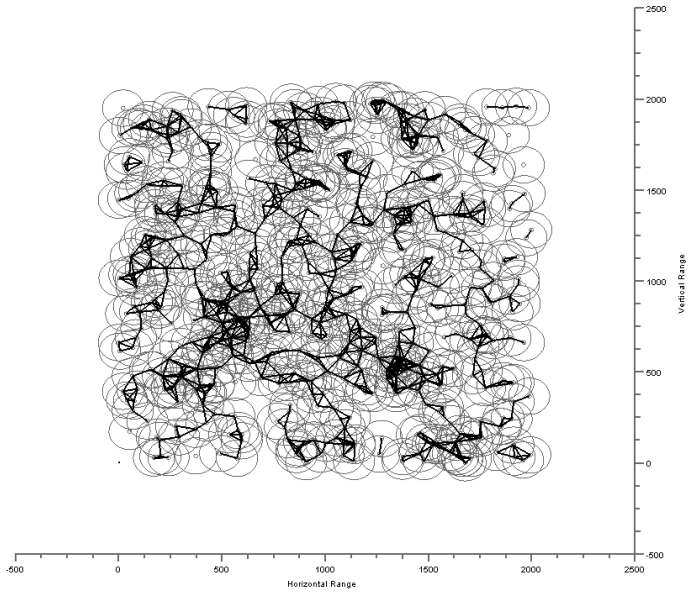


Figure 3. Initial ad-hoc wireless simulation over a 2km/square area.

III. Conclusions

This work-in-progress paper has provided a description of a framework for data roving for ubiquitous computing infrastructure. Three phases in the DR process have been described. Against a backdrop of increasing data demand requirements, relatively low quantities of storage on MDs compared to desktop counterparts, and an increasingly congested communications infrastructure, the concept of DR provides a means through which efficiencies may be feasible.

A number of experimental steps are necessary to validate the DR concept. Firstly, an implementation of DR is necessary to evaluate feasibility characteristics (an early version is shown in Fig. 3). Whereas the work in [3] uses a filesystem, we initially envisage a simpler model to transfer individual files based upon a unique identifier at this stage principally focused on the predictive element of DR. We do not consider heterogeneous traffic at this stage, however this is a future potential area for investigation. Figure 3 shows initial experimental work using a homogenous network type (though we note network selection is a more complex problem more generally, as illustrated in [4]).

A performance evaluation is also required, to assess the performance benefit of particularly the predictive element of DR. That is, the process of anticipating the files needed by a user and transferring this data through the infrastructure to a nearby node. An important parameter in whether the algorithm should consider short- or long-term usage data. Anticipating LDV nodes may lead to several redundant copies of data across the network. While this may be optimised through the use of file synchronisation protocols, there remains a potential inefficiency that can only be justified through the optimal prediction of files the user is likely to require.

Further, and as the introduction noted, employing DR may reduce the exposure of the mobile user to congestion effects on busy network links between the user and the source. In cases where access to the source is lost, DR may offer considerable benefits, and this should be characterised.

An important consideration, briefly mentioned in passing, is a means through which privacy of information transited through the UCI can be ensured. This is not addressed as part of this framework at this stage.

The primary difference between DR and other approaches such as caching, is that the process can be initiated in advance of the anticipated time of need. Particularly in the *reservation* step, an opportunity to make use of low utilisation of UCI, for instance during night time or early morning, when demand network traffic is lower, may create an opportunity to reduce any inefficiencies through the use of DR.

References

[1] M. Weiser, “Some Computer Science Issues in Ubiquitous Computing,” *Commun. ACM*, vol. 36, pp. 75–84, July 1993.

[2] S. Hamouda and B. Hamdaoui, “Dynamic spectrum access in heterogeneous networks: HSDPA and WiMAX,” in *Proceedings of the 2009 International Conference on Wireless Communications and Mobile Computing: Connecting the World Wirelessly*, IWCMC ’09, (New York, NY, USA), pp. 1253–1257, ACM, 2009.

[3] J. Flinn, S. Sinnamohideen, N. Tolia, and M. Satyanaryanan, “Data Staging on Untrusted Surrogates,” in *Proceedings of the 2nd USENIX Conference on File and Storage Technologies*, (Berkeley, CA, USA), pp. 15–28, USENIX Association, 2003.

[4] D. E. Charilas and A. D. Panagopoulos, “Network Selection Problem Multiaccess Radio Network Environments,” *IEEE Ve*, pp. 40–49, December 2010. Digital Object Identifier 10.1109/MVT.2010.939107.

 ISBN: 978-1-61208-164-9

Formalisms for Use Cases in Ubiquitous Computing

Richard E. Gunstone
Computing and Informatics, School of Design, Engineering and Computing
Bournemouth University, Poole, Dorset, United Kingdom
rgunstone@bournemouth.ac.uk

***Abstract*—Use cases have achieved widespread adoption in software requirements capture and representation in software system analysis and design. Ubiquitous Computing, a paradigm attributed to the late Mark Weiser, offers new challenges in the representation of user requirements. Such systems are likely to require a high degree of user-centricity if they are to meet sometimes demanding requirements of the modern computer user. This paper presents a review of relevant use case methodologies, principally those that augment or replace the Unified Modelling Language metamodel for use cases.**

***Index Terms*—use cases; software; ubiquitous computing**

I. Introduction

Use cases have achieved widespread adoption in software requirements capture and representation in contemporary software engineering and they are frequently integrated into business processes. The requirements engineering process is widely recognised as being crucial in the process of building a software system, by building a specification through iterative processes of elicitation, specification, and validation; and crucially this should ideally also integrate multiple viewpoints to foster objectivity [1].

A widely recognised shift in the usage of computing technology, focusing more on computing as a communications tool rather than a collection of discrete computational processing systems, leads toward a close relationship between user needs and software systems. This represents a change in the focus of attention, away from the computer system itself toward support for the activities of the user. *Ubiquitous computing*, and the software systems used to achieve it, shares this shift in emphasis toward user- and not system-centricity.

The concept of ubiquitous computing is acknowledged as beginning with the work of Mark Weiser at Xerox [2]. Weiser can be considered something of a visionary, describing a future paradigm of computing that eschewed then-accepted approaches in favour of a highly distributed, interactive and pervasive world. Similarly, we view ubiquitous computing as a move toward an environment where technology diffuses into the background and where software systems are used that adapt to user needs autonomously. We consider the current understanding of ubiquitous computing to be an evolution of Weiser's original design, reflecting some aspects but not all. Contemporary ubiquitous computing examples are wide and varied, but they tend to require relatively advanced functionality in networks (including Cloud computing), context-awareness, and interaction. A prevalent instance of ubiquitous computing to date has been the emergence of *converged devices* such as smartphones and tablet-class computing, and these devices are indicative of a widespread trend toward more ubiquitous computing architectures.

The use case approach to software engineering (as proposed by Jacobson) would appear to fit many of the requirements of ubiquitous computing. It encapsulates user requirements of a system in an easily-understood formalism, and caters for multiple viewpoints. However, limitations have been identified in several studies since its introduction.

The conventional use case representation is based on the concept of scenarios or collections of use cases that represent flows of events [1]. There are several prominent characteristics. The first is that a use case should consist of a *description* [3], which is typically unstructured or semi-structured text. The use case should also contain a *sequence of actions* (or transactions) performed by the system [3, 4, 5]. In addition, for the use case concept to be valid such actions should *lead to an observable result* [3, 4, 5] thereby excluding incomplete or multiple sequences [5]. Finally, the observable result should be of *value* to an actor [3, 4, 5]. Use cases can be represented using a graphical notation.

Use cases serve as projections of future system usage and as projected visions of interactions with a designed system [6]. They are considered by some authors to be scalable to large and complex systems, as they can be improved and expanded incrementally with little or no loss of prior information. Use cases can also be used in such a way as to facilitate requirements traceability throughout design and implementation [1]. Use cases typically make use of natural language, and because of this they are recognised as being a good medium through which requirements can be elicited and recorded, and because of their representation they also avoid the problems associated with purely narrative approaches, while at the same time permitting partial specifications [7]. Taking a user's viewpoint is of significant value when validating the 'adequacy of requirements' [7], and use cases can help facilitate such a process. Use cases are also considered useful for introducing abstraction into the requirements capture and design processes, to allow the system to be understood from structural, behavioural and interactive perspectives [8]. As a consequence of these characteristics, they can support a more effective elicitation process and consequently enable an agreement on the views of the users involved [9].

While use cases offer many opportunities, several of the key benefits offered through their use can also be considered drawbacks. As has been noted, they derive much of their flexibility from their use of natural language, however in

 ISBN: 978-1-61208-164-9

lacking formal lack formal syntax and semantics [1, 9] they may permit too much scope, introducing the potential for ambiguity or misinterpretation, and in in the case of plain narrative text this can lead to serious quality problems [7]. They also have limited support for structuring and managing large use-case models [9], which can have consequences for all parts of the development process. Similarly, while there has been recent efforts to represent checks on the quality of use case descriptions [10, 11], the use of text to describe the use case does not necessarily guarantee that the complete process is specified. Use case may also promote a highly localised perspective that can obscure business logic [5]. Finally, in their application, use cases permit the definition of system behaviour from the perspective of many different stakeholders, and, while such flexibility can be considered a benefit in some respects, it is also thought to lead to conflicting functional requirements [12] and contradiction.

In trying to find suitable augmentation or replacement of the methodology for the reasons described in Section I, a number of concepts including some applicable to ubiquitous computing, can be considered desirable. Lee et al. propose several that we propose are very relevant for ubiquitous computing and slightly adapt [1]:

- *clarity* (combining Lee et al.'s *comprehensibility* and *unambiguity*)–a replacement or augmented formalism should continue the theme of ease-of-understanding that Jacobson's original method provided, and also provide a method that is clear and unambiguous to all stakeholders and analysts.
- *scalability*–any change should scale well to the use case elicitation process, providing representational schemes to ensure abstraction is permitted.
- *partiality*–a proposed formalism should accommodate often incomplete information (and to not require a complete representation in order to function).
- *goal-based*–it is desirable for a use case methodology to include some kind of formalisation of user goals as part of the representation to accommodate the goals of the user.

These criteria provide a useful way of analysing the range of formalisms that have emerged since the original use case methdology was proposed. Ubiquitous computing has a strong association with privacy concerns, one aspect not included in the above summary. Incorporating privacy (including Non Functional Requirements - NFRs) from the initial requirements capture process is another important consideration, and we discuss this in more detail in Section III.

We propose the move towards more widespread use of ubiquitous computing technologies interactions between user and system will become focused on the activities of the users in their environment rather than with the system that may be used at a particular time. Therefore, we consider a representational scheme that can accommodate goals as being an important requirement as a means to partly, or fully, address this trend.

The remainder of this paper is structured as follows: In Section II we review a number of extensions that have emerged regarding use cases, with a particular focus on the aspects most relevant to ubiquitous computing. The paper is finished in Section III, where conclusions are made.

II. Review

A number of extensions have emerged that could be regarded as incremental refinements. Lee and Xue [9] propose a goal-based approach (GDUC) to specifying use cases, one that permits the representation of NFRs and consideration of interactions between requirements. NFR representation is a frequent desirable property in the engineering of complex systems, however the use case approach is not geared to representing NFRs easily. Moreover, ineffectively dealing with NFRs is thought to have led to a number of failures in software development (see [13] for a review and an overarching process to elicit NFRs). In GDUC for a given actor-specific and functional goal, it is classified with respect to their competence (complete or partial satisfaction is needed), view (actor- or system-specific), content (functional or non-functional), yielding a faceted break-down into *«extend»* extension use cases and extension goals. The representation used in GDUC allows for NFRs to be represented and also permits associations between indirectly associated goals.

Another process is that by Tan et al. [14] where Data Flow Diagrams (DFDs) are utilised in an augmented form to transform parts of an analysis model such as processes and data flows into an Object Oriented (OO) design and implementation in terms of classes, class attributes, and so on. Their aim is to harness DFDs for use-case realisation. In this approach, DFDs are enhanced to cater for candidate class operations being reflected in processes in the diagram, candidate class attributes, arguments and return values being represented, and control flow being fully specified without recourse to additional specification. The authors propose that by incorporating their augmented DFDs into the requirements analysis stage, by representing use-cases, a comprehensive method can than be used to translate use-cases into OO design, and in turn to implementation.

Glinz has proposed an approach that combines a structured textual representation and a statechart-based structure [7]. In this method there is a clear distinction between events produced by the actor and the responses of the system, and there are simple structuring constructs (such as *if*, *go to step*, etc.) to add further detail. Activity flows between scenarios can be achieved by arranging the scenarios as a directed graph, where edges have explanatory text describing the conditions required for that flow of activity. Glinz's statechart methodology appears to offer greater clarity compared to other approaches, and we return to this in our observations section.

Similar to Glinz's approach, Nebut et al. [15] propose a formalisation of the use case corpus by capitalising on pre- and post-conditions, to make contracts executable by using requirement-level logical expressions. This is a logical evolution of the use case approach and gives rise to a sequential structuring. In the case of Nebut et al.'s work, this is done with the goal of generating tests from a formalisation (in particular the detection of faults in embedded software). Logical expressions used in the conditions of use cases are

 ISBN: 978-1-61208-164-9

constructed from Boolean logic operators: *conjunction*, *disjunction*, and *negation*, in addition to quantifiers $\forall$ *(forall)* and $\exists$ *(exists*), and *implication* in post-conditions. The authors use the formalism to generate all possible orderings of use cases to form a 'transition system', and once constructed this allows a variety of structural and logical tests to be performed, exercising paths in the transition system. Tests are implemented as sequence diagrams, representing nominal or exceptional scenarios associated with use cases. The use of a formalised representation has particular benefits in terms of analysis (c.f. Lee et al.) however the authors note the difficulty it may impose on the requirements analyst, who is forced to be rigorous and to clearly specify conditions.

An interesting continuation of augmentation to the UML® metamodel is the work of Dias et al. [16] that introduces *Use Case Fragments* (UCFs) into the development process of use cases. They identify that successful use cases should include several elements, including the basic flow and sequence of steps, alternate flows, information exchange details, and also business rules associated with the interactions encapsulated in the model. Addressing the perceived requirements of both students and novice professionals, they consider the use of a catalogue of recurring use case fragments that can be composited quickly to address the majority of use cases needed in typical information systems. Such fragments are noted as having a similar organisation to software patterns, with each addressing the abstract concept such as *select one element from a set of existent elements*. A UCF template contains the fragment name, sub-goals, purpose, basic flow, alternative flows, input and output details, and rule details. In order to have an applied UCF reflect the situation where it is being used, there are customisation points in the UCF where business terms are substituted. Some of the advantages noted by Dias et al. include the UCF acting as a facilitator, supporting novice requirements professionals in developing use cases, improving writing speed and supporting the specification process.

Sutcliffe [6] introduces the scenario-based requirements engineering approach, of relevance to ubiquitous computing. In common with the original use case approach, this process involves use case elicitation from users and gives formatting guidelines. The final two stages in this process involve scenario generation (from a use case specification) and scenario validation. The process makes use of a reusable library of requirements and associated application classes, influencing factors, exception types, requirements specification(s), and validation frames to support method stages. The use case is modelled as a collection of actions with rules that govern connectivity between actions, and these actions lead to the attainment of a goal. Use cases are organised into a hierarchy to allow for refinement of higher-level use cases. Several action link types are available, including strict sequencing, part sequencing, and inclusion, offering a range of ways links can be established. Validation functionality is possible, to check conformance of the use case models using a schema, and if suitable abstraction is present the use case model can be linked to related systems allowing identification of abstract application classes. This work provides a series of functions that would aid the development of ubiquitous computing systems, particularly where there are common application functions across use cases. An example could be route finding and geo-social networking, which both require geo-location services through the Global Positioning System (GPS), Assisted GPS, Wi-Fi® and IEEE® 802.11 Positioning Systems, etc.

A justification for many of the research proposals that advocate a replacement formalism in the UML metamodel is that the use case method does not lend itself well to formal analysis [1]. In comparison, a formalised model would permit an analysis of the dependencies and inconsistencies (or flaws) in the model. Several research studies have proposed replacing the UML metamodel for use cases to achieve well-defined representations suitable for analysis and resolving imprecision, which we mention in passing.

The Petri net formalism is a technique that has been used to model software systems since the 1970s [17]. Devised as a means of modelling discrete-event systems, Petri nets are well-suited modelling software concepts and in particular for modelling concurrency, with no inherent requirement for synchronicity, and offer (through their asynchronous features) an abstraction above the flow of time, to present events in terms of their partial ordering [17]. They have been applied in a number of ways to enhance the use case approach. Xu & He [18] apply the concept of the Place Transition variant of the Petri net ((PT, PrT, P/T net) to the generation of test requirements, in the context of aspect-oriented use cases. In contrast, Lee et al. propose Constraints-Based Modular Petri Nets (CMPNs) (see [1] for mathematical definitions), which are suggested as improving on the drawbacks of P/T nets and coloured Petri nets. CMPNs are argued as addressing several shortcomings with other approaches, such as being cumbersome where incremental modification is required, to limited analysis techniques on interaction and dependency, and in the case of Finite State Automatons (FSAs) high state space complexity [1].

Replacement formalisms are geared toward a more radical change to the representational schemed used for use cases, and while offering benefits in terms of machine analysis it does present potential drawbacks. Such drawbacks can include a loss of clarity, and potentially more work on the part of the use case analyst.

III. OBSERVATIONS

Use cases have achieved widespread use in software engineering problems, and more widely as a generic term for user needs from products and services. In this paper we have identified the principal benefits of the original use case method, shortcomings and new approaches. For ubiquitous computing, the adoption of use cases provides an opportunity to ensure any systems developed are more closely aligned to user requirements.

This paper has identified several evolutions to the original use case representation, and these can be broadly classified as either *augmentation* or *replacement* (though we note this is not a clear separation and in reality some degree of overlap exists). Replacement formalisms tend to use a more formal representation to permit machine analysis. Regards this latter

 ISBN: 978-1-61208-164-9

point and our earlier observations in Section I, we note Glinz, that while formal representations are useful for analysis and can achieve high levels of precision, this comes at the expense of readability and effort to write the scenarios [7]. This shortcoming in comprehensibility of some approaches detracts from the comprehensibility of the original use case method (a characteristic identified in Section I as worthy of preservation).

Taking these concepts further requires the development of an extension to the UML metamodel that facilitates the desirable attributes needed for ubiquitous computing, in particular representing user goals. It is also necessary to construct and refine a suitable requirements example for ubiquitous computing that illustrates these requirements, and use this to validate the modifications to the UML metamodel. One example is the the class of situation awareness applications for ubiquitous computing (c.f. [19]) that provides a basis for an illustrative set of use cases from a user perspective.

One aspect not examined in depth, but noted in the introduction, is that of privacy. It is desirable particularly for ubiquitous computing systems to consider how privacy requirements on the parts of users can be represented and implemented early-on in the development process.

Ubiquitous computing takes privacy considerations for contemporary computing much further, because such systems are intrinsically based upon the collection and processing of information about their users, the environment, their property, actions, and so on. This information is collected all or most of the time, senses information types that are not readily accessible in contemporary computing paradigms (such as video camera feeds of rooms, contents of fridges via Radio Frequency Identification, etc.), and is extensively processed to yield new useful information that can aid the activities of its users. On a practical level moving toward building such complex systems inevitably requires specific technical developments to the system design to ensure privacy can be protected or enhanced. This necessitates the implementation of technical measures–Privacy Enhancing Technology (PET), discussed in a complementary paper [20].

REFERENCES

[1] W. J. Lee, S. D. Cha, and Y. R. Kwon, "Integration and Analysis of Use Cases Using Modular Petri Nets in Requirements Engineering," *IEEE Transactions on Software Engineering*, vol. 24, pp. 1115–1130, December 1998.

[2] M. Weiser, "Some Computer Science Issues in Ubiquitous Computing," *Commun. ACM*, vol. 36, pp. 75–84, July 1993.

[3] D. Leffingwell and D. Widrig, *Managing Software Requirements: A Use Case Approach*. Addison-Wesley Professional, 2 ed., May 2003. Print ISBN-10: 0-321-12247-X; Print ISBN-13: 978-0-321-12247-6.

[4] I. Jacobson, M. Griss, and P. Jonsson, *Software Reuse: Architecture, Process and Organisation for Business Success*. Reading, MA: Addison-Wesley/ACM Press, 1997.

[5] A. J. H. Simons, "Use Cases Considered Harmful," tech. rep., University of Sheffield, 1999.

[6] A. G. Sutcliffe, N. A. M. Maiden, S. Minocha, and D. Manuel, "Supporting Scenario-Based Requirements Engineering," *IEEE Transactions on Software Engineering*, vol. 24, pp. 1072–1088, December 1998.

[7] M. Glinz, "Improving the Quality of Requirements with Scenarios," in *Proceedings of the Second World Congress for Software Quality (2WCSQ)*, (Yokohama), pp. 55–60, September 2000.

[8] Object Management Group, "Introduction to OMG's Unified Modelling Language," July 2005. http://www.omg.org/gettingstarted/what_is_uml.htm, Last accessed June 2011.

[9] J. Lee and N.-L. Xue, "Analyzing User Requirements by Use Cases: A Goal-Driven Approach," *IEEE Software*, pp. 92–101, July/August 1999.

[10] K. Phalp, J. Vincent, and K. Cox, "Assessing the Quality of Use Case Descriptions," *Software Quality Journal*, vol. 15, pp. 69–97, 2007.

[11] K. Phalp, A. Adlem, S. Jeary, J. Vincent, and J. Kanyaru, "The Role of Comprehension In Requirements and Implications for Use Case Descriptions," *Software Quality Journal*, 2010.

[12] J. H. Hausmann and R. Heckel, "Detection of Conflicting Functional Requirements in a Use Case-Driven Approach - A static analysis technique based on graph transformation," in *ICSE 2002*, pp. 105–115, ACM Press, 2002.

[13] L. M. Cysneiros and J. C. S. do Prado Leite, "Nonfunctional Requirements: From Elicitation to Conceptual Models," *IEEE Transactions on Software Engineering*, vol. 30, pp. 328–350, May 2004.

[14] H. B. K. Tan, Y. Yang, and L. Bian, "Systematic Transformation of Functional Analysis Model into OO Design and Implementation," *IEEE Transactions on Software Engineering*, vol. 32, pp. 111–135, February 2006.

[15] C. Nebut, F. Fleurey, Y. L. Traon, and J.-M. Jezequel, "Automatic Test Generation: A Use Case Driven Approach," *IEEE Transactions on Software Engineering*, vol. 32, pp. 140–155, March 2006.

[16] F. G. Dias, E. A. Schmitz, M. L. M. Campos, A. L. Correa, and A. J. Alencar, "Elaboration of Use Case Specifications: an Approach Based on Use Case Fragments," in *Proceedings of SAC08*, (Fortaleza, Ceara, Brazil), pp. 614–618, ACM, March 2008.

[17] J. L. Peterson, "Petri Nets," *Computing Surveys*, vol. 9, pp. 223–252, September 1977.

[18] D. Xu and X. He, "Generation of Test Requirements from Aspectual Use Cases," in *Proceedings of WTAOP07 Workshop*, (Vancouver, British Columbia, Canada), pp. 17–22, ACM, March 2007.

[19] S. Minamimoto, S. Fujii, H. Yamaguchi, and T. Higashino, "Map estimation using GPS-equipped mobile wireless nodes," *Pervasive and Mobile Computing*, vol. 6, pp. 623–641, 2010.

[20] R. E. Gunstone, "Integrating Privacy During Requirements Capture for Ubiquitous Computing," in *Proceedings of the First International Conference on Social Eco-Informatics*, (Barcelona), 2011.

 ISBN: 978-1-61208-164-9

Development of a Context-Aware Information System for Baseball Service

Young-Tae Sohn, Jae Kwan Kim, Myon-Woong Park
Center for Bionics
Korea Institute of Science and Technology (KIST)
Seoul, Korea
{ytsohn, kimjk, myon}@kist.re.kr

Jae Kwon Lim, Soo-Hong Lee
Dept. of Mechanical Engineering
Yonsei University
Seoul, Korea
{ljk1225, shleee}@yonsei.ac.kr

***Abstract*—A context-awareness is one of the important issues for developing an intelligent information service system in order to provide the information most useful for the users. In this paper, a context-aware information service system for baseball game is described. To recognize the context of baseball play, a contextual knowledge model is suggested. An 'observation point' concept is also introduced to provide baseball information effectively and proactively. A proto-type context-aware information service system has been implemented on smart phone, and evaluated in the aspects of usefulness and appropriateness. The system was appreciated in understanding over the progress and immersion of the match.**

Keywords-context-awareness; contextual knowledge model; information service; observation point

I. INTRODUCTION

Rapid and remarkable progresses in computing environment such as World Wide Web and powerful yet affordable server systems enable the application programs to provide more intelligent and proactive information services. Software agent is usually the kernel of those information services, and intelligent agents for information services have been developed to satisfy user's requirement and to provide practical services in many different domains such as finance, traffic, e-health [6][7][8]. Among those agents, the agent for context-aware information service is supposed to recognize not only the current situation of the information content and user, but also the changes of their situation, in order to supply the information most useful for the user at the contextual aspect. The recognition of the situation is based on the overall assessment about all the related factors such as context of content, user location, surrounding objects, environmental conditions, and their semantic relations. As a consequence, for the recognition of the situation, contextual knowledge models and reasoning techniques based on ontology concept have drawn many attentions, and reported as highly effective [5][10][11]. For the representation of situation, the changes over time should be considered as one of the important factors since the agent needs to adapt to the dynamically changing situation and corrects its behavior [3][4]. These changes are even more significant in the case of the context-aware information service for baseball game.

Baseball game continues couple of hours, and the situation consisted of many factors like pitcher, batter, runner, inning, out count, ball count, etc. keeps changing. At every change, game audience might be interested in specific information through which the progress of the game is predicted. For the simplicity, in the rest of this article we refer to these kind of information as 'observation points (OP)'. There can be many observation points relevant to the specific situation of the baseball game. In addition, each audience might be interested in different observation points according to his or her knowledge of baseball game. For instance, the record on the stolen base of a runner might be more interesting than the hitting average of a batter if the runner is on the first base with out-count one at late inning while both team scored nil. In the broadcasting of baseball game, even though some observation points are usually provided by commentator, these information are only the commentator's observation points, so that likely to fail to satisfy every TV audience. Moreover, the commentator's observation points are very limited as the commentator should prepare the observation points before the specific situation take place. Currently, several mobile information services for baseball game have been developed and commercialized [9][12]. These services, however, focused on providing the current status of a game and live video clips.

In this paper, we investigate the use of a software agent which can systematically provide the observation points at the specific situation of baseball play in order to address the above-mentioned problem. The agent was regarded as the kernel of artifact being capable of playing the role of commentator and recommending proactively the observation points suitable for the context of game as shown in Figure 1.

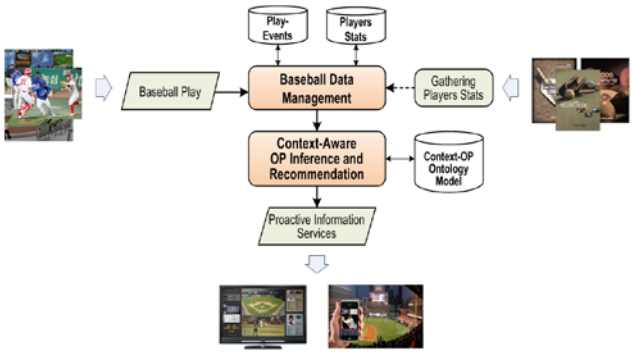


Figure 1. Conceptual diagram of the intended information service.

 ISBN: 978-1-61208-164-9

For the inference of the observation points at the specific situation of baseball play, a contextual knowledge model ('Context-OP Ontology Model' in Figure 1) representing the relations between contextual factors and observation points of baseball play is suggested. All the possible observation points at the specific situation of baseball play are collected from the interviews with baseball experts. Also, the records of players are gathered manually, and statistically managed to be linked with the observation points in the contextual knowledge model. Each situation of baseball play is entered by game recorder as a play event defined with situation factors. The agent has been implemented as a context-aware information service on a smart phone, which becomes the powerful platform for mobile internet services and the most useful personal assistant for watching baseball game.

The rest of this paper is organized as follows. Section 2 describes the context-awareness and functional requirement for information services. Section 3 illustrates the functional structure and its role of the proposed context–aware information service for baseball game. Section 4 explains the contextual knowledge model used for recognition of situation and inferring observation points. Section 5 discusses the service platform and a proto-type mobile application of the developed information service system. Section 6 sums up the work and concludes the paper.

II. CONTEXT-AWARE INFORMATION SERVICE

As the agent needs to recognize current situation and recommend information appropriate for the context, it requires number of functionalities such as context-awareness, inference of matching items, information retrieval, and priority indexing (see Figure 2) [1][2]. For the context-awareness, the formalization of the contextual factors defining the situation of the target content is necessary. The recognized contextual factors can be represented in a formalized context model by applying ontology technology. The function to recognize and assess current context by analyzing the model is identified as context recognition. Clear definition on the contextual elements is necessary for context-awareness. Moreover, grouping of the elements and converting them into knowledge are also necessary. Once current context is recognized, the agent refers to the knowledge model and infers the information appropriate for the context. The knowledge model defines context element and information element for generating the relations between two elements in knowledge map. Then, the agent retrieves the specific information relevant to the context from the database and supplies it to the user.

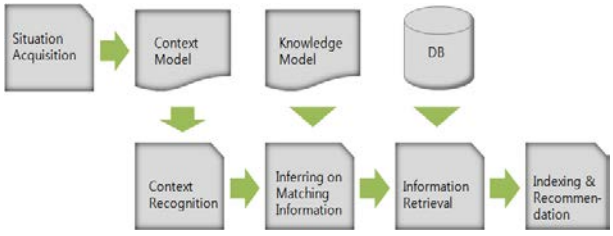


Figure 2. Functions required for a context-aware information service.

III. CONTEXT-AWARE INFORMATION SERVICE FOR WATCHING BASEBALL GAME

There are various records in baseball game as often referred as the game of data. They are commonly provided by caster or commentator. However, inevitably the service is uni-directional and limited in quality and quantity in many cases. Introduction of the 'observation point', and proactively providing the observation point and relevant statistics in suitable timing by the system would improve the service and add the zest to watching the game. The observation point means the group of the player's record and interesting information relevant to the current context of baseball play, which attracts audience's interest. The system structure of the context-aware information service applied to the baseball game watching is depicted in Figure 3. The role of each module on the structure diagram is introduced, and the flow from context recognition to providing observation points is explained in this section.

Provisioning of observation points starts when a play-event occurs, that means the change of the situation in baseball game. The event is recorded by game recorder through the interface as shown on Figure 4, and analyzed to match to the context model, which is then stored in the 'Play-Event DB'. The server process (i.e. agent) regularly checks the change of content in the 'Play-Event DB'. In case of any changes detected, the agent recognizes the current context through the 'Context Recognition Module', and extracts observation points relevant to the context. Then, the observation points are served through the user interface. The extraction of the observation points is carried out in the 'Observation Point Inference Module' by inferring the 'Context-Observation Point Ontology Model' representing the relations between context and observation points. The agent extracts all the instances of observation points matched to the context. Then, detail information of each extracted observation points are retrieved from the 'Observation Point DB'. As shown in Figure 5, the instance of the observation point has the name of the related database table, name of column, searching constraints, etc. as the 'Related Data' properties. The details of each observation points are stored in the database through 'Player Stats Analysis Module' when the players' records are imported.

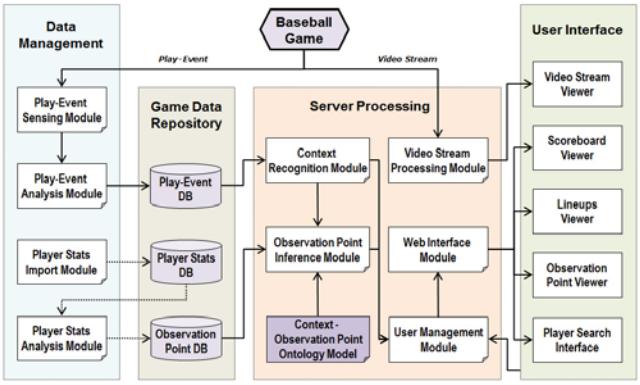


Figure 3. System structure of the context-aware information service for watching baseball game.

Figure 4. User interface for recording play event.

Also, a weight value is given to each observation point instance in order for comparing the importance of the observation points. The inferred observation points are sorted in the order of weight values given to the observation points. Finally, the derived observation points are changed into XML files, and provided to user in the 'Web Interface Module'. Through this integrated process, the agent proactively provides users with situation relevant observation points of baseball game, and users are able to get interesting information.

IV. CONTEXTUAL KNOWLEDGE MODEL FOR INFERENCE OF OBSERVATION POINTS

The structure of the contextual knowledge model for the representation of the relations between context and observation points is shown in Figure 5. The model consists of 2 classes for representing the change of context, and 2 classes for representing the relation between observation points. Also, the relation between context and observation point is represented in order to infer the observation point relevant to the current context. This relation is one-to-many relation as the one observation point might be relevant to many contexts. Moreover, an additional class, 'Additional OP', is introduced to represent the related multiple observation points. The context class is defined with contextual factors of baseball play, and the observation point class has a weight value and properties used for retrieval of detail information. In the context class, contextual factors are grouped to be recognized according to the data type property. Inning, runner, count, score difference, etc. are groups in high level. Inning has inning number and top/bottom as lower elements. This ontology model has been modeled in the way of contexts that can be specifically represented with these grouping elements by using Protégé as shown in Figure 6. The context with a series of events can be represented since the context class contains the relation with the previous context. The change in context along with the progress of events can be expressed as the sequence, which is controlled with 'previous context number'. The agent can identify the contextual change over time according to the representation defined in the model. Variety of the observation points might have been derived since the change of context over time can be expressed and the agent is able to recognize the contextual changes.

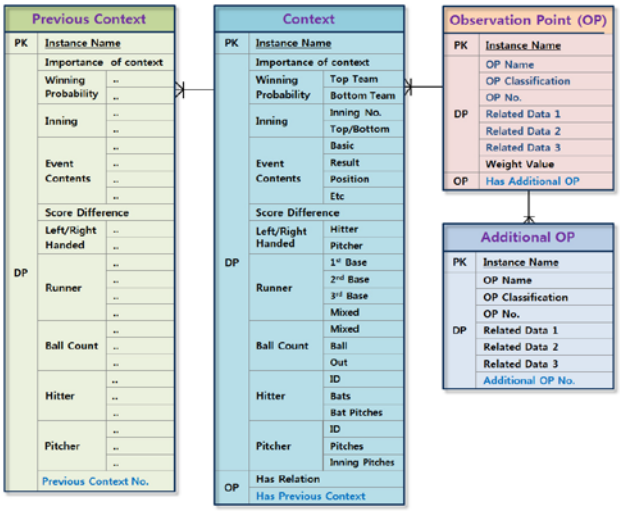


Figure 5. Structure of the Contextual Knowledge Model.

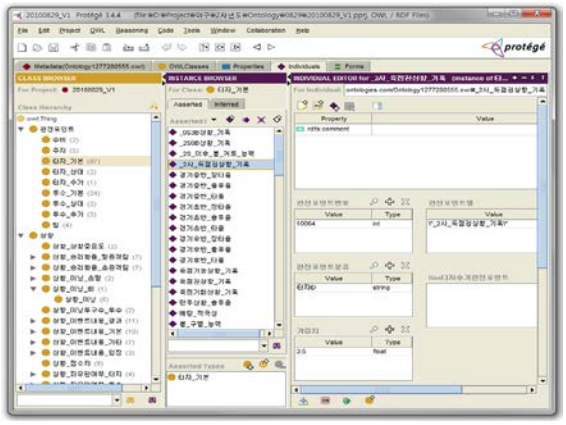


Figure 6. Ontology Modeling of Knowledge Model using Protégé.

V. IMPLEMENTATION

A proto-type information service system for providing observation points and the relevant information during the baseball play has been realized in order to assess the usability of the context based information agent. The system has been implemented in the server/client structure based on internet as shown on Figure 7. The near real-time play event on a game in progress and the video stream of the game are supplied to the server at the ball park. The observation points related to current situation are derived and indexed by the agent according to the level of relevance. The observation points and related statistics are stored along with the current video stream, and these are supplied to the client on request.

As an example, iPhone has been used as the client device for mobile service. A client App has been developed, which can supply video stream of the baseball game, observation points, game progress information, player information, etc. through the user interface as shown on Figure 8. The App regularly requests the server for the observation points along with the timing of the video streaming of a game. The game situation and the player information are supplied only on demand from user.

 ISBN: 978-1-61208-164-9

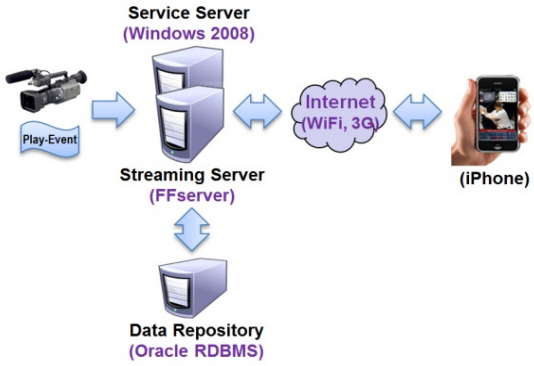


Figure 7. Implementation platform for the proto-type information service system.

Figure 8. Examples of baseball information service on the client App.

Only top three of the suggested observation points are displayed at the bottom considering screen size of smart phone. Once any of the observation points is selected by the user, the relevant data are displayed on the top of the screen. The information is overlaid on the video streaming of the ball game.

The system has been applied to the baseball games of Korean Series 2010, and evaluated in the aspects of usefulness of the service and appropriateness of the suggested observation points. The evaluation was carried out by using questionnaire survey from 70 university students after watching a baseball game through the system. They liked baseball game and known baseball knowledge quite well. Figure 9 shows the result of evaluation. Almost 90% users expressed positive opinions in the usefulness, and over 80% users thought the suggested observation points were helpful to understand the progress of the game. Therefore, it is enough to suggest that the information service might be enhance the understanding over the progress of the match, and increase the interest and immersion in the game.

However, some users mentioned the lack of the variety of observation points, caused by the interface of the proto-type application. The interface is designed for display only top three observation points because of the small screen of the device. Therefore, this issue might be resolved by modifying the interface through which the user can search the suggested observation points. Personalization of the observation points might be another resolution of the issue.

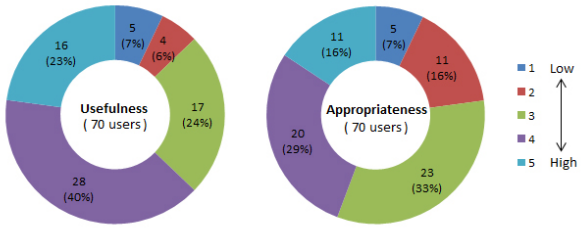


Figure 9. Evaluation results of the proto-type information service system

VI. SUMMARY AND DISCUSSION

In this work, an intelligent information service system for baseball game has been implemented as a context based information service agent. Recording interface for the acquisition of the progress information of baseball game and for the context modeling has also been designed. An ontology model has been structured for being aware of contextual change over time. The ontology is also designed for parsing data to knowledge and extracting observation points. The observation points which is the representation of groups of the information appealing to the user at a certain context are listed in the order of the relative importance and displayed on the mobile device as shown on Figure 8.

The implemented proto-type system satisfactorily offered information at every change of the situation while the system was applied to the real match of Korean Baseball League. In an actual situation, when runners on the first and third base at the out count one, for instance, the system suggested the possibility of hit, base steal, and double play. Once the possibility of double play was selected, the record of the batter on double play against the pitcher during the season was displayed with some other information. As the observation points were listed first, and the record information in detail was supplied only when the observation point was selected, the service was appreciated non-invasive and timely. The scope of this agent can be expanded to other domains like tourism, e-learning, and e-health since the ontology can be modified and expanded as required.

REFERENCES

[1] J. Anhalt, et al., “Toward context-aware computing: experiences and lessons”, IEEE Intelligent Systems, vol. 16, no. 3, pp. 38–46, 2001.

[2] A. K. Dey, G. D. Abowd, and D. Salber, “Conceptual Framework and a Toolkit for Supporting the Rigid Prototyping of Context-Aware Applications”, Human-Computer Interaction, vol. 16, pp. 97–106, 2001.

[3] X. Wang, J. S. Dong, and C. Y. Chin, “Semantic space: an infrastructure for smart spaces”, IEEE Pervasive Computing, vol. 3, no. 3, pp. 32-39, 2004.

[4] H. Chen, et al., “Intelligent agents meet the semantic web in smart space”, IEEE Internaet Computing, vol. 8, no. 6, pp. 69–79, 2004.

[5] O. Brdiczka, J. L. Crowley, and P. Reignier, “Learning Situation Models for Providing Context-Aware Services”, Proc. HCI 2007, LNCS, vol. 4555, pp. 23-32, 2007.

 ISBN: 978-1-61208-164-9

[6] M. Ganzha, et al., “Adaptive Information Provisioning in an Agent-Based Virtual Organization: Preliminary Considerations”, Proc. the SYNASC Conference, IEEE CS Press, pp. 235–241, 2007.

[7] H. Chang, J. Roh, and S. Cho, “Context based Use Control Model for Mobile Device”, Society of Information Science Journal, vol. 14, pp. 63-70, 2008.

[8] A. Marco, et al., “Location-based services for elderly and disable people”, Computer Communications, vol. 31, pp. 1055–1066, 2008.

[9] F. Bently and M. Groble, “TuVista: Meeting the Multimedia Needs of Mobile Sports Fans”, Proc. MM’09 ACM Multimedia Conference, pp. 471-480, 2009.

[10] C. Bettini, O. Brdiczka, and D. Riboni, “A survey of context modeling and reasoning techniques”, Pervasive and Mobile Computing, vol. 6, no. 2, pp. 161-180, 2010.

[11] D. Riboni and C. Bettini, “COSAR: hybrid reasoning for context-aware activity recognition”, Personal and Ubiquitous Computing, vol. 15, no. 3, pp 379-395, 2011.

[12] Major League Baseball (MLB), “Gameday”, http://www.mlb.com, 30.07.2011

 ISBN: 978-1-61208-164-9

Data Center Workload Analysis in Multi-Source RSMAD's Test Environment

Leszek Staszkiewicz, Michał Brewka, Małgorzata Gajewska, Sławomir Gajewski, Marcin Sokół
Faculty of Electronics, Telecommunications and Informatics
Gdansk University of Technology
11/12 G. Narutowicza St., Gdansk, Poland
e-mail: {leszek.staszkiewicz, michal.brewka, malgorzata.gajewska, slawomir.gajewski, marcin.sokol}@eti.pg.gda.pl

***Abstract*—The paper presents the system model of Radio System for Monitoring and Acquisition of Data from Traffic Enforcement Cameras, noting the used network connections and their effective throughput. The article presents the results of test verifying the performance of selected elements of designed system, in terms of image data transmission from multiple sources. Threats that could be caused by large amounts of data transmitted from multiple sources to the database server have been identified.**

Keywords-RSMAD; data center; workload; performance.

I. INTRODUCTION

Radio System for Monitoring and Acquisition of Data from Traffic Enforcement Cameras (RSMAD) provides innovative, integrated and extensive computerized system primarily used for transmission, archiving and exploration of data concerning traffic offenses. The RSMAD system is designed for the police and it is to cover the whole country with its range [1][2].

The research has been aimed at verifying the performance and capabilities of the one of the most important elements of the system – the Data Center (on the basis of the database server). It was very important to test workload of the database server in the multisource RSMAD's environment (presented in Section III).

There are many articles referring to database servers' workload. But it is difficult to find example related with testing database server workload, using many mobile sources of image data.

The paper contains system model and multi-source RSMAD's environment description, database server's performance measurments methods, and interpretation of results. The model of the system, detailing the tested elements and their parameters, will be discussed. The multi-source RSMAD's environment will be presented, including summarized possible throughput. The tests were conducted in variants presenting the actual conditions of the system performance and in variants allowing identification of the risks resulting from the larger amount of incoming data than the predicted one. Therefore, proposals to avoid such threat will be presented.

II. SYSTEM MODEL

The basic element of the RSMAD system is Traffic Enforcement Camera (TEC) enriched by Transmission Module (TM). The module's role is performed by a computer with parameters not worse than presented in Table I. The UMTS/GSM/TETRA modem (router) and dedicated application are also integral elements of the module. The application collects and processes image data from TEC and forms the transport block (compressed and cryptographically protected packet containing image data and included information (in XML format)). So, prepared data is sent by the application to the database server using FTP (*File Transfer Protocol* [3]). Transport block is transmitted using the secure VPN (*Virtual Private Network*) tunnel. The entire image data processing is also to partially relieve the Data Acquisition Center.

Database server performs several key functions in the system:

- Operates the FTP server software allowing recording the transport blocks by TM included in the RSMAD system,
- Processes the received transport blocks, providing decryption, decompression, verification and recording of image data, as well as adding the information included in transport block to database,
- Stores the image data for the application used to generate documentation of traffic tickets for traffic offenses.

The exact technical parameters of the database server are presented in Table I.

TABLE I. HARDWARE PARAMETERS OF SELECTED RSMAD'S ELEMENTS

	Transmission Module	Database Server
Processor	Intel Celeron M 1.00 GHz	Intel E7400 2.80 GHz
HDD	160 GB	250 GB
RAM	1024 MB	2048 MB
Operating system	Windows XP Home Edition x86	Windows Server 2008 Enterprise x86
Network Interface Card	Ethernet 10/100	Ethernet 10/100/1000

 ISBN: 978-1-61208-164-9

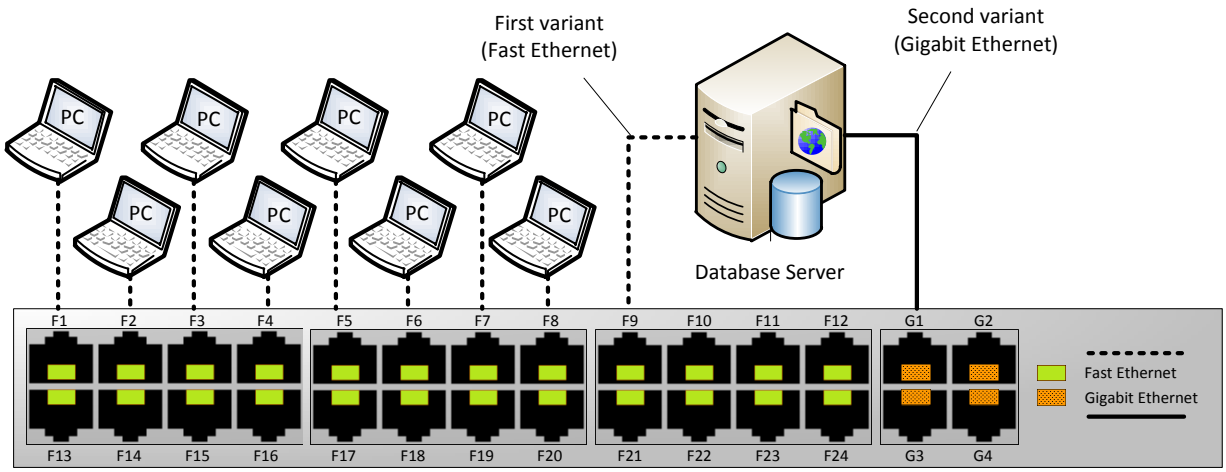


Figure 1. The test environment (2nd and 3rd variant)

Database server is just one of many elements included in the Data Acquisition Center which description exceeds the theme of this article. The detailed architecture of the RSMAD system is presented in [1][2][4].

To create the earlier mentioned VPN tunnels, the Security Gateways (ZyXEL ZyWALL 2 Plus) were used. This solution fundamentally improves the level of security of data transmitted over Internet. [5]

The transmission module, sending transport blocks from TEC, uses the selected subsystems of UMTS/GSM/TETRA for transmission of data. Selection of particular solution, as well as transmission rate offered, depends on specification of particular modem used in the system, configuration of used network and its instantaneous load. RSMAD does not limit technology of transmission which can be chosen by the operator of the system.

The target implementation of the RSMAD system should be characterized by a throughput of the database server link much greater than the throughput rate of transmission module of TEC. Such solution would significantly improve the capacity of the system and would make its work faster and more efficient [6].

III. THE MULTI-SOURCE RSMAD'S ENVIRONMENT

Traffic enforcement camera (mobile or stationary) can take 2 new photos and TM can generate 2 new transport blocks every 5 seconds. Maximum size of one transport block is 2 MB. According to these assumptions, maximum throughputs generated by 100 TMs has been presented in Table II.

TABLE II. MAXIMUM THROUGHPUT GENERATED BY 1 AND 100 TRANSMISSION MODULES

Transmission type used	Maximum throughput (1 TM)	Maximum throughput (100 TMs)
n.a. (maximum value)	6.4 Mbps	640 Mbps
HSUPA (release 6)	5.76 Mbps	576 Mbps
UMTS	384 kbps	37.5 Mbps

Maximum values (in the 1st row in Table II) were calculated on the basis of assumptions presented in the previous paragraph. Other maximum throughputs are dependent on transmission type used.

It is necessary to mention that average transport block size is much lower than 2 MB, real throughputs are lower than theoretical. In addition, mostly TECs do not take maximum number of photos, and not every localization allows for the best transmission type support. Generally, real throughputs are much lower that maximum value.

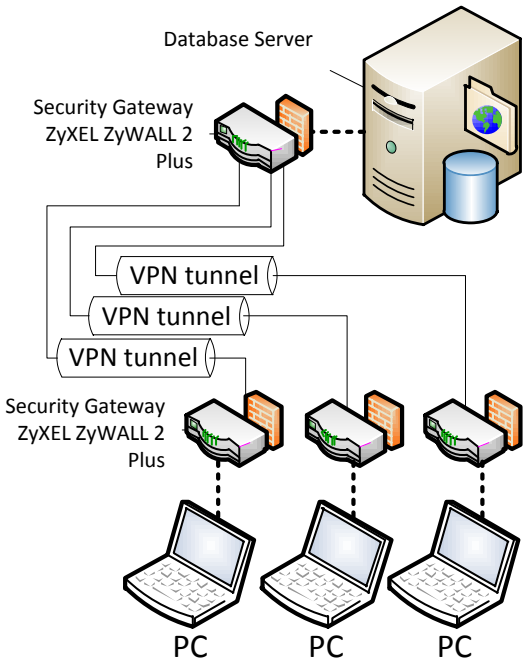


Figure 2. The test environment – the first variant.

 ISBN: 978-1-61208-164-9

IV. DATABASE SERVER'S PREFORMANCE MEASUREMENTS

Due to the expected large number of supported devices, as well as a large amount of incoming image data, the database server was subjected to various performance tests. These tests checked the efficiency of the database server under the various load of its Network Interface Card (NIC), and under load of the database itself. Following the preformed tests, it was decided to further investigate the performance of the hard drive used in the test server. All tests were carried out independently to each other. Analysis of tests' results is presented in Section V.

A. Test of database server performance under Network Interface Card load – the first variant

Tests have been conducted to verify the impact of the amounts of incoming image data on the database server performance. Methods for loading of the NIC were used. Tests were carried out using the FTP server software, running on the database server, and FTP clients, running on other computers (elements of the test environment). Each FTP client was able to establish up to ten simultaneous connections.

During testing the following options backgrounds, a load on the NIC was increased – files were sent from single computer, and then gradually the number of computers sending files was successively increased. Computers and servers were equipped with Gigabit Ethernet. Server performance was measured by registering on Database Server's NIC download rate. Also the percentage usage of CPU was recorded.

In the first variant of the test environment, it was taken into account a combination of 3 FTP clients (10 connections per client) with the server using a secure VPN tunnel (the set up by Security Gateways). The first variant of the test environment is presented in Figure 2.

The maximum rate was limited to only about 20 Mbps (as presented in Table III). This was due to a high degree of transmission security, low productivity and applied Security Gateways. Such a low rate has had a negligible impact on server load – CPU usage has increased by no more than 2 - 3% in comparison to CPU load without using NIC.

TABLE III. DOWNLOAD RATES IN TUNNEL MODE – THE FIRST VARIANT OF THE TEST ENVIRONMENT

Type of IPsec	Type of hash algorithm	Type of cipher	Average data transfer rate in [Mbps]	
			Test I	Test II
ESP	**SHA-1**	**3DES**	*17,12*	*19,36*
		AES-128	*18,48*	*20,64*
		AES-256	*18,00*	*20,32*

B. Test of database server performance under Network Interface Card load – the second variant

In subsequent versions of the test environment (presented in Figure 1), the use of Security Gateways and VPN tunnel was abandoned (for tests only) to check the server performance in a much larger influx of imaging data. In target implementation of RSMAD, security gateways' performance will not influence significantly on efficiency of system work.

The only network device that mediated the transfer was a switch equipped with Fast Ethernet and Gigabit Ethernet interfaces (such solution provided the least possible impact of other devices on the effective transmission rate). Computers – clients' NICs were connected to the Fast Ethernet interfaces, and server's NIC was connected to Fast Ethernet interface of the switch.

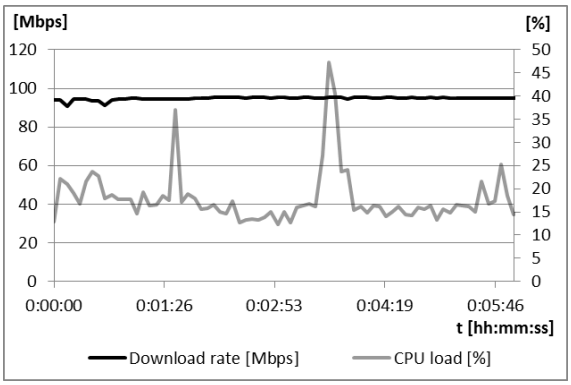


Figure 3. CPU load and NIC load (download rate) of Database Server – the second variant of the test environment.

The Database Server's NIC transmission rate at the limit standard Fast Ethernet was recorded – aggregate file transfer rate reached 96 Mbps (Figure 3). The CPU load of Database Server was in this case was about 15 – 20 %.

C. Test of database server performance under Network Interface Card load – the third variant

In the third scenario of the test environment, similar measurement methods as in the second variant were used.

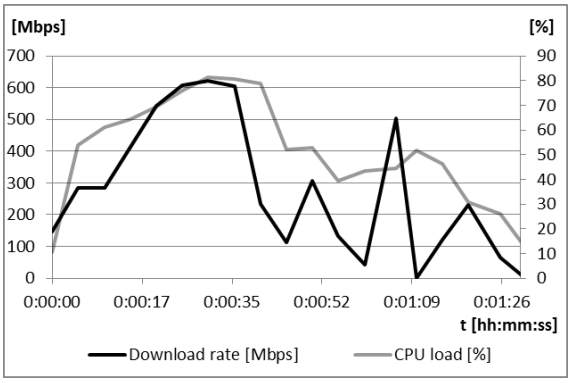


Figure 4. CPU load and NIC load (download rate) of Database Server – the third variant of the test environment.

Computers – clients' NICs were also connected to the Fast Ethernet interfaces but server's NIC was connected to Gigabit Ethernet interface of the switch. The third variant of the test environment is presented in Figure 1.

A significant load of the CPU, reaching up to 80% was observed (as presented in Figure 4). With the achievement

rate exceeding the level of 500 Mbps, the FTP server application has stopped responding and the performance of the operating system has also been disrupted.

D. *Testing of the efficiency of the database server in conditions of database load and testing of the database server's hard drive performance*

During testing, the local connection to the database took place. Also, neither of the available physical NICs was used. Database environment Microsoft SQL Server 2008 was used with AES-128 database encryption. To check the performance of the database server, toughest test conditions were prepared. Each query to the database began with the forging of a new connection.

The procedure was 8 times repeated, beginning with the 1st query, each time increasing the number of requests tenfold. Additionally, all tests were 3 times repeated, and the results averaged [7].

Data saving in the database took place in a continuous manner, taking 50% of the CPU power (continuous operation of one core). Average number of queries per second (with 50% CPU load) is presented in Table IV.

TABLE IV. RESULTS OF TESTING OF THE EFFICIENCYOF THE DATABASE SERVER IN CONDITIONS OF TABABASE LOAD

Number of queries	Processing time [s]	Average number of queries per second
1	0.0170	58.82
10	0.0180	555.56
100	0.0230	4347.83
1000	0.0727	13755.16
10000	0.4773	20951.18
100000	3.9303	25443.35
1000000	35.1163	28476.80
10000000	353.4117	28295.61

In real conditions of the system operations, a database query will not be done continuously. Between queries there will be made a decryption and a decompression of the transport block. Therefore database server's load caused database work will be much lower.

Hard drive performance was tested using a dedicated application. Testing software checked maximum possible speed of data saving to hard drive. With the greatest load write speeds on the hard drive have achieved 85 MBps (680 Mbps). It should be noted that the total load of hard disk with recording generated by one application or one process may prevent the stable operation of the operating system.

V. INTERPRETATION OF RESULTS

Tests conducted in the first scenario of the test environment shows that the transmission and processing of data sent with the maximum allowed rate (for this variant) do not interfere with the work of the database server.

Second, an intermediate variant of the test environment also allows the stable operation of the database server. However, it should be noted that throughput disrupting the work rate of the device (Environmental Test - variant 3), are also feasible to achieve. This is possible by using a sufficiently large number of sources of image data (TECs) using a high-speed links (eg. HSUPA or later subsystems).

After a thorough research of the third variant of the test environment, it was found that the recorded throughput, distorting the FTP server application performance and system application, can be close to the maximum possible data read / write speed to the hard disk of the database server. A disk performance test, has confirmed the hypothesis that the system stability problems really arise when achieving a comparable throughput with a maximum speed of read / write data to the hard drive.

The work of the database server is not only limited to receiving files from multiple image data sources and its processing (e.g., image data information adding to the database). The image data or information about a specific traffic offenses stored in the data base will be used by applications for conduct the process issuing the ticket. All of these processes will load the CPU, RAM, hard drive and the Database Server's NIC.

There is a possibility of increasing the efficiency of application server through:

- The use of a hard drive with higher read/write rate,
- The use of multiple hard drives connected in the matrix increasing read/write speed in comparison to a single hard drive,
- Replacement or installation of new CPUs,
- Installation of larger amount of memory.

In case of absence of a sufficient capacity to process incoming transport blocks this task can also be differentiated between the database server and an FTP server on two physically separate servers. This solution due to the design and architecture of the RSMAD system should not create problems other than a more complex configuration of the software currently drafted.

In order to become independent on hardware, the stability of their servers should be taken care of, through its appropriate configuration on the application layer. The maximum rate allowed on the FTP server should be also limited to the level which does not cause destabilization of the server operating system's work.

VI. CONCLUSION AND FUTURE WORK

The paper identified the risks resulting from the provision of large amounts of data from multiple sources to the database server. Modifications to hardware in order to prevent overloading of RSMAD system's servers were also proposed. Moreover, the need to control system's performance, not only carried out on the hardware level but primarily at the application layer, was pointed.

With the current state of implementation of subsystems for data transmission in mobile networks, the occurrence of described risks is not expected. But they should not be

 ISBN: 978-1-61208-164-9

ignored as in view of the rapid development of the 4G cellular networks, the risk of overloading the equipment with too large number of incoming data is becoming more real.

ACKNOWLEDGMENT

This research work is carried out under research and development grant No. N R02 0034 06 in 2009-2012, in the Department of Radiocommunication Systems and Networks, Faculty of Electronics, Telecommunications and Informatics in Gdansk University of Technology. The work is financed by the National Centre for Research and Development.

REFERENCES

[1] KSSR DT 01.100 v 1.0.1: General concept of RSMAD (in Polish), Gdansk University of Technology, Poland 2009.

[2] S. Gajewski, M. Gajewska, R. Katulski, A. Marczak, M. Sokół and J. Staniszewski, "Radio system for monitoring and acquisition of data from traffic enforcement cameras – features and assumptions of the system," International Conference Transport of 21st Century, Białowieża, September 2010.

[3] Network Working Group J. Postel, RFC959: File Transfer Protocol (FTP).

[4] KSSR DT 07.100 v. 1.0.1: General concept of RSMAD's DAC (in Polish), Gdansk University of Technology, Poland 2009.

[5] M. Sokół, M. Gajewska, S. Gajewski and A. Marczak, "Secure access control and information protection mechanisms in radio system for monitoring and acquisition of data from traffic enforcement cameras," International Conference Transport of 21st Century, Białowieża, September 2010.

[6] M. Brewka and L. Staszkiewicz, Computer Model for Image Data Registration (in Polish), Gdansk University of Technology, Poland 2010.

[7] KSSR RT 04.900 v. 1.0.0: IMOFOT-F and IMOFOT-C interfaces' performance analysis (in Polish), Gdansk University of Technology, Poland 2011.

 ISBN: 978-1-61208-164-9

Usability Evaluation Using Eye Tracking for Iconographic Authentication on Mobile Devices

Claudia de Andrade Tambascia, Ewerton Martins Menezes, Robson Eudes Duarte
CPqD Foundation
Campinas – SP, Brazil
{claudiat, emenezes, robsond}@cpqd.com.br

***Abstract*— This article aims to present the results of a usability evaluation for the use of iconographic authentication on mobile devices as a way to improve security aspects in handling information. This way of authentication was defined in a project called Multimodal Biometric and Iconographic Authentication for Mobile Devices. These assessments were carried out with eye tracking support tools as a means of proving the difficulty points and allow the design decision could be made more accurate to the application final purpose.**

Keywords-Usability evaluation; iconographic passwords; eye tracking observation

I. INTRODUCTION

Considering that human brain recognizes and reminds visual information better than textual information [1][2][3][4], the usage of authentication mechanisms that explore the later rather than the former represent a new paradigm for safety and usability issues in the context of mobile devices due to increasing demands for safer and more flexible new ways of authentication.

In this context, the iconographic authentication may be used to totally or partially lock the device, configuring one level of authentication, which can be then integrated to biometric techniques to increase safety to the process.

A system for local iconographic authentication, where the password verification happens within the device, with no external database access has been proposed. Usability and accessibility issues must be addressed considering the multiplicity of user profiles, including those with little or no familiarity to computing interfaces, with disabilities or with low literacy.

This paper will describe the methodology used to evaluate the usability of iconographic passwords on mobile devices, in the context of a Multimodal Biometric and Iconographic Authentication project [5], as well as presenting the results obtained taking into account aspects like ease of use and memorization, time spent for authentication and the strategies for password creation. This research is important to determine the viability and the benefits of iconographic passwords in this usage context.

In the following sections, the concepts of iconography and usability applied to mobile devices, the prototype employed for iconographic passwords creation and usages, the methodology of evaluation, results obtained and final considerations will be presented.

The section two will present graphical authentication concepts used in this project followed by section three that will present related works. The section four will present a prototype develop for iconographic passwords followed by section five with test methodology used. The section six will present the obtained results followed by section seven with some conclusions and future works.

II. GRAPHICAL AUTHENTICATION

Graphical authentication is a type of knowledge-based authentication that has been explored for over twelve years. It can be basically categorized in three groups [6]: (1) the recall-based authentication systems, in which the user is asked to recall and reproduce a secret drawing, (2) the cued-recall systems, where the user has to remember, (3) and target specific locations within an image and the recognition-based systems, which usually demand the users to memorize a group of images.

As a recognition technique, the iconographic authentication demands less cognitive load than recalls techniques and tends to increase the usability, the security and the user performance, besides being specially appealing in the mobile context, where typewritten input is less common than pointing at the screen.

While some recognition-based systems use faces [7], assuming that the brain has got a special ability to recognize them, other systems use abstract images [8], which are stronger from a security point of view, due to their difficulty of describing. Nevertheless, the use of icons brings a better compromise between usability and safety, once it facilitates mnemonic strategies and then consequently the memorization.

The security level offered by such systems depends on many factors, such as the length of the repository available to the user, the password length, the input method, and the icons themselves which must ideally show similar probabilities of choice avoiding dictionary attacks.

III. RELATED WORK

According to Nielsen and Mack [9], problems with usability found on an interface might be related to different aspects such as user difficulty in learning how to use a system; user delay to complete his/her tasks; deception of the user in operations caused by the system and non-attractive interfaces. Often, the use of inadequate language causes intelligibility problems, which combined with the above

 ISBN: 978-1-61208-164-9

aspects, further contribute to user dissatisfaction and interfering with the quality of experience.

We can find in literature studies like the Jun Gong's one [10], which is based on Shneiderman's "Golden Rules of Interface Design" that proposes a generic set of guidelines for mobile devices. In addition, it is possible to find main mobile manufacturers recommendations, but they usually cover specific cases [11][12], which evidences the lack of standards and consensus mobile usability field.

To ensure effective application of iconographic authentication and to fulfill the user expectations, specific usability aspects for mobile environment, such as extremely dynamic context use and limited user attention, must be taken into account. Applications must be capable to start, stop and resume with little or no human effort, and also have to provide multiple feedbacks and be customized under user needs.

The mobile devices present hardware limitations related to screen size, processing power and input methods. These facts draw attention to images and text size definition as well as buttons displayed on the interface, so the error and cognitive effort rates can be reduced, besides prioritize the choice of icons rather than text input.

IV. PROTOTYPE FOR ICONOGRAPHIC PASSWORDS

One of the main goals of the use of iconic passwords in process of authentication is to increase usability, assuming that the password memorization gets easier for the visual inclined users, and to ensure the user will keep this information in mind for a longer time.

To evaluate iconographic aspects of the authentication process, such as, quality of icons, repertory length, password length and amount of icons displayed it was implemented a prototype that runs in an Android emulator. With this prototype it was possible to do tests in a conventional desktop using an eye tracker device and analyze the user visual behavior during password creation and usage.

A repertory of seventy-two icons was considered as it is shown at Fig. 1. Each column was filled with a category of icons, totalizing twelve categories with six icons each. With the intention of reaching a balance between usability and security the order of the columns and the placement of the icons within the columns change at each interaction action.

Figure 1. Icons repertory considered in the prototype.

The categories considered for the icons were fruits, animals, technological devices, and means of transportation, balls, sea elements, musical instruments, scholar material, smiles, banners, body parts and hats.

The selection of such icons in each category was made after an analysis of possible strategies to help memorization of iconic passwords and, consequently, usability increase of iconographic authentication, without loss of mathematical security, as presented in [13].

The first screen of the prototype allowed the user to choose the option of creating or using an iconographic password. It was possible to configure the password length, as shown at Fig. 2.

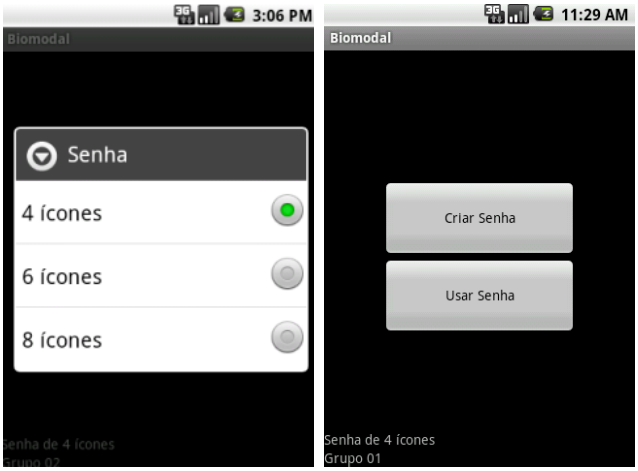


Figure 2. Prototype initial screen.

Passwords were categorized in fixed lengths of four, six and eight icons and tests with five users for each of this category were defined, according to the methodology presented in the following section.

V. TEST METHODOLOGY

A. User selection and tests frequency

The first step for the test realization was to determine the users that would participate on it as well as the frequency of the test sessions.

According to [14], practitioners of usability recommend many different quantities of interviews, for several different reasons. For shipping products, it was recommend six interviews, for the following reasons: i) six one-hour interviews can be conducted in one calendar day; ii) testing six respondents allows you to identify trends. For this project five users were selected for each category of password.

These users were divided into man and woman with under-thirty years old, thirty to forty-five year old and forty-five years old or older, as showed in Fig. 3.

 ISBN: 978-1-61208-164-9

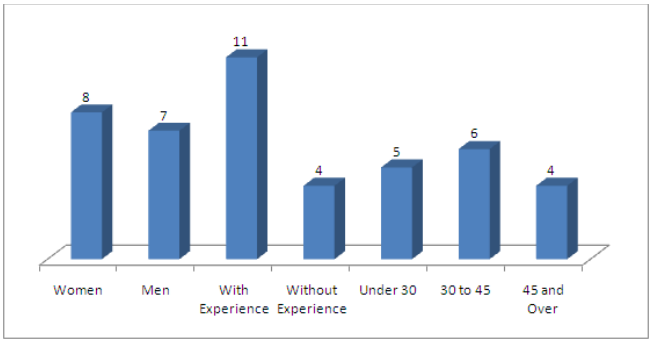


Figure 3. Summary of user profiles (gender, level of experience and age groups).

Such division was necessary once the memorization was to be measured, which implied the need of considering different age groups. It was also taken into account previous experience with mobile devices, for it could influence the usability requirements.

The users were separated into three groups, with five users each to optimize test sessions in a work day. After the four first interactions with the first group, the test sessions with group occurred.

Since one of the main criteria to be evaluated will be the memorization of passwords after a period of time. It was defined that the test would happen during fifteen days according to the schedule presented at Fig. 4.

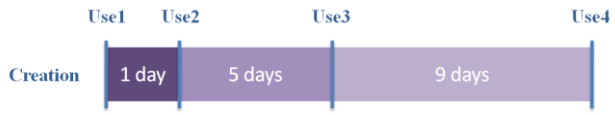


Figure 4. Tests schedule.

The first use was composed by password creation and confirmation. If the user could not be able to remember the password the creation process was restarted. The second use had to be one day after its creation. The third use was planned to happen five days after password creation, encompassing a weekend that could influence password memorization. Finally, the fourth use had to happen nine days after password creation for it was a period considered enough for the password is forgotten in case the memorization strategy failed.

B. *Test configuration*

All the tests were performed in a usability laboratory composed of two rooms: the participant room and the observation room. The eye tracking device employed was Tobii Eye Tracker T60 [15], which consists of a seventeen inches display with cameras and embedded infrared sensors. The iconographic authentication prototype run on this display and the interaction method for selection and browsing was the mouse.

During the test execution users were asked to keep a distance of about sixty centimeters far from the display to enhance eye tracking quality. The screen recording was made together with viewpoints using gaze plots, heat maps, and audio. User's expressions were also important in case the eye tracker might not capture relevant data.

C. *Results tabulation*

After the realization of four tests rounds, the user's performance was grouped according to password length and compared with the time spent during the creation and usage. For passwords with four icons, a concern in the password creation was observed, which lead a better time performance in confirmation task as shown at Fig. 5.

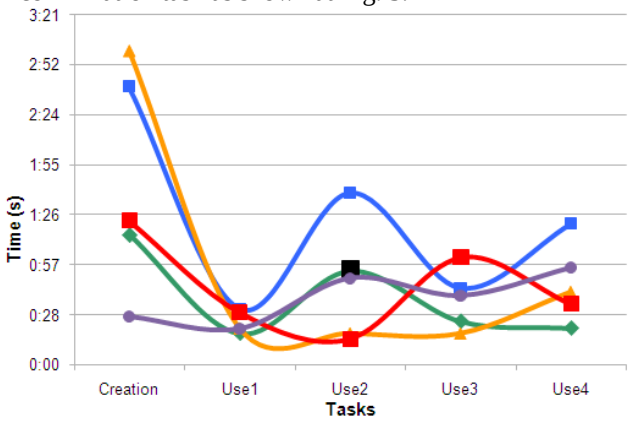


Figure 5. Evolution of users performance for passwords with four icons.

There was only one mistake by one user in his second interaction, identified by a black square in the following graphic shown at Fig. 5. Further interactions happened with no significant changes.

For six icons passwords, the user's behavior presented wider variation and more mistakes after the second interaction, as shown at Fig. 6.

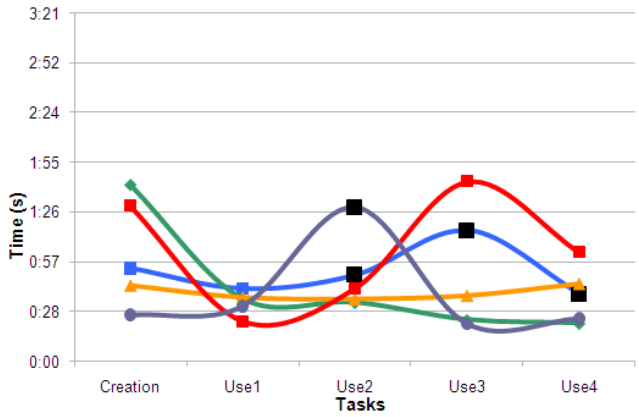


Figure 6. Evolution of users performance for passwords with six icons.

Password creation, compared with four icons password, was faster, which can explain why subsequent interactions lasted longer and had more mistake occurrences.

Finally, for eight icons password, it was possible to observe the best performance in interactions after password creation, except for a user who quickly created the password

 ISBN: 978-1-61208-164-9

and could not remember it in the subsequent interactions (as shown at Fig. 7.

It is important to highlight that the mentioned user made a mistake in the second interaction and last a long at the third one, but after that he would not forget the password anymore, showing satisfactory result in the last interaction

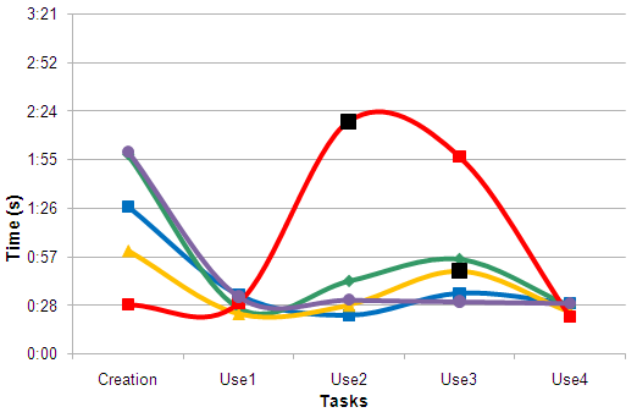


Figure 7. Evolution of users performance for passwords with eight icons.

D. Eye tracking data

Based upon Duchowski's recommendations [16] for experiments involving eye tracking data it was observed the visual behavior of the participants during the iconographic password creation and use processes.

The tests videos were separate into relevant video segments and using the gaze plot resource was possible to compare different ways of exploring visually the grid of icons. Three samples of interaction schemas by the time of password creation were shown in Fig. 8. The circles represent the visual fixation points and its size is proportional to the duration of the look.

(a) (b) (c)

Figure 8. Ways of eye interaction in password creation.

Fig. 8 (a) presents the behavior of a user who analyzes the icons in a chaotic or random way, searching for familiarization among the images for the creation of the password. Fig. 8 (b) presents the behavior of a restrained user that what limits his/her field of view to two categories, aiming at easing password creation. Fig. 8 (c) presents the behavior of a methodical user who observes invariably all icons available on the prototype to make his/her choices.

In all cases of user behavior it was possible to notice a concern in observing icons according to the strategies chosen to ease the memorization process.

E. Memorization strategies

After password creation, the test conductors tried to infer the strategy used to memorize it. At the end of the round of testes the users were briefly interviewed and asked about strategies used and face difficulties.

During the test realization it was observed that iconographic passwords enable a wide universe of possible combinations which favors the use of most varied strategies of memorization and creation of passwords that are stronger and less susceptible to dictionary attack.

The strategies are extremely dependents on the repertory of icons used and the amount of icons to be memorized and a direct relation among the level of difficulty perceived by each user was observed.

Table 1 shows a categorization of strategies used by each participant of the tests.

TABLE I. MEMORIZATION STRATEGIES

Strategies	Users
Peer association	5
Creating history	3
Category elimination	3
Icons with similiar colors	2
Cultural issues	2
Memorization of individual icons	2
Visual affinity with icon	2
None specific	9

Have a fixed or a free password order strongly influences the creations strategies and for this reason it was not announced to the participants that the icons order did not matter on creation process. Even so, many users created and used the password in order, especially in the cases where there were associations of the icons to stories or sentences. Nonetheless, the larger the password was, the less users selected icons in the created order.

VI. RESULTS

Many performance factors such as the execution time, error rates, screen browsing and password memorization in a time interval were evaluated.

In this context, passwords composed by four icons were the ones which presented best use performance, with lower error rates and time average in all interactions. The eight icons passwords were the ones with more occurrences of history creation as a way to improve the memorization and when histories were not created, the elimination of categories was used to reduce the possibilities and to ease the password creation.

It was possible to observe that there is not a direct relation between the quantity of icons in a password and the time necessary for authentication, for users with different

 ISBN: 978-1-61208-164-9

password sizes get very similar performance in authentication. However, the time spent to create passwords may influence directly its memorization, that is, users who spent more time creating the password had less error rates and tended to not forget it.

This fact was ratified by testimonies of the users who have not remembered the password and assigned it to the lack of attention by the time of the password creation. It indicates the need of thinking about strategies which may favor the password memorization.

As one of the premises of the test was that it did not need to be entered in the creation order, many users that had initially created strategies prioritizing the icons order finished by abstracting this strategy as the familiarization to the password increased. It corroborates researches that point as a usability factor the use of passwords by order and disorder, a different paradigm from the one used for traditional passwords.

The choice frequency of the icons to compose the iconographic passwords was another important factor. It was possible to observe a regular distribution of the icons where the most and less chosen categories kept constant, as shown at Figure 9.

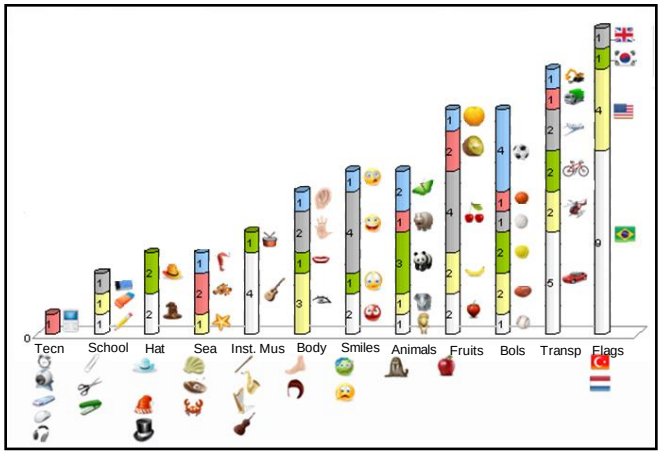


Figure 9. Distribution of icons chosen by the participants.

The user preference for gaudy icons was also observed. The icons under technology devices category, which were the less chosen ones, had grey as predominant color. However, cultural and esthetic aspects had major choice weight. For this reason, icons such as Brazilian banner, soccer ball, red car, and cherry were chosen much more than orders.

The appropriated feedback at the authentication proved to be a key factor to guide the user throughout the process. As observed in the survey the user needs to know how many icons have already been chosen, should be clear the number of screens, if there are more than one, and the current screen. This information must be presented to users in clear and non-blocking way avoiding the slowdown of the authentication process.

At last, it was identified a characteristic named “love at first sight” that consists of a situation in which the user thinks about choosing an icon to compose his/her password, but he/she does not do so. By the time of the authentication, the user then finds him/herself in doubt about chosen and will not know whether it compounds or not the password. This problem may be solved by considering a little training process just after password submission. This training may include, e.g., three authentication simulations before password registration. Alternatively, the system could itself choose the password for the user what, on the other hand, could reduce the usability of the solution.

VII. CONCLUSIONS AND FUTURE WORKS

As the use of icons to make authentication can be considered a new paradigm, it was adopted an initial approach considering six icons passwords to reach an adequate usability level and allowing the use of histories to ease memorization and reducing time spent for creation and use of the passwords.

On the other hand, it is reasonable to claim that as the familiarity of the users to this new paradigm increases, the results of usability tests will be better either. And higher levels of security will also be reached.

At last, in the performed tests it was also observed a preference for some icons instead of others, what may weaken the security of iconographic authentication. But there is a positive feature that reduces this risk which is the possibility of substituting some icons as well as letting to the system the definition of the password.

Considering the security and usability aspects presented in the Sections II and III, some project decisions were established to lead the prototype implementation. In a general way, the users that made the usability tests with four icons did not only have a very good performance but also reported that it was easy to memorize only four icons. As for the six and eight password users, they had very heterogeneous test result and performance.

The security comparison between iconographic and alphanumeric passwords in the evaluated scenarios, it was defined an initial password size of six icons within a repertory of ninety icons. This repertory was split into three screens with thirty icons each (six rows and five columns). The use of bigger repertory of icons impacts significantly in the usability of the solution, resulting in the increasing of the security not proportional to this impact, which makes impossible the use of a bigger repertory. As for the size of the password, it could be bigger, and it would result in an increment of security for the solution. But, as it deals with a paradigm change, it was decided to maintain initially the six icons for iconographic authentication.

The tests performed in the context of this project have presented an analysis, which contributes for the definition of the iconographic authentication solution that will be used for implementation of the functional prototype of project Multimodal Biometric and Iconographic Authentication for Mobile Devices.

Despite the missing standardization among the methods it’s possible to ensure that iconographic password systems meet the usability and security requirements. As a future work it’s necessary to provide a comparison between the

 ISBN: 978-1-61208-164-9

performance of iconographic and alphanumeric paradigms, besides test the proposed solution in real mobile devices with dynamic usage contexts.

ACKNOWLEDGMENT

We express our gratitude to FUNTTEL – Fundo Nacional das Telecomunicações, do Ministério das Comunicações, which funds this research.

REFERENCES

[1] B. Kirkpatrick. "An experimental study of memory". Psychological Review, 1894, 1:602-609.

[2] S. Madigan. "Picture memory". In J. Yuille, editor, "Imagery, Memory, and Cognition: Essays in Honor of Allan Paivio", cap.3, pp.65-89. Lawrence Erlbaum Associates, 1983.

[3] A. Paivio, T. Rogers, and P.C. Smythe. "Why are pictures easier to recall than words?" Psychonomic Science, 1968, 11(4):137-138.

[4] R. Shepard. "Recognition memory for words, sentences, and pictures". Journal of Verbal Learning and Verbal Behavior, 1967, 6: pp. 156-163.

[5] R.P.V. Violato, M.U. Neto, F.O. Simões, I.M.A. Ávila, M.A. Angeloni, T.F. Pereira, E.T. Nakamura, and R.S. Cividanes. "BIOMODAL Project – Multimodal Biometric Authentication and Image-Based Authentication for Mobile Devices". 8th Sumer School for Advanced Studies on Biometrics for Secure Authentication. 2011.

[6] R. Biddle, S. Chiasson, and P.C. Van Oorschot. "Graphical Passwords: Learning from the First Twelve Years", School of Computer Science, Carleton University, 2011.

[7] Passfaces Corporation. The Science Behind Passfaces. Write paper, http://www.realuser.com/enterprise/resources/white_ papers. htm, accessed May 2011.

[8] R. Dhamija and A. Perrig. "Déjà Vu: A user study using imagens for authentication". In 9th USENIX Security Symposium, 2000.

[9] J. Nielsen and R. Mack. "Usability Inspection Methods." New York, John Wiley & Sons, Inc.,1994.

[10] J. Gong and P. Tarasewich. "Guidelines for handheld mobile device interface design". College of Computer and Information Science, Northeastern University, Boston, USA, 2004.

[11] R. Blum, K. Khakzar, and W. Winzerling. "Mobile Design Guidelines in the Context of Retail Sales" Support Fulda University of Applied Sciences. Germany, 2008.

[12] S. Chan, X. Fang, J. Brzezinski, Y. Zhou, X. Shuang, and L. Jean. "Usability For Mobile Commerce Across Multiple Form Factors". Journal of Electronic Commerce Research, vol. 3, no. 3, 2002.

[13] I.M.A. Avila, "Estratégias Mnemônicas para Senhas Icônicas", Abril de 2011.

[14] S. Weiss, "Handheld Usability", Ed. John Wiley & Sons Ltd, New York, 2002.

[15] Tobii Eye Tracking. "Screen based eye tracking - Tobii T60 & T120". http://www.tobii.com/en/analysis-and-research/ global/products/hardware/tobii-t60t120-eye-tracker/, accessed May 2011.

[16] A. Duchowski. "Eye Tracking Methodology: Theory and Practice", Springer, 2nd edition, 2007.

 ISBN: 978-1-61208-164-9

Meeting the Challenge of Global Mobile Phone Usability

Design and practices

Yan Cimon
CIRRELT
Université Laval
Quebec City, Canada
yan.cimon@fsa.ulaval.ca

Fatima-Zahra Barrane
Faculty of Business Administration
Université Laval
Quebec City, Canada
fatima-zahra.barrane.1@ulaval.ca

Diane Poulin
CIRRELT
Université Laval
Quebec City, Canada
diane.poulin@fsa.ulaval.ca

***Abstract*— How can mobile phone design increase usability and the user's experience in a global setting? The purpose of this paper is to put forth general design principles to enhance usability. We use a structured literature review of 307 peer-reviewed papers. We find that shared interaction, intuitiveness and personalization should drive mobile phone design. Industry practices are examined and future research avenues are suggested.**

Keywords-usability; design; principles; global practices.

I. INTRODUCTION

How can mobile phone design increase usability and the user's experience in a global setting? Information technology (IT) and mobile phones, commonly referred to as Information and Communication Technologies (ICT) are now pervasive in our everyday lives and that pervasiveness increases as computers and phones converge as they are fast becoming information appliances. While there are still non-adopters, the non-adoption of ICTs may be explained by a range of attitudinal, skills-related and infrastructure-related factors [1]. However, non-adoption, as it relates to literacy issues, is weaker in the case of mobile phones [2]. Possible explanations are the availability of complementary infrastructure, efficient enough providers [e.g., 3], and easily accessible related innovations [4]. This implies that mobile phones are within reach of most of the world's population: Somalia – a failed state – ranks 16th out of 42 African countries in terms of penetration [5].

In the global arena, mobile phones are an empowerment tool as they help develop human capital [6]. In Mexico, they free users from geography-related institutional constraints [7]. In rural Indian areas, they help the less fortunate and are a catalyst for social change [8]. In the Middle East, they positively influence economic freedom [9]. Conversely, a lack of access to mobile technologies may hinder development [10] and even promote inequalities [11].

This paper is structured as follows. First, we state the problem we are addressing which is that of designing an optimal user experience in a global and diverse context. Second, we explain the context of this research as it relates to the impact of literacy on ICT use. Then we look at how users' familiarity with technology influences design in terms of accessing information. Third, we discuss the methods used which consist in a structured literature review of a large number of articles. Fourth, the resulting design principles are examined in terms of their contribution to the improvement of usability and some industry practices are presented. Finally, we conclude and discuss potential implications for academics, businesspeople and policymakers.

II. PROBLEM STATEMENT

Mobile phones – and ICTs in general – have proven very popular in recent years. Globally, they are used in a variety of contexts: from highly educated and "sophisticated" users who exploit every single functionality, to utilitarian users that focus exclusively on core functions (i.e. dialling calls only or simple web surfing), to users with low or no literacy who have developed strategies to overcome this obstacle in order to make mobile phones a tool for meaningful interaction and/or community engagement. Also, there are a wide variety of geographical and cultural challenges to mobile phone usage as pictograms and input-output methods and devices may be culturally biased even if they are globally standardized.

Notwithstanding this, accessing information is highly dependent on interface design and usability. Tools and technique are evolving at a rapid pace and user adaptation would be easier if interface design followed sets of principles that would be consistent globally and culturally, among other variables.

Thus, in extremely diverse usage contexts as reflected by an ever-increasing global user base, it is important to determine how mobile phones may be designed for optimal usability notwithstanding the potential users' levels of literacy or familiarity with technology.

III. CONTEXT

A. Literacy

In the global arena, the challenge of literacy is often overlooked when considering usability issues. The traditional concept of literacy tied to the written language is obsolete as it is not essential to access information and knowledge anymore if only because of the penetration of television and mobile phones [see 12]. Literacy is now considered to be « digital » [13] insofar as access to ICT is now deemed

 ISBN: 978-1-61208-164-9

essential by UNESCO [14]. Thus, designers need to provide an information-rich experience to users of ICTs notwithstanding their literacy level, in synch with the growing migration of content on these platforms.

B. Familiarity with technology

Similarly, users' familiarity with technology is another design challenge. Traditionally, user interface tools for accessing information have consisted in series of menus that may be classified in three categories. First, hierarchical menus are a commonly used tool that reflects the organization of information. Although they are not necessarily intuitive and require a great deal of both classical and digital literacy skills, they have the benefit of being systematic. Second, fisheye menus blow up certain parts of the information a user is querying to make it easier to discriminate between the choices offered to this specific user. They often rely on text and icons. Third, tag clouds organize information according to its popularity. This is very useful for users with preferences and search patterns aligned on those of related groups of users. Designers need to find ways to divorce familiarity with technology and access to content so that ICTs become an enabler for all potential or actual users.

IV. METHODS

We used a structured literature review process in order to determine a set of principles and practices for the design of highly usable mobile phones and ancillary applications. We reviewed a total of 307 peer reviewed papers that focused on keywords related to mobile phone usage and design like: literacy, usability, sales, adoption, etc. Papers were then grouped by themes that were clustered around design principles and industry applications.

With regards to the validity and reliability [15] of this review, the breath of articles we surveyed allows for a good level of reliability, while the keywords chosen and our preoccupation for examining diverse research fields and cases [16] provides validity to this research. We were thus able to extract design principles and industry practices that increased usability and user experience.

V. DESIGN PRINCIPLES AND INDUSTRY PRACTICES

Our structured literature review first led to general design principles. Then, real-life examples were extracted in light of those general principles.

A. Designing interfaces for usability

Usability is thought to be a driver for interface adoption. As a general rule, interface design should be centered on the user to allow increased confidence and meaning in their interactions with – and through – mobile phones. Especially in the case of interface design for low literacy users [17], any sustainable solution must be coherent with their geographic or cultural contexts [18]. Thus, design principles may be drawn around three core ideas: shared experience, intuitiveness, and personalization.

1) Allow users a shared interactive experience: A common experience, in the form of a common platform or as an ability to access and exchange similar content, is conducive to higher quality knowledge flows between users [19]. It must also allow for various modes of interaction (i.e., with keyboards, touchscreens, videos, or other input/output modes, etc.) [see 20, 21]. Massive multiplayer online video games are a case in point [see 22]. Thus, social relationships and interface interactions benefit greatly from being designed like human relationships [23].

2) Focus on intuitiveness: Reducing the users' cognitive load increases their interaction accuracy. This implies that processes underlining a given interface have to make it user-friendly and rich [e.g., 24]. Minor changes to handsets and a streamlining of the interface (colors, symbols, etc.) significantly improve usability [17]. Furthermore, an elegant interface influences the perception of usability [e.g., 25] as this concept has both objective and subjective components [26].

3) Personalization and user-control matter: Some systems offers choices for increased personalization [27] that go as far as being sensitive to a particular context [28]. Usability may also be improved by increasing the level of control exerted by the user and through a « help » section [29]. Social tagging [30] and the combination of visualization and voice activated features are another set of useful tools [31].

These principles also find their way into industry practice.

B. How global industry leaders do it: some examples

Global industry leaders have embodied these principles in their practices and products to varying degrees. These may be grouped into three categories: 1) the relationship between literacy and the handset; 2) interacting with the mobile phone; and 3) leveraging the possibilities offered by a mobile phone.

1) The handset and literacy: Few global leaders have focused on low literacy customers. Yet in many emerging markets, these customers constitute the bulk of mobile phone users. In developed countries, many users have literacy-related challenges as well. In 2008, Nokia and Motorola have come up with different ways of addressing these challenges from intensive R&D effort into the matter. While Nokia focused on the physical ergonomics of the handset, simpler menu functions, and an icon-based menus; Motorola opted for a larger screen with bigger letters and numbers as well as more classical, yet streamlined, hierarchical menus. At that time, both handset makers focused on trying to increase intuitiveness while providing limited user control through a less complex phone.

2) User interaction with/via a handset. Global industry leaders have often sought efficient ways to leverage user interaction with – and through – mobile phones. Web browsing, email, social media, film and picture applications – and eventually 3D applications – provide additional

 ISBN: 978-1-61208-164-9

richness and connectedness beyond the voice transmission capabilities of traditional mobile phones. Handset makers offer users differentiated modes for interaction (reticular screen vs mini-keyboard) and different means for scrolling content. While the iPhone and traditional Blackberry handsets offered very different user experiences, the Blackberry Torch constitutes an example of the possible convergence between these products as it offers a reticular screen and a physical QWERTY keyboard. It also offers "apps", just like Android-based phones and Apple devices do; the latter having been the pioneer in marketing "apps" efficiently.

3) Leveraging the handset. Mobile phones offer increasing possibilities for interaction trough web-browsing and "apps". Google is a case in point. Its search engine interface design is extremely simple, streamlined yet very efficient, especially for mobile devices. It is intuitive as it only has one text-entry box. It allows for personalization as users may filter search results according to certain preferences. It is easy to use on mobile platforms and is a recognized leader in keyword searches. The Amazon Kindle also embodies these principles. While one can procure a physical e-reader, it is also possible to download a Kindle application that allows one to read Kindle books on an iPhone, an Android handset or a Blackberry.

VI. CONCLUSION AND IMPLICATIONS

More broadly, a continuous focus on design for user-centered usability has many implications for business, government and research.

A. Implications for business and government

Business and governments will benefit greatly from better mobile phone and interface design. This will provide for better, high-context and information-rich transactional environments. Service delivery will be improved for all customers, including those with literacy challenges. Furthermore, it will enable implementation mechanisms that will not be sensitive to geography or context, thus providing value-added solutions to emerging economies as well as developed ones.

B. Implications for academics

This research has a range of implications for academics. First, it is a multidisciplinary foray into mobile phone usability in a global context. Second, it helps refocusing this research field beyond engineering requirements or marketing issues, which are useful but provide a partial picture.

C. Future research

In conclusion, the literature, the design principles and industry practices derived from this research call for a return to basic user-oriented principles to drive future research: 1) more should be done to find ways to reduce the cognitive load attached to mobile phone usage; 2) both engineering and the social science aspects need to be simultaneously considered to tackle the issues that surround usability; 3) simplicity and interaction richness need to become a focus for industry leaders. Immediate future steps for this research involve a careful examination of the impact of literacy that could help explain the effect of this variable on mobile phone usage for complex tasks and transactions, especially for users with no or limited levels of literacy. Furthermore, the related technical challenges need to be dealt with following the determination of a roadmap detailing the engineering requirements and technology development that will flow from this research. In-depth cases, examples and user life stories are tools that will lead to an increased understanding of the particular technologies that are bound to shape the implementation of these principles. From healthcare management challenges in the emerging world [32] to more participation and citizen engagement in Africa [33], mobile phones have become a part of global communities and thus should be on a path to increased usability by the widest range of customers on a global scale.

ACKNOWLEDGMENTS

The authors would like to thank the *Ministère des services gouvernementaux du Québec;* The *Fonds québécois de la recherche sur la société et la culture* (FQRSC) and the Research start-up fund of *Université Laval*. The comments of the three anonymous reviewers are gratefully acknowledged as they have improved this paper. The usual caveats apply.

REFERENCES

[1] P. Verdegem and P. Verhoest, "Profiling the non-user: Rethinking policy initiatives stimulating ICT acceptance," *Telecommunications Policy*, vol. 33, pp. 642-652, 2009/12// 2009.

[2] V. Andonova, "Mobile phones, the Internet and the institutional environment," *Telecommunications Policy*, vol. 30, pp. 29-45, 2006.

[3] H. Junseok, C. Youngsang, and L. N. Viet, "Investigation of factors affecting the diffusion of mobile telephone services: An empirical analysis for Vietnam," *Telecommunications Policy*, vol. 33, pp. 534-543, 2009.

[4] P. Rouvinen, "Diffusion of digital mobile telephony: Are developing countries different?," *Telecommunications Policy*, vol. 30, pp. 46-63, 2006.

[5] B. Powell, R. Ford, and A. Nowrasteh, "Somalia after state collapse: Chaos or improvement?," *Journal of Economic Behavior & Organization*, vol. 67, pp. 657-670, 2008.

[6] A. Ashish and A. Suma, "The software industry and India's economic development," *Information Economics and Policy*, vol. 14, pp. 253-273, 2002.

[7] A. Veneta, "Mobile phones, the Internet and the institutional environment," *Telecommunications Policy*, vol. 30, pp. 29-45, 2006.

[8] S. Mehta and M. Kalra, "Information and Communication Technologies: A bridge for social

 ISBN: 978-1-61208-164-9

equity and sustainable development in India," *The International Information & Library Review*, vol. 38, pp. 147-160, 2006.

[9] F. Shirazi, R. Gholami, and D. Añón Higón, "The impact of information and communication technology (ICT), education and regulation on economic freedom in Islamic Middle Eastern countries," *Information & Management*, vol. 46, pp. 426-433, 2009.

[10] R. S. Subba, "Bridging digital divide: Efforts in India," *Telematics and Informatics*, vol. 22, pp. 361-375, 2005.

[11] E. Forestier, J. Grace, and C. Kenny, "Can information and communication technologies be pro-poor?," *Telecommunications Policy*, vol. 26, pp. 623-646, 2002.

[12] A. Robinson-Pant, "Changing discourses: Literacy and development in Nepal," *International Journal of Educational Development*, vol. 30, pp. 136-144, 2009.

[13] I. P. Pandey, "Literate lives across the digital divide," *Computers and Composition*, vol. 23, pp. 246-257, 2006.

[14] M. Shiohata, "Exploring literacy and growth: An analysis of three communities of readers in urban Senegal," *International Journal of Educational Development*, vol. 29, pp. 65-72, 2009.

[15] E. G. Carmines and R. A. Zeller, *Reliability and Validity Assessment* vol. 07-017. Newbury Park CA: Sage, 1979.

[16] R. K. Yin, *Case Study Research* vol. 5. Thousand Oaks CA: Sage, 1994.

[17] Z. Lalji and J. Good, "Designing new technologies for illiterate populations: A study in mobile phone interface design," *Interacting with Computers*, vol. 20, pp. 574-586, 2008.

[18] S. Tino, "E-Government in developing countries: Experiences from sub-Saharan Africa," *Government Information Quarterly*, vol. 26, pp. 118-127, 2009.

[19] Y. Cimon, "Designing Computer Supported Collaborative Work around Knowledge Flows," *Proceedings of the Decision Sciences Institute Annual Meeting*, pp. 3891-3896, 2009.

[20] S. Bodker and Y. Sundblad, "Usability and interaction design - new challenges for the Scandinavian tradition," *Behaviour & Information Technology*, vol. 27, pp. 293-300, 2008.

[21] J. York and P. C. Pendharkar, "Human-computer interaction issues for mobile computing in a variable work context," *International Journal of Human-Computer Studies*, vol. 60, pp. 771-797, 2004.

[22] C. S. Nam, S. Johnson, Y. Li, and Y. Seong, "Evaluation of human-agent user interfaces in multi-agent systems," *International Journal of Industrial Ergonomics*, vol. 39, pp. 192-201, 2009.

[23] H. Schaumburg, "Computers as Tools or as Social Actors? The Users' Perspective on Anthropomorphic Agents," *International Journal of Cooperative Information Systems*, vol. 10, p. 217, 2001.

[24] Y. Zou, Q. Zhang, and X. Zhao, "Improving the Usability of E-Commerce Applications using Business Processes," *IEEE Transactions on Software Engineering*, vol. 33, p. 837, 2007.

[25] N. Tractinsky, A. S. Katz, and D. Ikar, "What is beautiful is usable," *Interacting with Computers*, vol. 13, pp. 127-145, 2000.

[26] K. Hornbæk, "Current practice in measuring usability: Challenges to usability studies and research," *International Journal of Human-Computer Studies*, vol. 64, pp. 79-102, 2006.

[27] A. Bunt, C. C., and M. J., "Mixed-Initiative Interface Personalization as a Case Study in Usable AI," *AI Magazine*, vol. 30, p. 58, 2009.

[28] G. Calvary, J. Coutaz, D. Thevenin, Q. Limbourg, L. Bouillon, *et al.*, "A Unifying Reference Framework for multi-target user interfaces," *Interacting with Computers*, vol. 15, pp. 289-308, 2003.

[29] S. Y. Chen and R. D. Macredie, "The assessment of usability of electronic shopping: A heuristic evaluation," *International Journal of Information Management*, vol. 25, pp. 516-532, 2005.

[30] F. Carmagnola, F. Cena, L. Console, O. Cortassa, C. Gena, *et al.*, "Tag-based user modeling for social multi-device adaptive guides," *User Modeling and User - Adapted Interaction*, vol. 18, p. 497, 2008.

[31] M. Howell, S. Love, and M. Turner, "Visualisation improves the usability of voice-operated mobile phone services," *International Journal of Human-Computer Studies*, vol. 64, pp. 754-769, 2006.

[32] H. Lucas, "Information and communications technology for future health systems in developing countries," *Social Science & Medicine*, vol. 66, pp. 2122-2132, 2008.

[33] A. J. Njoh, "African cities and regional trade in historical perspective: Implications for contemporary globalization trends," *Cities*, vol. 23, pp. 18-29, 2006.

 ISBN: 978-1-61208-164-9

Communication Needs of Japan and the United States: A Comparative Analysis of the Use of Mobile Information Services

Qazi Mahdia Ghyas, Fumiyo N. Kondo, Takayuki Kawamoto
Dept. of Social Systems & Management
University of Tsukuba
Ibaraki, Japan
E-mail: s1030160@u.tsukuba.ac.jp, kondo@sk.tsukuba.ac.jp, takaroom1118@yahoo.co.jp

***Abstract*—Mobile marketers are anxious to gain knowledge about the use of mobile services in different cultures and countries. The aim of this research is to construct a method for comparing consumer demand for mobile information services in different countries. We attempted to gain a understanding of the cross-national needs structure through a comparison of use intentions between the United States (at the University of California at Los Angeles) and Japan. Toward this end, we extracted use intention factors from both the locations. The results confirmed the following four factors: the information-intensive factor, the entertainment factor, the low penetration service factor, and the communication tool factor. This study also found that the two countries have different needs characteristics for a certain mobile communication service, i. e., mobile e-mail, and roughly the same needs characteristics for mobile entertainment services and for mobile information services except "radio".**

Keywords- Cross-national study; Information intensive; Low penetration service; Entertainment; Communication.

I. INTRODUCTION

By the end of 2010, there were 5.3 billion mobile subscriptions worldwide. That equates to 77% of the world's population [1]. This represents a large increase from the 4.6 billion mobile subscriptions in existence at the end of 2009. The increase in mobile service usage around the world has been driven by both advanced technologies and the growing number of service options available to consumers. For the most part, these services include mobile searches, news and sports information, music and video downloads, e-mail, and instant messages [1]. The explosive growth in the use of mobile devices is frequently noted in research studies [2], [3].

Despite the growing importance of mobile devices, few studies have been conducted using a cross-national approach. The usages of mobile devices vary considerably among different countries [4]. The adoption of mobile services and technology does not appear to follow any single universal logic or pattern for different countries [2]. Harries *et al.* [5] investigated the role that culture plays in explaining differences in adoption, usage, and attitudes with regard to mobile services by comparing the United Kingdom and Hong Kong. Cho [6] explored how mobile phone users in the United States and Korea adopt both existing and potential mobile services. Lee *et al.* [7] investigated the different usage patterns among mobile users in Korea and Japan and interpreted these patterns within the framework of a value structure. Vrechopoulos *et al.* [8] conducted sociological research and found Finland to be the most mature mobile market when compared with Germany and Greece. They identified critical success factors and noted that these factors vary among the countries. Bohlin [9], on the other hand, identified new policy implications for the future European mobile market through an analysis of the success factors in the Japanese mobile Internet market.

As mobile carriers and content providers perform on a global scale [10], empirical cross-national research on mobile services has become increasingly relevant. A clear understanding of the mobile service needs of consumers can be achieved by investigating the structure of mobile services across different countries. In order to learn about consumer needs with respect to mobile information services, and whether consumers in different countries perceive these needs differently, we performed a comparison between the mobile information services needs structure of young people in Japan and the United States. These locations were selected for this international comparison because they are the two leading countries in the mobile market and because they use mobile services differently. Mobile users in Japan are the "most connected," with more than 75% using connected media (browsed the Internet, accessed applications, or downloaded content) compared to 43.7% in the United States and 38.5% in Europe [13]. At the end of June 2010, there were 111 million mobile subscribers in Japan [11], and there were 302.95 million subscribers in the United States at the end of December 2010 [12]. There is a need to identify a practical systematic framework of different structures in mobile information service needs in the United States and Japan by way of a cross-national comparison. If there are differences in service needs structures between two countries, mobile companies need to vary their international marketing strategies and tactics between the countries by adjusting for the differences. The following strategies can be used: introduce very high-spec devices, offer multiple technologies (picture messaging, mobile Flash, GPS, etc.), provide better network quality and coverage, etc. By understanding the differences in consumer needs with respect to mobile information services, mobile

 ISBN: 978-1-61208-164-9

companies will have a better chance of success. Therefore, the following hypotheses are presented:

H1. There is no change in the mobile phone service structure over a two-year period in Japan.

H2. The mobile service structure for information is same in both the United States and Japan.

H3. The mobile service structure for entertainment is same in both the United States and Japan.

H4. The mobile service structure for communication needs is same in both the United States and Japan.

This paper is composed as follows: In the following section, we cite relevant literature surveys to introduce various mobile services, and then in Section III, we identify the behaviors of mobile service users in the United States and Japan. In Section IV–VI, we describe how we measured information services in terms of the use intention data via an online survey. Section VII presents a factor analysis that is based on Japanese data and was carried out over a two-year period, and Section VIII presents a comparative factor analysis using data from Japan and the United States. The paper ends with a discussion of the results in terms of information, entertainment, and communication service factors, as well as the limitations of the study.

II. LITERATURE REVIEW

The use of mobile communication devices is increasing rapidly, and devices based on mobile technology are commonplace in everyday life [14]. A mobile information *service* is defined as use of the Internet via a handheld device [15]. The consumer pays for the desired mobile content or services [16]. Existing and potential services vary depending on the developments in mobile technology [9]. M-businesses offer more efficient markets and value system services, customized offering services, building community services, disrupting pricing services, and a radically extended reach in values services [17]. There are three main types of mobile services:

Information services: As a source of information, a mobile phone can have a significant impact on user behavior. Therefore, the quality of information provided by the device (e.g., maps or driving directions, restaurant guides, and promotional ads) often serves to expedite search efforts and stimulate the intention to use the mobile phone further. Users often log on to the Web to check e-mail, get news, obtain maps or driving directions, consult restaurant guides, etc. [9].

Entertainment services: Millions of people use their mobile devices for play [18]. Mobile phone usage in this context is an enjoyable activity that allows for an escape from reality. Users of these services may perceive mobile phone usage as more entertaining than informative [19].

Communication services: Short message service (SMS), multimedia message service (MMS), voice mail, and e-mail all fall under the category of communication services. These services may be classified as either utilitarian or hedonic, depending on the way they are used and the motivation behind their use [9].

III. MOBILE INFORMATION SERVICES IN THE UNITED STATES AND JAPAN

Most people today have a mobile phone, and many use them for mobile services beyond just calling and messaging. The mobile service market is growing rapidly. However, there are also many new service providers competing for customers, so it is very important to understand consumer usage behavior. This can be achieved easily by comparing user behavior on the basis of regions. The needs and uses of these services differ from country to country.

The United States is in its early stages of M-commerce development and adoption as compared to many European and Asian countries (e.g., Sweden, Japan, and Korea) [20]. This is true despite the fact that mobile application usage is slightly higher in the United States than that in Japan, as indicated in Table 1. In the United States, 19% fewer application users utilize their browser than in Japan, while 19% fewer browser users utilize applications. Messaging methods also vary. The United States displays the highest rate of text messaging, with 68.0% of users sending text messages compared with just 41.60% in Japan. Japanese users exhibit the highest reach in the e-mail category at

TABLE 1. MOBILE SERVICE USER BEHAVIOR IN JAPAN AND THE UNITED STATES IN OCTOBER–DECEMBER 2010: PERCENTAGE OF TOTAL MOBILE AUDIENCE (AGE: 13+ YEARS)

Countries	USA	Japan
Used connected media		
Browser, app or download	46.70%	76.80%
Used browser	36.40%	55.40%
Used application	34.40%	53.30%
Used messaging		
Sent text message	68.00%	41.60%
Instant messaging	17.20%	3.60%
Email	30.50%	57.10%
Accessed entertainment/social media		
Took photos	52.40%	62.90%
Social networking or blog	24.70%	19.30%
Played games	23.20%	16.30%
Recorded video	20.20%	15.80%
Listened to music	15.70%	12.90%
Watched TV and/or video	5.60%	22.80%
Accessed financial services		
Bank accounts	11.40%	7.00%
Financial news or stock	10.20%	16.50%
Accessed news, sports, weather, search, retail, travel, reference		
News and information	39.50%	57.60%
Weather reports	25.20%	34.70%
Search	21.40%	31.50%
Maps	17.80%	17.10%
Sports news	15.80%	18.20%
Restaurant info	10.00%	9.70%
Traffic reports	8.40%	14.00%
Classifieds	7.30%	3.60%
Retail site	6.50%	8.50%
Travel service	4.40%	2.90%

 ISBN: 978-1-61208-164-9

57.10%, while consumers in the United States are most likely to use text messaging services on their mobile devices (68.00%). Social networking/blogs reached the greatest percentage of mobile users in the United States at 24.70%, followed by Japan at 19.30%. Japanese users were most likely to capture photos (62.90%) and watch TV/video (22.80%) on their mobile devices, while users in the United States were most likely to listen to music (15.70%) and play games (23.20%) [10]. Table 1 displays mobile service user behavior in Japan and the United States [16].

Mobile phone users in certain countries, such as Japan, use integrated services, such as receiving messages regarding credit card usage, enjoying windows live messenger and other instant messenger systems, receiving messages from online community services, receiving promotional price discounts for family restaurants, and receiving coupons. Pioneers of location-based services—such as Korea and Japan—have created precise combinations of infrastructure and applications needed to ensure success [6].

IV. METHODOLOGY

We conducted two consecutive studies on Japanese customers in 2008 and 2009. In 2008, we conducted in-depth interviews with 30 mobile phone users who had adequate experience using mobile information services to explore their information needs and to identify the crucial factors that influence their mobile phone usage. On the basis of this qualitative study, we developed an instrument for survey research. We employed a professional market research firm in Japan to collect data under a random sampling framework from a panel of mobile information service users between the ages of 16 and 79. Data were collected online in the period between September 18 and September 24, 2008. A questionnaire focusing on the "use of information services via mobile phone" was distributed to a randomly selected Internet research panel with a sample size of 20000. From this sample, 5567 effective responses (27.8% of the total sample) were obtained. Out of these 5567 effective responses, the number of people who had mobile phones was 5222, which amounted to 93.8% of the effective responses. The following 21 services were examined: mobile e-mail, SMS, MMS, TV phone, radio, Internet, 1-seg TV (mobile terrestrial digital audio/video and data broadcasting service), music, ring tones, video streaming, games, learning (dictionary, translation services, and encyclopedia), health, infotainment content (movies, nightclubs, and celebrity gossip), mobile chat (push to talk), stock trading, shopping services, coupon and advertising information services, online storage services (Internet data storage services), reservation or booking (hotel rooms or airline seats), and location-based services (GPS or maps).

V. DATA COLLECTION

In 2009, an Internet research panel with a sample size of 3500 was randomly selected from the original 5222 respondents. Data were collected online during the period between July 10 and July 14. We obtained 1854 effective responses (53.0% of the sample).

For a two-year comparative study, a sample of 1854 effective responses from Japanese users was compiled. This sample consisted of the same people who responded in both 2008 and 2009; these respondents were designated as our analysis subjects.

In 2009, an Internet study was conducted with a sample size of 499 students at the University of California at Los Angeles (UCLA) in the United States. Out of the 499 respondents, 389 were in their 20's.

For a comparative study of young people in their 20's between Japan and the United States, we had an effective sample of 169 out of 1854 respondents from Japan and a sample of 389 students from UCLA. These respondents comprised our final set of sample data.

VI. FACTOR ANALYSIS BASED ON USE INTENTION DATA

A. Measurement

Aaker and Alvarez Del Blanco [21] have indicated that brand awareness indirectly affects purchase behavior. Likewise, an awareness of newly emerging services will affect purchase behavior and the intention to use these services. Therefore, this was a good place to begin extracting factors based on use intention. We conducted a factor analysis on 5222 respondents from the 2008 data and 1854 respondents from the 2009 data to extract common factors that exist among similar services in terms of the "use intention" of the 21 mobile information services. The following is a summary of the measurement:

The phrase, "Please rate your intention to use the following mobile information service" was used to operationalize use intention. A five-point Likert-type scale was anchored by low/high use intention for the 21 services.

VII. FACTOR ANALYSIS BASED ON JAPANESE USE INTENTION DATA OVER A TWO-YEAR PERIOD

In order to focus on the needs of the present and potential customers, we analyzed only "use intention." For this comparative two-year period, 1854 effective responses were designated as the analytical subjects. We conducted a factor analysis to extract common factors that existed among similar information services in terms of the "use intention" of the 21 mobile information services.

The factor analysis was conducted on the basis of the use intention by the principal factor method using varimax rotation. Kondo et al. [22] identified three dimensions of mobile services—information intensiveness, amusement, and service penetration rate—on the basis of the data of "awareness," "past use behavior," and "use intention." Here, we extracted four factors from the 21 information services, leading to the addition of one factor from the previous analysis. There was no change between 2008 and 2009 in the services affected by these factors except for "radio" (because of a missing value), which confirmed hypothesis H1.

Table 2 summarizes the factor loadings for each service

 ISBN: 978-1-61208-164-9

TABLE 2. FACTOR LOADINGS AND USER RATIO FOR EACH SERVICE IN 2008 AND 2009

Service Items	2008				2009				User
	1	2	3	4	1	2	3	4	Ratio
Reservation or booking	**.793**	.160	.136	.170	**.775**	.189	.211	.162	45.00%
Shopping services	**.739**	.291	.163	.098	**.712**	.302	.241	.076	47.90%
Coupon 'advertisement	**.620**	.312	.090	.243	**.602**	.350	.099	.228	59.80%
On-line storage services	**.617**	.272	.419	.088	**.551**	.258	.500	.084	24.40%
Health	**.613**	.398	.317	.054	**.534**	.394	.437	.072	39.50%
Learning	**.594**	.416	.179	.134	**.532**	.442	.243	.167	53.80%
Location based services	**.590**	.255	.154	.206	**.634**	.284	.205	.175	48.70%
internet	**.587**	.428	-.035	.229	**.574**	.408	.038	.205	84.50%
Infotainment content	**.560**	.498	.275	.108	**.539**	.472	.374	.150	50.90%
Stock trading	**.501**	.122	.343	-.024	**.542**	.099	.388	-.045	29.70%
Radio	**.417**	.330	.255	.130	.266	.350	**.397**	.140	37.50%
Ring tones	.198	**.743**	.136	.257	.210	**.724**	.170	.233	67.00%
Music	.375	**.714**	.117	.157	.297	**.748**	.205	.164	58.30%
Video streaming	.379	**.678**	.260	.179	.360	**.682**	.314	.190	50.80%
Games	.308	**.568**	.225	.119	.371	**.514**	.248	.125	57.20%
1 seg TV	.284	**.495**	.187	.197	.269	**.498**	.174	.174	46.50%
Mobile chat	.338	.296	**.754**	.070	.308	.267	**.774**	.039	17.70%
TV phone	.254	.356	**.479**	.297	.217	.311	**.506**	.279	35.40%
MMS	.092	.261	.158	**.719**	.095	.258	.159	**.700**	84.70%
Mobile email	.130	.144	-.147	**.615**	.143	.118	-.149	**.643**	97.10%
SMS	.085	.049	.139	**.449**	.062	.077	.193	**.451**	80.30%

(A) Year 2008: 1707 respondents; age group: 16~79
(B) Year 2009: 1686 respondents; age group: 16~79

in the case of each factor. The differences between the two years were due to a number of missing values. The results showed that in both 2008 and 2009, four factors were confirmed to be the primary factors affecting the mobile information service needs in Japan. These four factors were as follows:

Factor 1: information intensiveness;
Factor 2: entertainment;
Factor 3: low penetration service; and
Factor 4: communication service.

Factor 1 refers to services that require a high degree of information, such as making a reservation or stock trading. Factor 2 represents services with *entertainment* characteristics, such as ring tones. Factor 3 represents services with *low penetration* characteristics where the use ratio is low, such as a TV phone. Factor 4 represents services having *communication tool* characteristics, such as SMS, e-mail, and MMS, i.e., e-mail with pictures. Services within the factor are listed as follows:

Factor 1: radio, the Internet, learning, health, infotainment content, stock trading, shopping services, coupon and advertising information services, online storage services, reservation or booking, and location-based services;
Factor 2: 1-seg TV, music, ring tones, video streaming, and games;
Factor 3: TV phone and mobile chat;
Factor 4: mobile e-mail, SMS, and MMS.

VIII. COMPARATIVE FACTOR ANALYSIS BASED ON USE INTENTION DATA FROM JAPAN AND THE UNITED STATES

In order to focus on the needs of both present and potential customers, we analyzed only the "use intention." For this comparative research of young people in Japan (169) and the United States (389), effective responses were designated as the analytical subjects. We conducted a factor analysis to extract common factors that exist among similar information services in terms of the "use intention" of the 20 mobile information services, excluding 1-seg TV, which does not exist in the United States.

We used factor analysis as the statistical technique to analyze the data. We examined the data to check for inconsistencies due to random error by running a reliability test, ensuring that the integrity of the data was at a manageable level. Table 3 shows that the overall factor analysis was significant for Japan as the Kaiser-Meyer-Olkin statistics were greater than 0.50 and the chi square statistics were significant with a probability of less than 0.05 [23]. In the case of UCLA (Table 6), the overall factor analysis was not significant because there was a considerable amount of missing data from the Internet questionnaire, and hence, we ran the factor analysis without including the Internet data. Without the Internet data, the chi square statistics became significant, as shown in Table 8. We conducted a factor analysis on the 20 mobile information services by further excluding Internet data and extracted four factors that explained the 68.23% cumulative variance for Japan and three factors that explained the 57.6% cumulative variance for UCLA (eigenvalues greater than 1 are shown in Tables 4 and 9). The Cronbach α coefficient, the reliability coefficient of the measured value of questionnaire items for each construct from the point of view of internal consistency, is used for verifying whether each item had common parts for the same factor. If the

value of this coefficient was 0.7 or more, the internal consistency of the measurement scale was considered to be high and the reliabilities were adequate. The coefficients for each factor are shown in Tables 5 and 10. Since all values exceeded 0.7, it was concluded that the items of each information service of these factors had common parts.

We extracted four factors from 20 information services. There was no change in the first two factors between UCLA and Japan, except for some services belonging to the *communication* factor. In Table 10, the factor structure is presented on the basis of the identification of items that have loadings on the same factor, with a factor loading greater than 0.4. The service item Internet (for the United States) did not satisfy the abovementioned requirement and hence was omitted. For UCLA and Japan, the same items that significantly loaded on the first factor were reservations and booking, coupon advertisements, Internet storage services, shopping services, stock trading, learning, and location-based services. These six items represented the information services that customers could access by using a mobile device. Therefore, this factor was referred to as a mobile *information-intensive* service. Reservations and booking and coupon advertisements were very significant in the *information-intensive service* for both countries. The most information-intensive service items were loaded on the same factor; this implied that the service structure for an information-intensive service was same between the United States and Japan. This supported hypothesis H2.

The common items for the United States and Japan that were loaded as the second factor were music, games, ring tones, and video streaming. All of these items had an entertainment factor. These items indicated that customers prefer to be entertained by their mobile devices. Therefore, this factor could be named "*entertainment.*" Users in both countries were always satisfied by mobile entertainment services that enabled them to listen to music and download ring tones on their mobile phones. There existed an entertainment factor for both counties, and the entertainment service items were loaded on the same factor except for some items. This resulted in the rejection of hypothesis H3. We could interpret this as the existence of a same structure of the entertainment factor with slight differences in its members.

In Table 11, the order of the third and fourth factors was different for the United States and Japan. The identified items (for the United States and Japan) of the last two factors were fewer than those of the first two factors. Therefore, they were relatively old services and did not explain data variability well as compared to the first two factors. For the United States, the third factor consisted of two items, SMS and MMS, which facilitate basic communication; hence, this factor was named "*communication services.*" For Japan, factor 3 represented services with *low penetration* characteristics where the loading value was low. The third factor consisted of mobile chat and TV phones. These were classified as the "*low penetration factor,"* which specified advanced communication tools. The fourth factor was composed of MMS, mobile e-mail, and SMS, all of which facilitated basic communication. We found that the service structure for communication was different between the United States and Japan, which resulted in the rejection of hypothesis H4.

TABLE 3. KMO AND BARTLETT'S TEST FOR JAPAN

Kaiser-Meyer-Olkin measure of sampling adequacy		0.912
Bartlett's test of sphericity	Approx. Chi-square	2085.6
	Df	190
	Sig.	0.00

TABLE 4. KMO AND BARTLETT'S TEST FOR JAPAN

	Rotation sum of squared loadings		
factor	Total	% variance	Cumulative %
1	9.379	46.896	46.896
2	1.936	9.682	56.578
3	1.255	6.276	62.854
4	1.076	5.380	68.234

TABLE 5. RELIABILITY STATISTICS FOR INFORMATION SERVICE IN JAPAN

For Japan	Cronbach's α	No. of items
Information-intensive service	0.898	8
Entertainment	0.913	7
Low penetration service	0.774	2
Communication tools	0.654	3

TABLE 6. KMO AND BARTLETT'S TEST FOR UCLA (WITH INTERNET)

Kaiser-Meyer-Olkin measure of sampling adequacy		0.905
Bartlett's test of sphericity	Approx. Chi-square	60.443
	Df	190
	Sig.	**1.0**

TABLE 7. TOTAL VARIANCE EXPLAINED FOR UCLA (WITH INTERNET)

	Rotation sum of squared loadings		
Factor	Total	% variance	Cumulative %
1	7.123	35.616	35.616
2	2.529	12.645	48.261
3	1.330	6.650	54.911
4	1.034	5.168	60.079

 ISBN: 978-1-61208-164-9

TABLE 8. KMO AND BARTLETT'S TEST FOR UCLA (WITHOUT INTERNET)

Kaiser-Meyer-Olkin measure of sampling adequacy		0.906
Bartlett's test of sphericity	Approx. Chi-square	1246.164
	Df	171
	Sig.	0.00

TABLE 9. TOTAL VARIANCE EXPLAINED FOR UCLA (WITHOUT INTERNET)

	Rotation sum of squared loadings		
Factor	Total	% variance	Cumulative %
1	7.108	37.409	37.409
2	2.505	13.183	50.592
3	1.330	6.998	57.590

TABLE 10. RELIABILITY STATISTICS FOR UCLA INFORMATION SERVICE (WITHOUT INTERNET)

For UCLA	Cronbach's α	No. of items
Information-intensive service	0.930	10
Entertainment	0.883	7
Communication tools	0.642	2

TABLE 11. FACTOR LOADINGS OF EACH MOBILE SERVICE: COMPARISON BETWEEN THE UNITED STATES AND JAPAN

Rotated Factor Analysis							
Mobile service items	Japan(169): factor				UCLA(389): factor		
	1	2	3	4	1	2	3
Reservation and booking	**.813**	.169	.294	.083	**.823**	.133	.021
Location based services	**.704**	.290	.148	.122	**.391**	.366	.290
Shopping services	**.657**	.302	.276	.018	**.737**	.236	.044
Stock trading	**.620**	.168	.520	-.060	**.667**	.195	-.156
data storage services on Internet	**.586**	.316	.557	.068	**.761**	.126	-.036
Coupon· advertisement	**.556**	.312	.087	.113	**.796**	.041	.037
Internet	**.504**	.387	-.079	.288	N/A	N/A	N/A
Radio	**.440**	.252	.396	.126	.286	**.585**	-.041
Ring tones	.144	**.735**	.167	.217	.074	**.574**	.103
Video streaming	.318	**.694**	.284	.225	.353	**.659**	.076
Music	.423	**.661**	.164	.218	.134	**.723**	.185
Infotainment content	.508	**.596**	.373	.059	**.690**	.213	.022
Games	.366	**.536**	.207	.082	.139	**.699**	.137
Learning	.519	**.522**	.184	.067	**.566**	.329	.114
Health	.472	**.516**	.365	.021	**.710**	.273	-.034
Mobile chat	.324	.276	**.842**	.040	**.495**	.332	.063
TV phone	.209	.389	**.490**	.128	.345	**.512**	-.058
MMS (Text messaging)	.075	.219	.106	**.762**	-.020	.226	**.481**
Mobile email	.184	.041	-.317	**.605**	.159	**.543**	.330
SMS (Text messaging)	-.016	.093	.173	**.505**	-.057	.045	**.821**

A) USA (UCLA): 19 items; 389 respondents
B) Japan (National): 20 items; 169 respondents

For the large sample size of 1854 for 2008 and 2009 in Japan and the medium sample size of 389 for UCLA, health and learning were loaded in the first factor, which was an information-intensive factor. Because of the small sample size of 169 for Japan, the first factor and the second factor for the services of learning and health could not be easily differentiated. They were very close and had little influence on the entertainment factor. In Table 11, we concluded that a factor in TV phones was less assertive in Japan. Infotainment content was loaded in the entertainment factor for Japan and in the information-intensive factor for the United States.

IX. CULTURAL AND TECHNOLOGICAL DIFFERENCES IN MOBILE INFORMATION SERVICES BETWEEN JAPAN AND THE UNITED STATES

Differences in service needs between Japan and the United States were found in our research. From the previous section, we concluded that the intention to use mobile services for chat, mobile e-mail, and Internet access was considerably higher in Japan than in the United States. The reasons may be summarized as follows:

In the United States, a mobile phone is often viewed as a necessary tool rather than a luxury [24]. People in the United States are just as enthusiastic about mobilizing technology, but they often think in terms of shrinking and mobilizing the PC and the Internet, rather than expanding the mobile phone. Young people in the United States are much more likely to use SMS than e-mail. Sending an SMS was often considerably cheaper than sending an e-mail. The U.S. market has traditionally favored smart devices, such as the BlackBerry, which target business users as a path for potential growth.

On the other hand, mobile service sales in Japan have been consumer driven: people use their phones for e-mail, music downloads, games, and mobile-wallet services, in which financial transactions are carried out via the mobile phone [25]. Thus, Japan has developed a sophisticated mobile phone market earlier than the United States. Many Japanese people look to their mobile device as a central source of information gathering. This leads to a "Keitai (mobile phone) Culture" that is more obvious in Japan than in other countries, partially because of the Japanese people's affection for technology in general. The citizens of Japan are very technologically savvy, with considerable technological research, development, and manufacturing occurring in their country. Similarly, Japanese adults and teenagers rely on their mobile phones for communication and for the types of functions that a laptop or desktop computer would normally provide. With so many types of services and phones available, they may have one phone solely for the purpose of talking and another phone just for e-mail and accessing the Internet, or for other capabilities. The increase in texting via e-mail is the natural extension of the mobile phone culture and etiquette, which dictates the correct and appropriate usage of phones because Japanese people do not want to listen to other passengers chatting incessantly on their phones while they are riding a train home from work. As technology grows and develops, the mobile phone appears to

 ISBN: 978-1-61208-164-9

be at the forefront of both exponential growth and the evolution of culture. In this sense, the Japanese mobile market is years ahead of the U.S. market and is leading the way with respect to the mobile phone culture. The smart phone market in Japan expanded in 2010 with innovative and diverse formats such as personal/governmental/corporate communications. Our data dealt with mobile information service needs in 2009 instead of the actual usage. When we look at the rapid growth of smart phones in 2010, our analysis based on the 2009 needs data successfully predicted the potentiality of the mobile information service needs.

X. CONCLUSION

We identified four dimensions of need determinants for Japan and the United States: the *information-intensive* dimension, the *entertainment* dimension, the *communication* dimension, and *the service penetration rate (advanced communication)* dimension. Each factor was very closely related to the device generation, 1G or 2G (*communication*), 3G (*entertainment*), and 3G or 4G (*information intensive*). Therefore, we could conclude that our measurements were generally appropriate for extracting factors with respect to the need for mobile information services in technologically advanced countries. These results might be considered reliable largely because of the consistency of the sample questions.

The dimension of the *service penetration rate* was related to services with advanced technology, and there would not be many people who experienced using certain services. The dimension of *information-intensive* services had the largest variability. This might be due to the fact that this dimension was specific to customers who were interested in the specific services and would require appropriate segmentation identifying the relationship between the interests in the service and the characteristics of the customers.

From the comparative study, we found that the service structure for the information service was same and the service structure for the entertainment service was roughly same in the case of both the United States and Japan.

However, the scenario for service with communication factors appeared to be due to the cultural differences between the United States and Japan. Japan is more advanced than the United States in the use of communication tools. Daily life in Japan is not conceivable without an Internet connection. Mobile users in Japan were the "most connected," with more than 75% using connected media (browsed the Internet, accessed applications, or downloaded content) in June 2010 as compared to 43.7% in the United States. Japanese mobile users also displayed the highest usage of both applications and browsers, with 59.3% of the entire mobile population accessing their browsers in June 2010, and 42.3% accessing applications. Comparatively, 34.0% of the mobile users in the United States used their mobile browsers, and 31.1% accessed applications. The use of messaging methods also varied. The United States had the highest use of text messaging, with 66.8% sending a text message in June 2010 compared with just 40.1% in Japan. Japanese users exhibited the highest reach in the e-mail category at 54%, while consumers in the United States were most likely to use instant messaging services on their mobile phones (17.2%). Mobile operators in developed countries could begin to lose money in the next two to four years if they do not change their business models [26]. In this competitive mobile market, companies need to come up with innovative ideas and implement them around the world.

This study confirmed that mobile information services could be categorized into three types: information, entertainment, and communication. In our comparison of these services in Japan and the United States, communication displayed the largest difference.

XI. LIMITATION

Our research has some limitations with respect to the generalized ability of its findings. In 2009, an Internet study with a sample size of 499 was administrated to UCLA students, while the sample for Japan was randomly drawn. We could not exclude the impact of country-specific factors such as governmental legislation and other regulations. Controls on these effects could lead to cross-cultural studies. However, this might prove difficult because regulations were not enforced simultaneously in the considered countries, and the rate of development was not equal.

REFERENCES

[1] <http://www.idc.com/about/viewpressrelease.jsp?containerId=prUS22110509&sectionId=null&elementId=null&pageType=SYNOPSIS>12.03.2011.

[2] Barnes S. J. and Scornavacca E.: "Mobile marketing—The role of permission and acceptance," International Journal of Mobile Communication, vol. 2, no. 2: pp. 128-139, 2004.

[3] Massoud S. and Gupta O. K.: "Consumer Perception and Attitude toward Mobile Communication", International Journal of Mobile Communication, vol. 1, no. 1: pp. 89-118, 2003.

[4] Pedersen P. E.: "An adoption framework for mobile commerce," 1st IFIP conference of E-commerce, Minitrack on mobile commerce, Switzerland, 2001, available at <http://ikt.hia.no/perep/publications.htm>.

[5] Patricia H., Ruth R., and Cheung C. K.: "Adoption and usage of M-commerce: A cross-cultural comparison of Hong Kong and the United Kingdom," Journal of Electronic Research, vol. 6, no. 3: pp. 210-214, 2005.

[6] Yoon C. C.: "Assessing user attitudes toward mobile commerce in the US vs. Korea: Implications for M-commerce CRM," Journal of Business & Economic Research, vol. 6, no. 2: pp. 91-100, 2008.

[7] Yeonosoo L. and Inseong L.: "A cross cultural study on the value structure of mobile Internet usage: Comparison between Korea and Japan," Journal of Electronic Commerce Research, vol. 3, no. 4: pp. 227-235, 2002.

[8] Vrechopoulos A., Constantiou I., Sideris L., and Doukidis G.: "The critical role of consumer behaviour research in mobile commerce," International Journal of Mobile Communications, vol. 1, no. 3: pp. 329-340, 2003.

 ISBN: 978-1-61208-164-9

[9] Bohlin E., Björkdahl J., and Lindmark S.: "Strategies for making mobile communications work for Europe," Proceedings of the European Policy Research Conference (EuroCPR), Barcelona, Spain, 2003, available at <http://www.chalmers.se/tme/SV/organisation/personliga-sidor/bohlin-erik>.

[10] Haghirian P. and Madlberger M.: "Consumer attitude toward advertising via mobile devices—an empirical investigation among Austrian users," European Conference of Information Systems, 2005, Regensburg, Germany, pp. 1-2.

[11] <http://www.tca.or.jp/english/database/2010/02/index.html> 10.02.2011.

[12] <http://en.wikipedia.org/wiki/List_of_countries_by_number_of_mobile_phones_in_use> 08.02.2011.

[13] <http://mobithinking.com/mobile-marketing-tools/latest-mobile-stats> 18.07.2011.

[14] Sridhar Balasubraman, Robert A. Peterson, and Sirkka L. Jarvenpaa: "Exploring the implications of M-commerce for markets and marketing," Journal of the Academy of Marketing Science, vol. 30, no. 4: pp. 348-361, 2002.

[15] Cho C. H. and Cheon H. J.: "A cross cultural comparison of interactivity on corporate websites: US, UK, Japan & Korea," Journal of Advertising, vol. 34, pp. 99-115, 2005.

[16] Consumer Behavior Statistics of Mobile Telephone Services, F R I D A Å H S L U N D Master of Science Thesis, Stockholm, Sweden, 2006.

[17] Mohammed R. A., Fisher R. J., Jaworski B. J., and Cahill A. M.: "Internet marketing: Building advantage in a networked economy," McGraw-Hill Irwin, Marketspace U, 2002.

[18] Danet B., Wachenhauser T., Bechar-Israeli H., Cividalli A., and Rosenbaum-Tamari Y.: "Curtain time 20:00 GMT: Experiments with virtual theater on Internet relay chat," Journal of Computer-Mediated Communication (JCMC), vol. 2, no. 3, December 1996, http://jcmc.huji.ac.il/vol2/issue3/.

[19] Zhou Zheng: "Users' attitudes toward web advertising: Effects of Internet motivation and Internet ability," Advances in Consumer Research, vol. 29: pp. 71-78, 2002.

[20] Chew A. A.: "The adoption of M-commerce in the United States," Honors thesis, California State University, Long Beach, CA, pp. 2-22, 2006.

[21] Aaker and Alvarez Del Blanco: "Brand Leadership," London: Free Press, 2009.

[22] Kondo F., Hirata J., and Akter S.: "Exploring loyalty in mobile information services: The role of sound amusements," International Journal of Mobile Marketing, vol. 5, no. 1: pp. 125-140, 2010.

[23] Naresh K. Malhotra: Marketing Research, 4th edition, Pearson Education, 2004.

[24] <http://mobileopportunity.blogspot.com/2006/09/european-vs-american-mobile-phone-use.html> 17.07.2011.

[25] <https://wiki.smu.edu.sg/digitalmediaasia/Digital_Media_in_Japan>19.07.2011.

[26] <http://www.tellabs.com/news/2011/index.cfm/nr/142.cfm> 04.02.2011.

 ISBN: 978-1-61208-164-9

Cloud Systems and Their Applications for Mobile Devices

Jin-Hwan Jeong
Cloud Computing Division
Electronics and Telecommunication Research Institute
Daejeon, South Korea
jhjeong@etri.re.kr

Hak-Young Kim
Cloud Computing Division
Electronics and Telecommunication Research Institute
Daejeon, South Korea
h0kim@etri.re.kr

***Abstract*— In this paper, we discuss what mobile cloud is and what the differences between traditional cloud systems and mobile cloud systems are. At first, we point out mobile peculiarities such as battery constrains and computation weakness, and then we envision cloud system architecture for mobile devices that addresses these two issues. Specifically, we introduce mobile sensor virtualization and remote execution as essential components of mobile cloud system. At last, we show canonical services benefiting from our mobile cloud system.**

Keywords-Cloud computing; Mobile; Android; Sensor; Remote execution

I. INTRODUCTION

Today, computer science communities face very big two tides called cloud computing and mobile computing.

Since a few years, major software companies such as Amazon, Google, Microsoft have built large-scaled cloud systems and most SP (Service Provider) have transferred their services to the cloud systems. As cloud systems provide various software platforms from IaaS (Infrastructure as a Service) to SaaS (Service as a Service) and guarantees endless computing resources with scale up/out flexibility, SP can implement scalable services easily at the low cost. It is no doubt that SNS (Social Network Service) is one of live evidences.

Another big tide is mobile computing lead by smartphones. Smartphones, as a representative of mobile devices are very popular, and their hardware configuration is outstanding, for example, 1.2 GHz CPU, 1024 MB ram, 802.11g/n WIFI, 3G, and various sensor devices. These hardware components make recent smartphones possible to do everything that desktop can do. Especially, high-end smartphones (e.g., Apple, Android, and Window Phone 7) that have their unique application ECO system can do beyond desktop's capabilities on the behalf of various sensor devices.

In a sense, it might be natural to collaborating mobile devices with cloud system. As the performance of smartphone is superior, users want to use their smartphone as a client of new SNS as well as legacy services. Also, cloud system administrators begin to enhance the cloud systems with additional servers that help mobile devices. Typical examples of this efforts are PUSH (e.g., C2DM (Cloud To Device Messaging) from Google) servers and SYNC servers. Specifically, C2DM is a wonderful tool for SNS developers and it is good for network operators, because it prevents from wasting bandwidth for Keep Alive messages.

Nevertheless, both service developers and users found some limitations on services running in cloud system with smartphones. At first, users always keep attention for one resource, battery power, which is a most crucial resource. High performance smartphone can process complicated services, but it drains battery power much faster. Therefore, users cannot help recharging the battery, and then, this degrades the portability. Secondly, service developers have different difficulties. Most services including SNS require various sensor data in real-time such as geographical location or motion speed. This means that software part running on cloud system needs to query user's sensor data in real time. However, there is no uniform platform/library to aggregate and to deliver sensor data from various sensor devices of individual mobile devices for the server side system, so this is a big hurdle for SNS developers.

To address the first issue, Chun [7] proposed an augmented execution model. This approach has strength in alleviating computation loads of smartphones, but it has some constraints. In augmented execution, when application runs on server side, it is hard to synchronize intermediate data with mobile side. As mentioned in the previous paragraph, mobile services need various sensor data ceaselessly. However, augmented execution model does not provide a well-defined platform how server module can read mobile sensor data.

As a result, a cloud system for mobile devices should provide a tool that help work together easily and provide a way how to send mobile sensor data for rich mobile services. Therefore, we adopt augmented execution model as a cloud execution model and we propose sensor virtualization layer that gives transparent sensor data delivery. The latter can enhance the former's usability.

In this paper, we propose a virtual machine based-remote execution module with sensor device virtualization. With these tools, server module can co-work client modules with easy and utilizes client's sensor devices seamlessly. We start from building a virtual machine for Android, and implement remote execution mechanism for Java and also implement virtual sensor devices into Linux kernel as a form of kernel

 ISBN: 978-1-61208-164-9

modules. As virtual sensor devices are implemented in Linux, the devices can be used transparently by Android system.

II. BACKGROUND AND RELATED WORK

As illustrated in Fig. 1, the current mobile cloud system is much similar with typical cloud system except PUSH and SYNC servers. PUSH and SYNC servers are of "sugar" servers, because the servers enhance just mobile user experience much better and help other core servers. In a sense, two servers can be just regarded as a mobile gateway for better data exchange with cloud servers.

Now, the question is what the "mobile" does mean. Some cloud systems can be regarded as mobile cloud systems because their clients are mobile devices or because their deploying services are targeting for mobile devices or because they contains PUSH and SYNC servers as mentioned above.

Before defining the meanings of mobile, we point out the differences between mobile devices and desktop systems in the view point of clients (service target). About ten years ago, mobile devices were quite different from desktop. They had low performance processor, low memory, low quality displayer, so they was not able to run normal operating systems which meant that mobile applications were somewhat dedicated for mobile platforms. Therefore, the key philosophy for mobile applications was that the less mobile devices do, the better services are. Ten years change everything, and then, today, there is no quite difference between desktop and mobile devices such as iPhone or Android, and mobile devices can process most jobs that desktop can do in terms of performance.

However, there is one thing left, portability. The faster mobile processor is, the faster battery drains. In spite of brilliant material technology, the development speed of rechargeable battery technology is very slow, so everyone who has an iPhone or Android phone now is commonly worried about the shortage of battery power.

Also, a new issue comes out. Unlike the past, stock mobile devices have many sensor devices, and mobile services read values in real time. Most mobile platforms (e.g., iOS, Android, or Windows Phone) provide well defined API to handle sensor devices. However, the method for transferring sensor data from mobile devices to the cloud systems still is nothing changed. Developers just use TCP/IP communication with their own socket libraries.

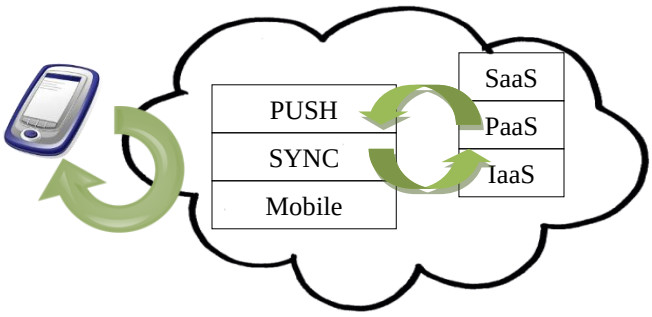


Figure 1. Typical mobile cloud system

There are some researches [6][7] for mobile issues. In [6], authors addressed battery lifetime with cloud computing and proposed off-loading computation for energy-saving. In [7], authors introduced clone computing to enhance smartphone capabilities. They [6][7] tried to catch two goals, computation limitation and energy limitation by Dalvik VM cloning (Augmented Execution) for Android. This work is remarkable, because augmented execution can leverage the degree of distribution between mobile device and cloud system on the same programming environment. However, they do not refer methods how to synchronize runtime execution environments (including run time data) with cloned Dalvik VM and real VM. Since sensor data have a tendency to change ceaselessly in mobile services, streaming sensor data is crucial. Nevertheless, augmented execution is worthy to evolve and we adopt it as a starting point of our remote execution.

III. MOBILE CLOUD SYSTEM ARCHITECTURE

Our system targets at two tools, remote execution and sensor device virtualization. Remote execution takes a responsibility for helping mobile device execution, which means that some sections of code are run on cloned VM instead of on mobile device, and sensor device virtualization that is used as sensor data sources for remote execution.

To simply remote execution model, we also restrict remote execution scope as following: 1) it has the same programming environment as mobile devices; 2) it does not share data between local modules and remote modules – stateless model; 3) only mobile side can initiates application. 4) only sensor data read from virtualized sensor devices are allowed for server. Like augmented execution in [7], our model is based on virtual machine.

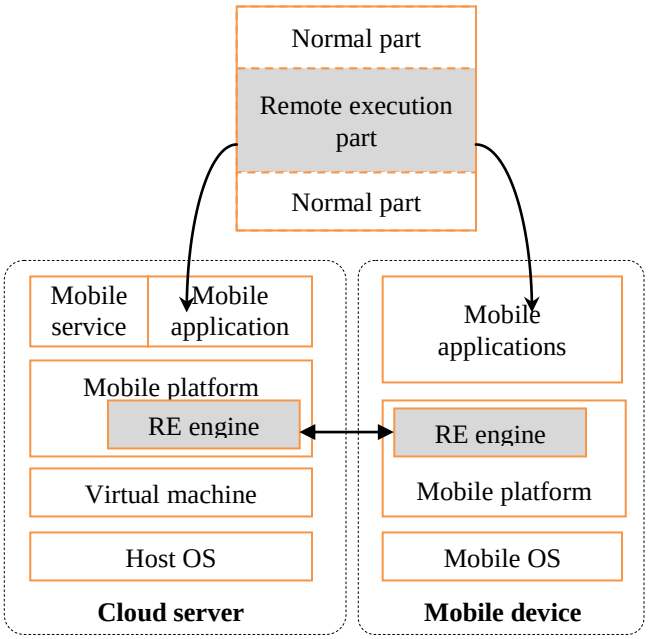


Figure 2. Mobile Cloud System

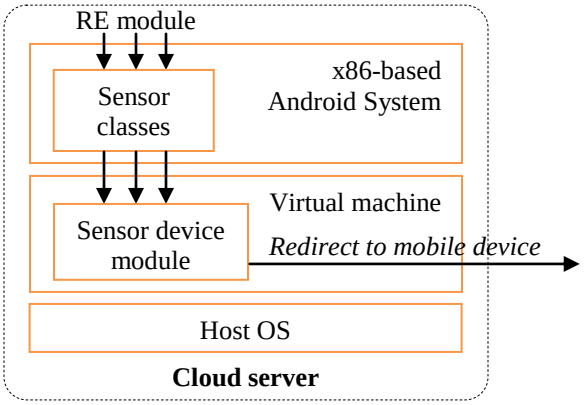


Figure 3. Sensor device virtualization

Fig. 2 illustrates overall architecture of our mobile cloud system. We adopted x86-based Android system for mobile virtual machine, and RE (Remote Execution) is implemented as Java library, which is similar with JAVA Remote Thread library. Like other remote thread libraries, RE is also based on a stateless model. Instead, our RE thread function takes one "Serializable" Java class as a function parameter.

In the cloud system side, RE modules are often in need of sensor data for mobile services. So far, developers should make a connection to read them, and they should endure various overheads such as checking device capabilities or network status. To alleviate such overheads, we introduce the sensor device virtualization that lets developers use client (remote) sensor devices locally. Fig. 3 shows the sensor device virtualization flow. As Android system is based on Linux, we hooked sensor devices I/O at kernel modules. This implies that applications (RE modules) can invoke sensor device I/O calls transparently. Specifically, when a request arrives from RE modules, virtual devices redirect a request to a peer mobile device. Although not presented in Fig. 3, mobile device also should have corresponding modules to process a redirected request. The corresponding modules receive a request and finally reply with live sensor value.

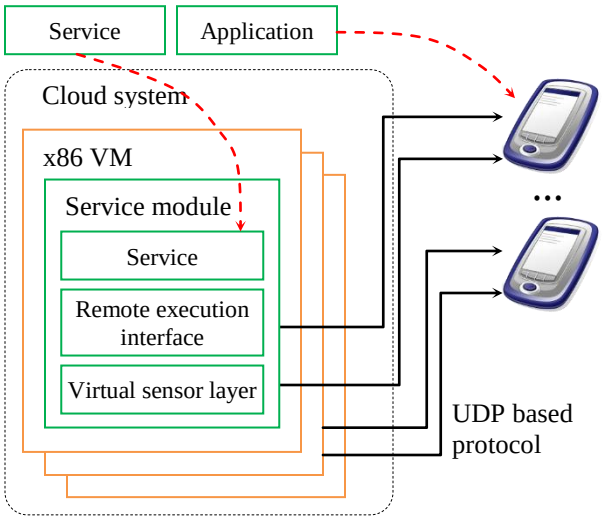


Figure 4. Overall cloud system for mobile devices

Our mobile cloud system with remote execution and sensor device virtualization is shown in Fig. 4. As one of the goals of our cloud system is to provide the same application environment as mobile devices to developers, our cloud supports mobile virtual machines. In case of Android phones, our cloud system allocates an Android x86 virtual machine per Android client for server part modules.

Specifically, for server module, service is linked with remote execution callee interface and sensor virtualization (requester) interface, and then it is initiated while VM creation implicitly (or explicitly by user's request). These interfaces eventually are connected on-the-fly with incoming mobile device using simple protocol. Similarly, mobile application has two modules, remote execution caller and sensor replier. Remote execution caller has various methods for instantiation / destruction / synchronization of remote thread object. Sensor replier simply replies sensor data requests from peer server.

Although our service execution model is based on the stateless model, it can process rich mobile services thanks to sensor device virtualization which provides live sensor data. In order to reduce response time and lightness, our model uses UDP based protocol with packet sequencing and limited ARQ features.

IV. APPLICATION

In this section, we show a typical example of services benefiting from our remote execution and sensor device virtualization. The example is about security. The scenarios are supposed that one security developer wants to verify both physical and logical security simultaneously. He/she needs sensor data for physical authentication and wants to run decoder for logical authentication on server. The main goal of this example is not to show a fact that security is enhanced due to our system. Instead, the goal is to show a fact that the developers can use mobile sensor data easily for physical authentication by sensor device virtualization just like they implement and run it on a local device.

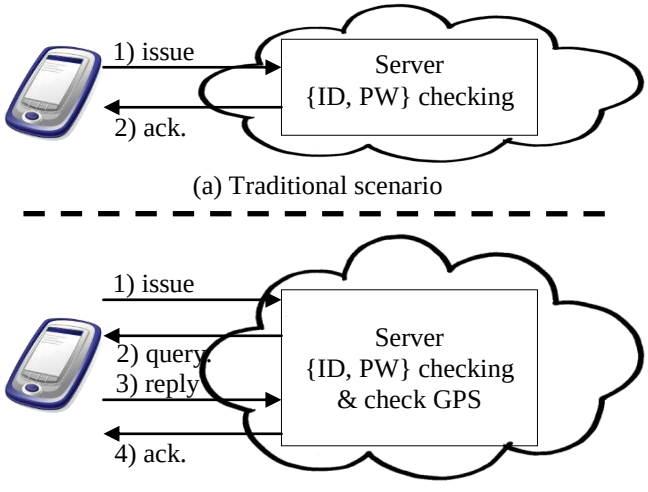


Figure 5. Two authentication scenarios

 ISBN: 978-1-61208-164-9

Mobile authentication with GPS: User authentication is one of eternal problems. Basically, if the tuple, {ID, Password} is correct, system regards an incoming user as being verified. However, suppose that a hacker got a password, and try to login. How can system detect this?

In the lower scenario (b) in Fig. 5, the developer of servers implemented verification module with additional features. Whenever the ID/Password arrives, server-side verification module queries GPS data without intervention of mobile user via local (virtualized) sensor devices and constructs location DB {user/location area}. One day, the user requests login, and verification module notices that his/her location is far from the usual location area. With only this fact, it is enough to suspect the user, and system can proceed to more robust verification step.

This scenario also can be implemented with traditional manner (a), but it is not simple. Developers should consider how to read GPS data and how to transfer to server. Also, developers should take care of upgrading of client part. However, if developers use our remote execution, they don't need TCP/IP to send ID/Password data, and if they use virtualized sensor devices, they don't need to know user's device model and how to access.

V. CONCLUSION

In this paper, we addressed two peculiarities called portability and sensor devices that make mobile cloud systems different from typical cloud systems, and we emphasize on the facts that mobile cloud system should take a care of two peculiarities. Especially, manipulating of various sensor data is considered as one of decision factors for whether the services are categorized as mobile services or not. For portability, we think the battery consumption. As a result, we propose remote execution with stateless model.

We are currently building Linux kernel supporting virtualized sensor devices and also implementing JAVA interfaces for remote execution. While we are still working on this project and are building prototypes, we take a look at the possibility that it can be indeed a good candidate of mobile cloud system.

REFERENCES

[1] http://www.android.com, April, 2011.

[2] http://developer.android.com/sdk/index.html, April, 2011.

[3] Bornstein, D. Dalvik virtual machine. http://www.dalvikvm.com, April, 2011.

[4] Kozuch, M. A., Ryan, M. P., Gass, R., Schlosser, S. W., O'Hallaron, D., Cipar, J., Krevat, E., López, J., Stroucken, M., and Ganger, G. R. "Tashi: location-aware cluster management," In Proceedings of the 1st Workshop on Automated Control For Datacenters and Clouds (Barcelona, Spain, June 19 - 19, 2009).

[5] Avetisyan, A.I., Campbell, R., Gupta, I., Heath, M.T., Ko, S.Y., Ganger, G.R., Kozuch, M.A., O'Hallaron, D., Kunze, M., Kwan, T.T., Lai, K., Lyons, M., Milojicic, D.S., Hing Yan Lee, Yeng Chai Soh, Ng Kwang Ming, Luke, J-Y., and Han Namgoong, "Open Cirrus: A Global Cloud Computing Testbed," in IEEE Computer, Vol. 43, Issue 2, pp. 35 ~ 43, April, 2010.

[6] Kumar, K., Y.-H. Lu, "Cloud Computing for Mobile Users: Can Offloading Computation Save Energy?," in Computer, Vol. 43, Issue 4, pp. 51~56, April 2010

[7] B.-G. Chun and P. Maniatis, "Augmented smartphone applications through clone cloud execution," in Proceedings of the 12th conference on Hot topics in operating systems (HotOS'09), pp. 8~12, 2009

 ISBN: 978-1-61208-164-9

New Scheduler with Call Admission Control (CAC) for IEEE 802.16 Fixed with Delay Bound Guarantee

Eden Ricardo Dosciatti
GETIC-NATEC-UTFPR
Federal University of Technology
Pato Branco - Parana - Brazil
Email: edenrd@utfpr.edu.br

Walter Godoy Jr.
NATEC-CPGEI-UTFPR
Federal University of Technology
Curitiba - Parana - Brazil
Email: godoy@utfpr.edu.br

Augusto Foronda
DAELN-NATEC-UTFPR
Federal University of Technology
Curitiba - Parana - Brazil
Email: foronda@utfpr.edu.br

***Abstract*—The IEEE 802.16 Working Group is developing a standard for broadband wireless access in Metropolitan Area Networks (MAN) known as WiMAX. One of the features of the MAC layer, in this standard, is that it is designed to provide differentiated servicing for traffic with multimedia requirements. Based on these assumptions, and considering that the standard does not specify a scheduling algorithm, a new scheduler with call admission control was proposed based on Latency-Rate (LR) server theory and with system characteristics as specified by the system standard using the WirelessMAN-OFDM (Orthogonal Frequency Division Multiplexing) air interface. The proposed scheduling algorithm calculates the time frame (TF) in order to maximize the number of stations allocated in the system while managing the delay required for each user. Properties of this proposal have been investigated theoretically and through simulations. A set of simulations is presented with both Constant Bit Rate (CBR) and Variable Bit Rate (VBR) traffic, and performance comparisons are made between cases with different delays and different TFs. The results show that an upper bound on the delay can be achieved for a large range of network loads, with bandwidth optimization.**

Keywords-IEEE 802.16; scheduling algorithm; delay bound; optimization; Call Admission Control (CAC).

I. INTRODUCTION

The deployment of high-speed Internet access is often cited as a challenge for the second decade of this century. Known as broadband Internet, it is effective in reducing physical barriers to the transmission of knowledge, as well as transaction costs, and is fundamental in fostering competitiveness. However, wired access to broadband Internet has a very high cost and is sometimes unfeasible, since the investment needed to deploy cabling throughout a region often outweigh the service provider's financial gains. One of the possible solutions in reducing the costs of deploying broadband access in areas where such infrastructure is not present is to use wireless technologies, which require no cabling and reduce both implementation time and cost [1].

This was one of the motivations behind the development by the IEEE (Institute of Electrical and Electronics Engineers) of a new standard for wireless access, called 802.16 [2], also known as Worldwide Interoperability for Microwave Access (WiMAX). It is an emerging technology for next generation wireless networks which supports a large number of users, both mobile and nomadic (fixed), distributed across a wide geographic area.

Motivated by the growing need for ubiquitous high-speed access, wireless technology is an option to provide a cost-effective solution that may be deployed quickly and easily, providing high bandwidth connectivity in the last mile. However, despite the many advantages of wireless access networks, such as low deployment and maintenance costs, ease of configuration and device mobility, there are challenges that must be overcome in order to further advance the widespread use of this type of network.

To achieve this purpose, the IEEE 802.16 standard introduces of a set of mechanisms, such as service classes and several coding and modulation schemes that adapt themselves according to channel conditions. However, the standard leaves open certain issues related to network resource management and scheduling algorithms.

This paper presents a new scheduler with admission control of connections to a WiMAX Base Station (BS). We developed an analytical model based on Latency-Rate (LR) server theory [3], from which an ideal frame size, called Time Frame (TF), was estimated, with guaranteed delays for each user. At the same time, the number of stations allocated in the system is maximized. In this procedure, framing overhead generated by the MAC (Medium Access Control) and PHY (Physical) layers was considered when calculating the duration of each time slot. After developing this model, a set of simulations is presented for constant bit rate (CBR) and variable bit rate (VBR) streams, with performance comparisons between situations with different delays and different TFs. The results show that an upper limit on the delay may be achieved for a wide range of network loads, thus optimizing bandwidth.

The remainder of this paper is organized as follows. In Section II, a brief description of the IEEE 802.16 standard is presented. Our analytical model of packet scheduling is proposed and explained in Section III. Evaluation of the

 ISBN: 978-1-61208-164-9

capacity of the new scheduler with Call Admission Control (CAC) is shown in Section IV. Conclusions are in Section V.

II. THE IEEE 802.16 STANDARD

A. Overview of Fixed WiMAX

The basic topology of a IEEE 802.16 network includes two entities that participate in the wireless link: Base Stations (BS) and Subscriber Stations (SS), as shown in Figure 1.

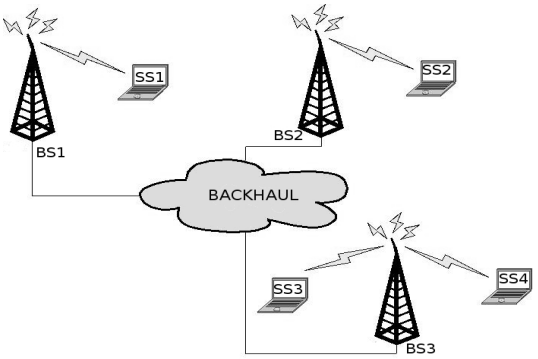


Figure 1. IEEE 802.16 Network Architecture

The BS is the central node, responsible for coordinating communication and providing connectivity to SSs. BSs are kept in towers distributed so as to optimize network coverage area, and are connected to each other by a backhaul network, which allows SSs to access external networks or exchange information between themselves.

Networks based on the IEEE 802.16 standard can be structured in two schemes. In PMP (Point-to-MultiPoint) networks, all communication between SSs and other SSs or external networks takes place through a central BS node. Thus, traffic flows only between SSs and the BS (see Figure 1). In Mesh mode, SSs communicate with each other without the need for intermediary nodes; that is, traffic can be routed directly through SSs. Thus, all stations are peers which can act as routers and forward packets to neighboring nodes. This article only considers the PMP topology.

The communication between a BS and SSs occurs in two different channels: uplink (UL) channel, which is directed from SSs to the BS, and downlink (DL) channel, which is directed from the BS to SSs. DL data is transmitted by broadcasting, while in UL access to the medium is multiplexed. UL and DL transmissions can be operated in different frequencies using Frequency Division Duplexing (FDD) mode or at different times using Time Division Duplexing (TDD) mode.

In TDD, the channel is segmented in fixed-size time slots. Each frame is divided into two subframes: a DL subframe and an UL subframe. The duration of each subframe is dynamically controlled by the BS; that is, although a frame has a fixed size, the fraction of it assigned to DL and UL is variable, which means that the bandwidth allocated for each of them is adaptive. Each subframe consists of a number of time slots, and thus both the SSs and the BS must be synchronized and transmit the data at predetermined intervals. The division of TDD frames between DL and UL is a system feature controlled by the MAC layer. Figure 2 shows the structure of a TDD frame. In this paper, the system was operated in TDD mode with the OFDM (Orthogonal Frequency Division Multiplexing) air interface, as determined by the standard.

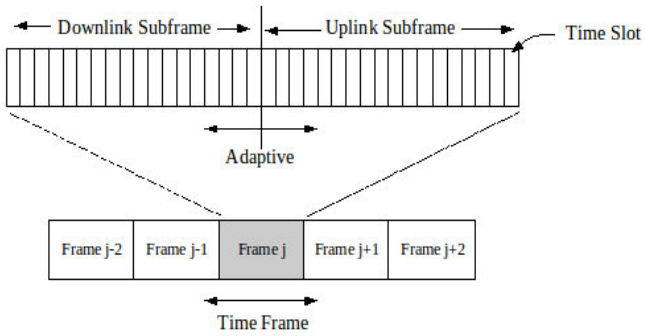


Figure 2. IEEE 802.16 Frame Structure

Figure 3 shows an example OFDM frame structure in TDD mode. As seen earlier, each frame has a DL subframe followed by a UL subframe. In this structure, the system supports frame-based transmission, in which variable frame lengths can be adopted. These subframes consists of a fixed number of OFDM symbols. Details of the OFDM symbol structure may be found in [1].

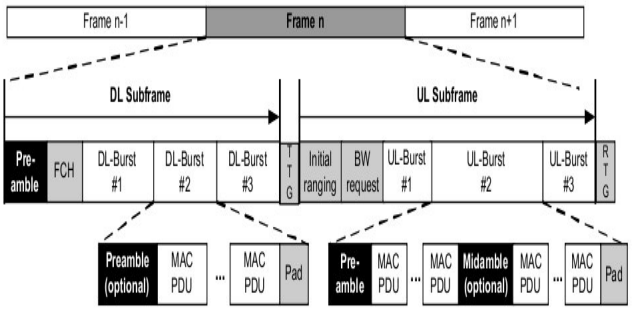


Figure 3. OFDM Frame Structure with TDD

The DL subframe starts with a long preamble (two OFDM symbols) through it the SSs can synchronize with the network and check the duration of the current frame. Instantly after DL long preamble, the BS transmits the Frame Control Header (FCH), which consists of an OFDM symbol and is used by SSs to decode the MAC control messages transmitted by BS.

 ISBN: 978-1-61208-164-9

The UL subframe consists in contention intervals for initial ranging and bandwidth request purposes and one or several UL transmission bursts, each from a different SSs. The initial ranging slots allow an SS to enter the system, by adjusting its power level and frequency offsets and by correcting its timing offset. Bandwidth request slots are used by SSs to transmit bandwidth request headers.

Two gaps separate the DL and UL subframes: the Transmit/Receive Transition Gap (TTG) and Receive/Transmit Transition Gap (RTG). These gaps allow the BS to switch from the transmit to receive mode, and vice versa.

B. Related Research

Since the standard only provides signaling mechanisms and no specific scheduling and admission control algorithms, some scheduling algorithms have been proposed to provide QoS (Quality of Service) for WiMAX. However, many of these solutions only address the implementation or addition of a new QoS architecture to the IEEE 802.16 standard. A scheduling algorithm decides the next packet to be served on the waiting list and is one of the mechanisms responsible for distributing bandwidth among several streams.

In [5], a packet scheduler for IEEE 802.16 uplink channels based on an hierarchical queue structure was proposed. A simulation model was developed to evaluate the performance of the proposed scheduler. However, despite presenting simulation results, the authors overlooked the fact that the complexity of implementing this solution is not hierarchical, and did not define clearly how requests for bandwidth are made. In [7], authors proposed a QoS architecture to be built into the IEEE 802.16 MAC sublayer, which significantly impacts system performance, but did not present an algorithm that makes efficient use of bandwidth. In [8], authors presented a simulation study of the IEEE 802.16 MAC protocol operating with an OFDM (Orthogonal Frequency Division Multiplexing) air interface and full-duplex stations. They evaluated system performance under different traffic scenarios, varying the values of a set of relevant system parameters. Regarding data traffic, it was observed that the overhead due to the physical transmission of preambles increases with the number of stations. In [9], a polling-based MAC protocol is presented along with an analytical model to evaluate its performance. They developed closed-form analytical expressions for cases in which stations are polled at the beginning or at the end of uplink subframes. It is not possible to know how the model may be developed for delay guarantees. Finally, in [10], the author presents a well-established architecture for QoS in the IEEE 802.16 MAC layer. The subject of this work is the component responsible for allocating uplink bandwidth to each SS, although the decision is taken based on the following aspects: bandwidth required by each SS for uplink data transmission, periodic bandwidth needs for UGS flows in SSs and bandwidth required for making requests for additional bandwidth.

Considering the limitations exposed above, these works form the basis of a generic architecture, which can be extended and specialized. However, in these studies, the focus is in achieving QoS guarantees, with no concerns for maximizing the number of allocated users in the network. This paper presents a scheduler with admission control of connections to the WiMAX BS. We developed an analytical model based on Latency-Rate (LR) server theory [3], from which an ideal frame size called Time Frame (TF) was estimated, with guaranteed delays for each user and maximization of the number of allocated stations in the system. A set of simulations is presented with CBR and VBR streams and performance comparisons are made for different delays and different TFs. The results show that an upper bound on the delay may be achieved for a large range of network loads with bandwidth optimization.

III. Analysis of the Analytical Model

A. System Description

Figure 4 illustrates a wireless network operating the newly proposed scheduler with call admission control, which is based on a modified LR scheduler [3] and uses the token bucket algorithm. The basic approach consists on the token bucket limiting input traffic and the LR scheduler providing rate allocation for each user. Then, if the rate allocated by the LR scheduler is larger than the token bucket rate, the maximum delay may be calculated.

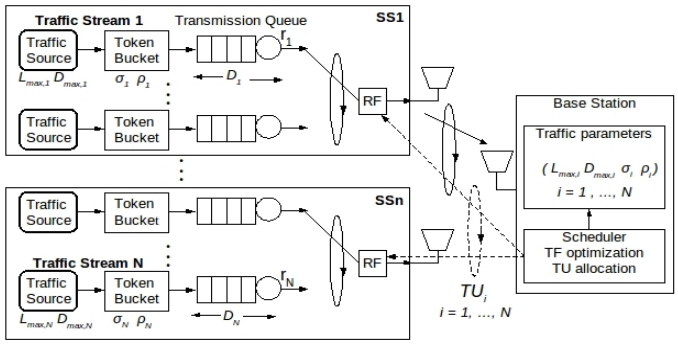


Figure 4. Wireless Network with New Scheduler

The behavior of an LR scheduler is determined by two parameters for each session i: latency θ_i and allocated rate r_i. The latency θ_i of the scheduler may be seen as the worst-case delay and depends on network resource allocation parameters. In the new scheduler with call admission control, the latency θ_i is a TF period, which is the time needed to transmit a maximum-size packet and separation gaps (TTG and RTG) of DL and UL subframes. In the new scheduler, considering the delay for transmitting the first packet, the latency θ_i of is given by

$$\theta_i = T_{TTG} + T_{RTG} + T_{DL} + T_{UL} + \frac{L_{max,i}}{R} \quad (1)$$

where T_{TTG} and T_{RTG} are DL and UL subframes gaps durations, T_{DL} and T_{UL} are the DL and UL subframes duration, $L_{max,i}$ is the maximum packet size and R is the outgoing link capacity.

Now, we show how the allocated rate r_i for each session i may be determined, and how to optimize TF in order to increase the number of connections accommodated with Call Admission Control (CAC).

B. CAC Description

An LR scheduler can provide a bounded delay if the input traffic is shaped by a token bucket. A token bucket [1] is a non-negative counter which accumulates tokens at a constant rate ρ_i until the counter reaches its capacity σ_i. Packets from session i can be released into the queue only after removing the required number of tokens from the token bucket. In an LR scheduler, if the token bucket is empty, arriving packets are dropped; however, our model ensures that there will always be tokens in the bucket and that no packets are dropped, as described in Section IV. If the token bucket is full, a maximum burst of σ_i packets can be sent to the queue. When the flow is idle or running at a lower rate as the token size reaches the upper bound σ_i, accumulation of the tokens will be suspended until the arrival of the next packet. We assume that the session starts out with a full bucket of tokens. In our model, we consider IEEE 802.16 standard overhead for each packet. Then, as we will show below, the token bucket size will decreased by both packet size and overhead.

The application using session i declares the maximum packet size $L_{max,i}$ and required maximum allowable delay $D_{max,i}$, which are used by the WiMAX scheduler to calculate the service rate for each session so as to guarantee required delay and optimize the number of stations in the network. Incoming traffic passes through a token bucket inside the user terminal during an interval, as shown in Figure 5.

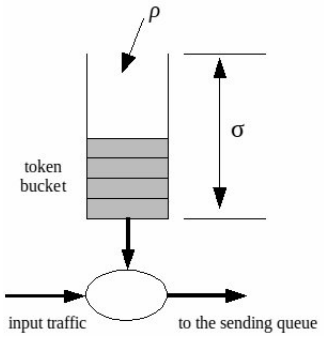


Figure 5. Input Traffic with Token Bucket

This passage of data traffic by the token bucket is bounded by

$$A_i(t) \leq \sigma_i + \rho_i t \quad (2)$$

where σ_i is the bucket size and ρ_i is the bucket rate.

Then, the packet is queued in the station until it is transmitted via the wireless. Queue delay is measured as the time interval between the receipt of the last bit of a packet and its transmission. In the new scheduler with call admission control, queuing delay depends on token bucket parameters, network latency and allocated rate. In [3], it is shown that if input traffic is shaped by a token bucket and the scheduler allocates a service rate r_i , then an LR scheduler can provide a bounded maximum delay Di:

$$D_i \leq \frac{\sigma_i}{r_i} + \theta_i - \frac{L_{max,i}}{r_i} \quad (3)$$

where r_i is the service rate, σ_i is the token bucket size, θ_i is the scheduler latency, $L_{max,i}$ is the maximum size of a package. $\frac{\sigma_i}{r_i} + \theta_i - \frac{L_{max,i}}{r_i}$ is the bound on the delay D_{bound}.

Equation (3) is an improved bound delay for LR schedulers. Thus, the token bucket rate plus the overhead transmission rate must be smaller than the service rate to provide a bound on the delay. The upper bound delay D_{bound} should be smaller or equal to the maximum allowable delay:

$$\frac{\sigma_i}{r_i} + \theta_i - \frac{L_{max,i}}{r_i} \leq D_{max,i} \quad (4)$$

Therefore, three different delays are defined. The first is the maximum delay D_i, the second is the upper bound on the delay D_{bound} and the third is the required maximum allowable delay $D_{max,i}$. The relation between them is $D_i \leq D_{bound} \leq D_{max,i}$.

So, the delay constraint condition of the new scheduler is

$$\frac{(\sigma'_i - L'_{max,i})TF}{r'_i TF - \Delta R + L'_{max,i}} + TF + \frac{L'_{max,i}}{R} + T_{TTG} + T_{RTG} \leq D_{max,i} \quad (5)$$

where σ'_i is the token bucket size with overhead, $L'_{max,i}$ is the maximum size of a packet with overhead (preamble+pad), TF is the time frame, r'_i is the rate allocated by the server with overhead, R is the outgoing link capacity, T_{TTG} is the gap between downlink and uplink subframes, T_{RTG} is the gap to between uplink and downlink subframes, $D_{max,i}$ is the maximum allowable delay and Δ is the sum of initial ranging and BW request, which is the uplink subframe overhead. Physical rate, maximum packet size and token bucket size are parameters declared by the application. However, TF and total allocated service rate must satisfy Equation (5).

Figure 6 shows a frame structure with TDD allocation formulas as described by Equation (5). Physical rate, maximum packet size and token bucket size are parameters

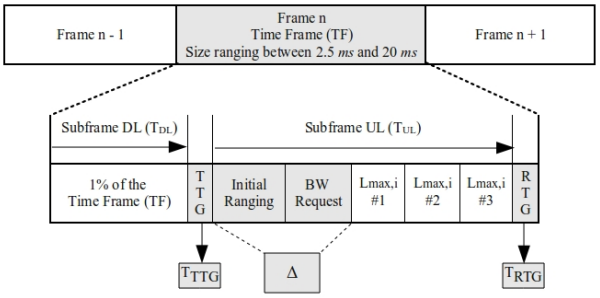


Figure 6. Frame structure with TDD allocation formulas of Equation (5)

declared by the application. However, TF and total allocated service rate must satisfy Equation (5).

Equation (6) is the second constraint condition to TF and service rate. Token bucket rate plus the rate to transmit overhead and a maximum-size packet must be smaller than the service rate to place a bound on delay. Thus, the second constraint condition is

$$\rho_i + \frac{\Delta R + L'_{max,i}}{TF} \leq r'_i \qquad (6)$$

where ρ_i is the bucket rate, Δ is the uplink subframe overhead, R is the outgoing link capacity, $L'_{max,i}$ is the maximum packet size with overhead, TF is the time frame and r'_i is the rate allocated by the service with overhead.

Previous schedulers do not provide any mechanism to estimate the TF needed to place a bound on delay or to maximize the number of stations, because each application requires a TF without the use of criteria to calculate the time assigned to each user. TF estimation is important because a small TF reduces maximum delay, but increases overhead at the same time. On the other hand, a large TF decreases overhead, but increases delay. Therefore, we must calculate the optimal TF to allocate the maximum number of users under these both constraint. The maximum number of users is achieved when the service rate for each user is the minimum needed to guarantee the bound on the delay D_{bound}. Different optimization techniques may be used to solve this problem. In this study, we have used a step-by-step approach, which does not change the scheduler's essential operation. We start with a small TF, for example, $2.5ms$, calculate r'_i and repeat this process every $0.5ms$ until we find the minimum r'_i that satisfies both equations.

IV. Performance Analysis

To analyze the IEEE 802.16 MAC protocol behavior with respect to the new scheduler with call admission control, this section presents numerical results obtained with the analytical model proposed in the previous section. Then, with a simulation tool, the proposed analytical model is validated by showing that the bound on the maximum delay is guaranteed. In this section, two types of delays are treated: required delay, in which the user requires the maximum delay, and the guaranteed maximum delay, which is calculated with the analytical model.

A. Calculation of Optimal Time Frame

All PHY and MAC layer parameters used in simulation are summarized in Table I.

Table I
PHY and MAC parameters

PARAMETER	VALUE
Bandwidth	20MHz
OFDM Symbol Duration	13,89 μs
Delay	5 / 10 / 15 and 20 ms
Δ (Initial Ranging and BW Request) $\rightarrow$ 9 OFDM Symbols	125,10 μs
TTG + RTG $\rightarrow$ 1 OFDM Symbol	13,89 μs
UL Subframe (preamble + pad) $\rightarrow$ 10% OFDM Symbol	1,39 μs
Physical Rate	70 Mbps
DL Subframe	1% TF

Performance of the new scheduler with call admission control is evaluated as the delay requested by the user and assigned stations. Station allocation results, in the system with an optimal TF, limited by the delay requested by the user, are described in sequence. The first step is define token bucket parameters, which are estimated in accordance with the characteristics of incoming traffic and are listed on Table II.

Table II
Token bucket parameters

	Audio	VBR video	MPEG4 video
Token Size (bits)	3000	18000	10000
Token Rate (kb/s)	64	500	4100

Thus, the optimal TF value is estimated according to the PHY and MAC layer's parameters (see Table I), token bucket parameters (see Table II), required maximum allowable delay, physical rate and maximum package size.

The graph in Figure 7 shows the optimal TF value, for four delay values required by users (5, 10, 15 and 20 ms).

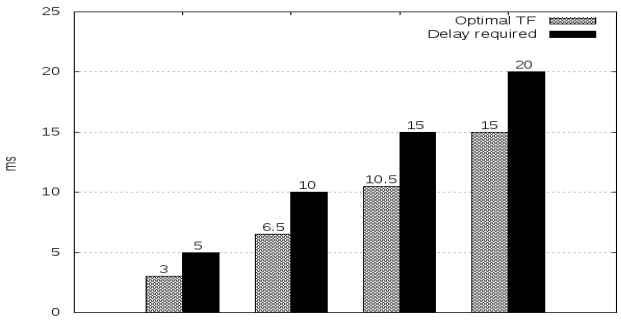


Figure 7. Optimal TF

Next, we show the number of SSs assigned to each traffic type. As an example, Figure 8 show that when the user-requested delay is of 20 ms, an optimal TF of 15 ms is calculated and 50 users can be allocated for audio traffic, or 30 users for VBR video traffic, or 13 users for the MPEG4 video traffic.

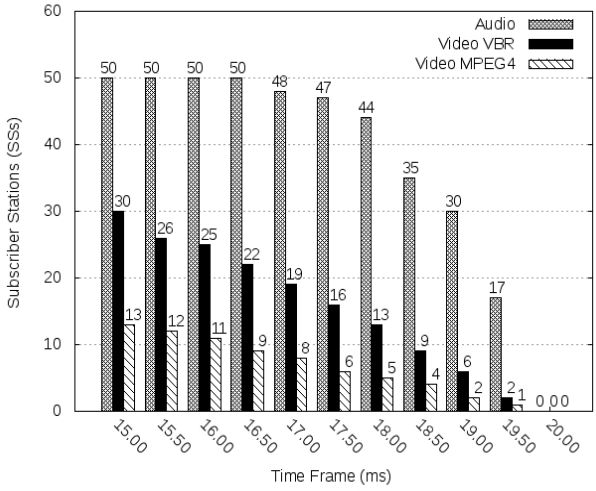


Figure 8. Number of subscriber stations for 20 ms of delay

Two important observations from Figure 8 should be highlighted:

1) With a requested delay of 20 ms, we cannot choose a TF of less than 15 ms, since the restrictions placed by Equation (5) (which regards delay) and Equation (6) (which regards the token bucket) are not respected and thus no bandwidth allocation guarantees exist.
2) We also cannot choose a TF greater than 15 ms, even though it complies with Equations (5) and (6) with respect to guaranteed bandwidth, because there will be a decrease in the number of users allocated to each traffic flow due to increase of the delay.

The same philosophy holds true for other delay values of 5, 10 and 15 ms.

B. Guaranteed Maximum Delay

In this article, only UL traffic is considered. To test the new scheduler's performance, we have carried out simulations of an IEEE 802.16 network consisting of a BS that communicates with eighteen SSs, with one traffic flow type by SS and the destination of all flows being the BS. In this topology, six SSs transmit on-off CBR audio traffic (64 kb/s), six transmit CBR MPEG4 video traffic (3.2 Mb/s) and six transmit VBR video traffic. Table III summarizes the different types of traffic.

On Figure 9, with an optimal TF of 3 ms and an user-requested delay of 5 ms, the average guaranteed maximum delay for audio traffic is 1.50 ms. For VBR video traffic, whose packet rate is variable, the average maximum delay is 1.97 ms. For MPEG4 video traffic, the average maximum delay is 2.00 ms.

Table III
Description of the different traffics

Node	Application	Arrival Period (ms)	Packet size (max) (bytes)	Sending rate (kb/s) (mean)
$1 \rightarrow 6$	Audio	4.7	160	64
$7 \rightarrow 12$	VBR video	26	1024	≈ 200
$13 \rightarrow 18$	MPEG4 video	2	800	3200

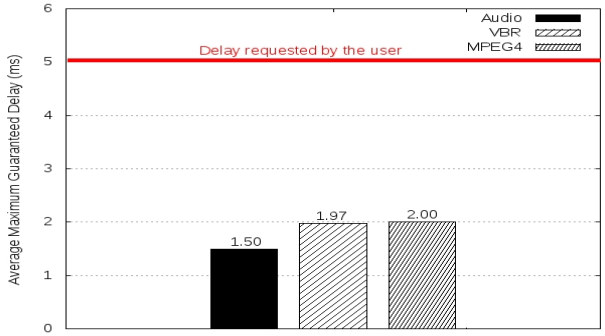


Figure 9. Maximum Guaranteed Delay

C. Comparison with other Schedulers

The new scheduler with call admission control, here called *New Scheduler*, was compared to those of [9], here called $Scheduler\_1$, and [5], here called $Scheduler\_2$. The comparison was accomplished through the ability to allocate users in a particular time frame (TF). Table IV shows the parameters used for comparisons.

Table IV
Parameters used for comparisons

PARAMETER	$Scheduler\_1$	$Scheduler\_2$
Bandwidth	20 MHz	20 MHz
OFDM symbol duration	13.89 μs	13.89 μs
Time Frame (TF)	5 ms	10 ms
Delay Requested by the user	0.12 ms	20 ms
Maximum Data Rate	70 $Mbps$	70 $Mbps$
Traffic type	Audio	Audio

In the graph of Figure 10, we compare the *New Scheduler* with the $Scheduler\_1$. A maximum delay of 0.12 ms was requested by the user, and the duration of each frame (TF) was set at 5 ms. Other parameters are listed in Table IV. In comparison, the *New Scheduler* allocates 28 users in each frame, while the $Scheduler\_1$, allocates 20 users. Thus, the *New Scheduler* presents a gain in performance of 40% when compared with the $Scheduler\_1$.

In the graph of Figure 11, we compare the *New Scheduler* with the $Scheduler\_2$. A maximum delay of 20 ms was requested by the user, and the duration of each frame (TF) was set at 10 ms. Other parameters are listed in Table IV.

 ISBN: 978-1-61208-164-9

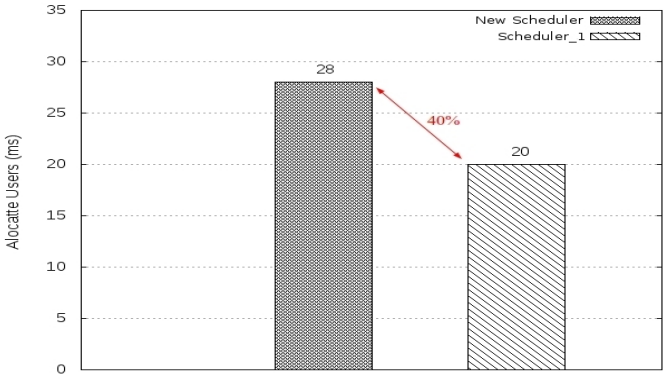


Figure 10. Comparison of allocation of users with $Scheduler\_1$

The comparison was extended by also considering frame duration values of 7.00 ms, 8.00 ms and 9.00 ms to demonstrate the efficiency of the *New Scheduler*. For a TF of 10 ms, the *New Scheduler* allocates 41 users in each frame, while the $Scheduler\_2$ allocates only 33 users. This represents 24.24% better performance for the *New Scheduler*. Similarly, the *New Scheduler* also allocates more users per frame in comparison with the $Scheduler\_2$ for all other frame duration values.

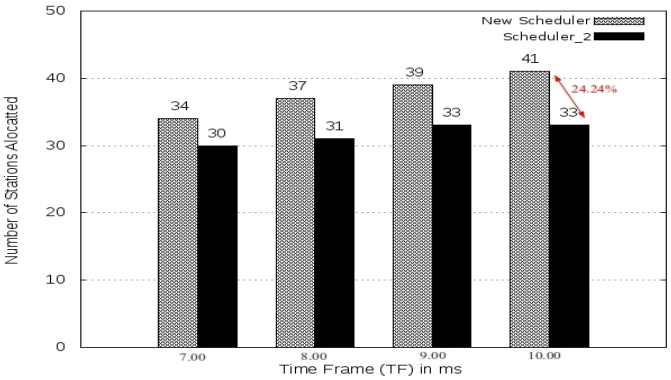


Figure 11. Comparison of allocation of users with $Scheduler\_2$

V. Conclusion

This work has presented the design and evaluation of a new scheduler with call admission control for IEEE 802.16 fixed networks, that guarantees different maximum delays for traffic types with different QoS requisites and optimizes bandwidth usage. Firstly, we developed an analytical model to calculate an optimal TF, which allows an optimal number of SSs to be allocated and guarantees the maximum delay required by the user. Then, a simulator was developed to analyze the behavior of the proposed system.

To validate the model, we have presented the main results obtained from the analysis of different scenarios. Simulations were performed to evaluate the performance of this model, demonstrating that an optimal TF was obtained along with a guaranteed maximum delay, according to the delay requested by the user. Thus, the results have shown that the new scheduler with call admission control successfully limits the maximum delay and maximizes the number of SSs in a simulated environment.

Acknowledgment

We thank all the researchers at the Advanced Nucleous of Communication Technology at UTFPR.

References

[1] A. Gosh, D. Wolter, J. Andrews, and R. Chen, "Broadband wireless access with WiMAX/802.16: current performance benchmarks and future potential," In IEEE Communications, v. 43(2), Feb. 2005, pp. 129-136, doi:10.1109/MCOM.2005.1391513.

[2] IEEE 802.16-2004, "IEEE Standard for Local and Metropolitan Area Networks - Part 16: Air Interface for Fixed Broadband Wireless Access Systems," IEEE Std., Rev. IEEE Std802.16-2004, Oct. 2004.

[3] D. Stiliadis, and A. Varma, "Latency-Rate Servers: A General Model for Analysis of Traffic Scheduling Algorithms," In IEEE-ACM Transactions on Networkink, v. 6, Oct. 1998, pp. 611-624, doi:10.1109/90.731196.

[4] E. R. Dosciatti, W. Godoy Jr., and A. Foronda, "A New Scheduler for IEEE 802.16 with Delay Bound Guarantee," The Sixth International Conference on Networking and Services (ICNS 2010), Cancun, Mexico, v. 1, Mar. 2010, pp. 150-155, doi:10.1109/ICNS.2010.27.

[5] K. Wongthavarawant, and A. Ganz, "Packet Scheduling for QoS Support in IEEE 802.16 Broadband Wireless Access Systems," In Internacional Journal of Communications Systems, v. 16, Feb. 2003, pp. 81-96, doi:10.1002/dac.581.

[6] C. Hoymann, "Analysis and performance evaluation of the OFDM-based metropolitan area network IEEE 802.16," In Computer Networks, v. 49, Oct. 2005, pp. 341-363, doi:10.1016/j.comnet.2005.05.008.

[7] G. Chu, D. Wang, and S. Mei, "A QoS architecture for the MAC protocol of IEEE 802.16 BWA system," In IEEE Conference on Communications, Circuits, and Systems, v. 1, Jun./Jul. 2002, pp. 435–439, doi:10.1109/ICCCAS.2002.1180654.

[8] C. Cicconetti, A. Erta, L. Lenzini, and E. Mingozzi, "Performance Evaluation of the IEEE 802.16 MAC for QoS Support," IEEE Transactions on Mobile Computing - TMC07, v. 6, Jan. 2007, pp. 26–38, doi:10.1109/TMC.2007.250669.

[9] R. Iyengar, P. Iyer, and B. Sikdar, "Delay Analysis of 802.16 Based Last Mile Wireless Networks," Global Telecommunications Conference - GLOBECOM'05 - IEEE, v. 5, Dec. 2005, pp 1-5, doi:10.1109/GLOCOM.2005.1578332.

[10] S. Maheshwari, "An Efficient QoS Scheduling Architecture for IEEE 802.16 Wireless MANs," Master Degree, K R School of Information Technology, Bombay, India, Jan. 2005.

A Hardware Architecture for MAP Decoding Based on Nibble Alignment

Seungkwon Cho, Sok-Kyu Lee
Short Range Radio Transmission Research Team
Electronics and Telecommunication Research Institute
Daejeon, Korea
{skcho, sk-lee}@etri.re.kr

Youngnam Han
Department of Electrical Engineering
Korea Advanced Institute of Science and Technology
Daejeon, Korea
ynhan@kaist.ac.kr

Abstract—**In IEEE 802.16-2009 standard, a Subscriber Station (SS) has to meet timing constraints imposed by uplink and downlink MAP relevance. Thus, the standard adopts nibble alignment in MAP Information Elements (IEs) in order to support fast MAP decoding. However, some MAP IEs are not nibble-aligned with increased implementation complexity of a MAP decoder. In this paper, we present a hardware architecture designed to efficiently decode MAP message on a nibble basis while suppressing the increase in implementation complexity. The feasibility of the proposed architecture is verified by FPGA synthesis and its performance is presented in terms of MAP decoding time.**

Keywords-IEEE 802.16-2009; MAP relevance; nibble alignment; MAP decoding

I. INTRODUCTION

Recently, the IEEE 802.16 Working Group on Broadband Wireless Access (BWA) Standards released IEEE 802.16-2009 [1] as a new base standard that supersedes and makes obsolete IEEE standard 802.16-2004 and all of its subsequent amendments and corrigenda. The new standard is expected to provide high data rate communications in metropolitan area networks (MANs). In order to support such a high speed transmission, careful considerations should be given to the implementation architecture of both physical (PHY) layer and medium access control (MAC) layer.

From the traditional hardware and software partitioning methodologies point of view, most of MAC functions are implemented by software because it enables a system to have more flexibility and reduced time-to-market. However, as transmission speed increases, more and more timing critical MAC functions have been migrated from software to hardware due to the inherent latency of software processing such as interrupt latency and relatively low performance compared with the dedicated hardware processing [2]. One of such examples is MAP decoding, where MAP is a MAC message that defines frame structure with the information about the scheduled access both downlink MAP (DL-MAP) and uplink MAP (UL-MAP) sub-frames. Each subscriber station (SS) can access the frame only after correctly receiving and decoding the MAP message. Since all the MAP processing should be completed to meet the MAP relevance, MAP decoding is regarded as one of timing critical MAC functions in SS. As a result, the IEEE 802.16-2009 standard took the fast map decoding into consideration and adopted nibble alignment as an underlying concept when each MAP IE is designed. It is because nibble alignment reduces the implementation complexity of MAP decoder [3] and thus motivates hardware implementation.

The standard explicitly requests that both hybrid automatic repeat request (HARQ) MAP IE and subburst IE should be nibble-aligned. However, most of subburst IEs, in IEEE 802.16-2009, are not nibble-aligned. Even though there was an effort to fix this inconsistency in the standard [4], the standardization body rejected this opinion due to the compatibility issue with the existing legacy SS. The inconsistency is less likely to be resolved even in the future and it is an obstacle to implementing the MAP decoder by hardware. Therefore, in order to achieve fast MAP decoding with low implementation complexity, it is necessary for a MAP decoder to deal with the MAP in nibble-by-nibble fashion whether the MAP IE is nibble-aligned or not. In this paper, we propose a hardware acceleration architecture for MAP decoding that maintains nibble processing whether the MAP IE is nibble-aligned or not. The proposed architecture is fully implemented by the hardware and its implementation results are provided in this paper.

The rest of this paper is structured as follows. Section II provides background and description of MAP decoding focused on nibble alignment. The definition of nibble alignment is also introduced. The proposed hardware acceleration architecture for MAP decoding and its detail operations are presented in Section III, and FPGA implementation results are covered in Section IV. Section V shows the performance of the implemented MAP decoder in terms of MAP decoding time. Section VI concludes this paper.

II. BACKGROUND AND DESCRIPTION OF MAP DECODING IN IEEE 802.16-2009

MAP itself is simply a MAC management message, which is carried in a payload of MAC Protocol Data Unit (PDU). The inband signaling nature of MAP message and its nontrivial signaling overhead have made the original MAP message obsolete in most of implementations [5]. The compressed MAP is a simplified version of the original MAP adopted to reduce the size of original MAP message. Fig. 1 shows the hierarchical structure of the compressed MAP in IEEE 802.16-2009 focusing on the subburst allocations by HARQ DL/UL

 ISBN: 978-1-61208-164-9

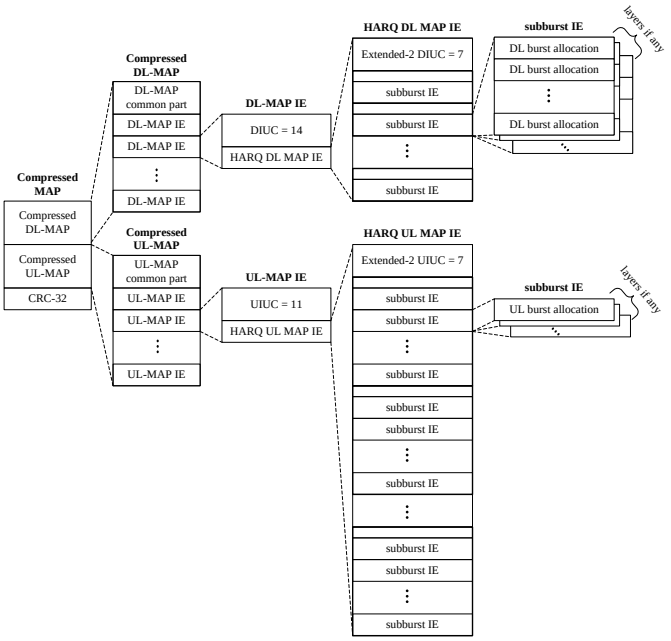


Figure 1. Hierarchical structure of the compressed MAP in IEEE 802.16-2009.

MAP IEs. Contrary to the original MAP message, the compressed map has no generic MAC header (GMH) and the compressed DL-MAP appears first with the first three bits set to 110. The compressed DL-MAP indicates the presence of the compressed UL-MAP by a single bit in the compressed DL-MAP data structure. The compressed DL-MAP includes several DL-MAP IEs identified by a downlink usage interval code (DIUC). The DIUC defines the type of DL access and the DL burst profile associated with that access. A DL-MAP IE entry with a DIUC = 14 indicates that it carries special information and conforms to the DL-MAP Extended-2 IE format [1]. Among the DL-MAP Extended-2 IEs, HARQ DL MAP IE and Persistent HARQ DL MAP IE are extraordinary MAP IEs in that they have a HARQ region for burst allocations with different modulation and coding scheme. Each burst inside the HARQ region is separately encoded and referred to as a subburst. These subburst allocations are specified by subburst IEs included in either HARQ DL MAP IE or persistent HARQ DL MAP IE. The compressed UL-MAP has similar hierarchical structure to the compressed DL-MAP with differences on how to allocate multiple subbursts, as is shown in Fig. 1. In case of DL subburst allocations, multiple DL subbursts with the same HARQ and coding scheme can be allocated by a single subburst IE. On the contrary, every single UL subburst allocation requires its own subburst IE.

At every frame, the compressed MAP is broadcast to all SS in a sector and each SS begins MAP decoding by checking the CRC appended at the end of the compressed MAP. The MAP is discarded and the SS does not access the relevant DL subframe and UL subframe when a CRC error is detected. Otherwise, the SS parses the MAP and extracts all the control information included in MAP IEs. This MAP decoding process is timing-critical due to MAP relevance. Fig. 2 shows MAP relevance with orthogonal frequency-division multiple access

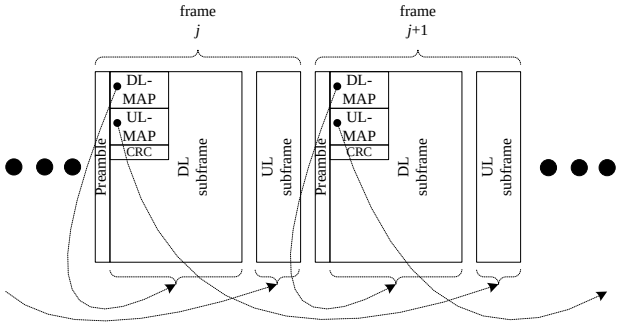


Figure 2. TDD MAP relevance for OFDMA PHY

(OFDMA) PHY in Time Division Duplex (TDD) system. According to the standard, information in DL-MAP always pertains to the same frame in which the MAP was received. In case of UL-MAP, all the information in the UL-MAP applies to UL subframe starting at the Allocation Start Time measured from the beginning of the current frame. TDD system has the ability to control the transition time between DL subframe and UL subframe by adjusting the Allocation Start Time included in the UL-MAP [1]. However, the maximum value of Allocation Start Time is bounded by twice the frame duration and the minimum value is equal to or greater than one frame duration. Referring to Fig. 2, this means that the UL subframe relative to UL-MAP in frame *j* always exists in the next frame (*j*+1). Therefore, all the uplink processing should be completed no later than the beginning of the UL subframe in the next frame in IEEE 802.16-based TDD-OFDMA system.

UL-MAP decoding must be regarded as one of the timing-critical uplink operations because of the upper bound for the UL-MAP relevance. A successful MAP reception without CRC error initiates the decoding process. After decoding the received MAP, the SS can identify the allocated uplink resources. Then the SS constructs MAC PDU by carrying out packing, fragmentation, GMH generation, payload encryption, and CRC calculation. The constructed MAC PDUs are finally concatenated to build an uplink burst and handed over PHY for transmission. It should be noted that all of these uplink processing cannot be done simultaneously but sequentially. Moreover, it is not until MAP decoding is completed that all of these uplink processing can be initiated. Therefore, fast UL-MAP decoding is very important, because the SS cannot access the scheduled uplink medium when it fails to keep the UL-MAP relevance even if SS succeeds in MAP decoding.

Fast MAP decoding is also necessary for downlink. The SS should not ignore any data following MAP message because it has no information about whether there are its own bursts or not. The downlink PHY of SS should buffer all the received data until it receives the result of DL-MAP decoding that indicates the detail allocation information about DL bursts. Thus, fast DL-MAP decoding has a beneficial effect on downlink PHY of the SS in the reduction of the required size of buffer memory. Besides the advantage in memory size, fast DL-MAP decoding enables the SS to send ACK/NAK of DL HARQ burst using UL ACK region in the next uplink subframe. The ACK/NAK signal of DL HARQ burst is sent by

 ISBN: 978-1-61208-164-9

the SS after fixed delay specified by HARQ DL ACK delay for DL burst field in Uplink Channel Descriptor (UCD) message. The standard allows three values for this delay, one, two, or three frame offset. Most implementation adopts the minimum delay of one frame because the fast HARQ retransmission is preferred. In this case, before the beginning of next UL subframe, the SS should complete the DL HARQ processing such as the checking of 16 bits CRC appended at each DL HARQ burst and figure out the offset in the UL ACK region in order to transmit ACK/NACK signals. The offset in the UL ACK region is determined by the order of HARQ-enabled DL burst in the DL-MAP. Besides, UL ACK region can be defined by not only UCD message but also HARQ ACK Region Allocation IE included in UL-MAP. Thus, when the UL ACK region is allocated by UL MAP IE, entire MAP decoding is needed to initiate DL HARQ processing. As a result, fast MAP decoding is also helpful for DL HARQ operations because the minimum HARQ DL ACK delay can be adopted.

III. DESCRIPTION OF THE PROPOSED NIBBLE ALIGNMENT SCHEME

As we mentioned earlier, fast MAP decoding is strongly required and the standard supports it by adopting nibble alignment in MAP IE. The nibble alignment in MAP IE can be classified into two broad categories according to nibble alignment, either at the end of MAP IE and or inside MAP IE. Nibble alignment at the end of MAP IE means that the length of entire MAP IE should be multiples of 4 bits. In the standard, there are three kinds of nibble alignment inside MAP IE. One is nibble alignment at the end of an if-else clause inside MAP IE regardless of whether the if-clause or the else clause is executed. The others are nibble alignment before a loop and inside the loop. Nibble alignment before a loop means that the first bit of a loop should be the MSB (Most Significant Bit) of the nibble and nibble alignment inside the loop means that the length of a loop should be multiples of 4 bits. UL Sounding Command IE is the best example that shows how to accomplish the nibble alignment inside MAP IE. UL Sounding Command IE appends padding bits for byte alignment at the end of MAP IE and utilizes reserved bits 11 times for the nibble alignment inside MAP IE.

Even though all the DL/UL MAP IEs in IEEE 802.16 are byte-aligned or nibble-aligned at the end of MAP IE, nibble alignment inside MAP IE is rarely kept by most of MAP IEs. It is because the nibble alignment inside MAP IE is not mandatory in general MAP IEs. However, the standard explicitly mandates subburst IEs to be nibble-aligned [1]. In spite of this requirement for subburst IEs, unfortunately, most of subburst IEs are not nibble-aligned. Therefore, the current IEEE 802.16-2009 loses consistency throughout the standard with respect to the nibble alignment inside subburst IEs. As a result, the fact that not all the MAP IEs are nibble-aligned should be taken into consideration especially when the MAP decoder is implemented by hardware for the sake of fast MAP decoding. The contribution of our paper is to present a hardware architecture designed for fast MAP decoding on a nibble basis.

Fig. 3 shows the proposed MAP decoder architecture for nibble alignment. According to the control of nibble alignment

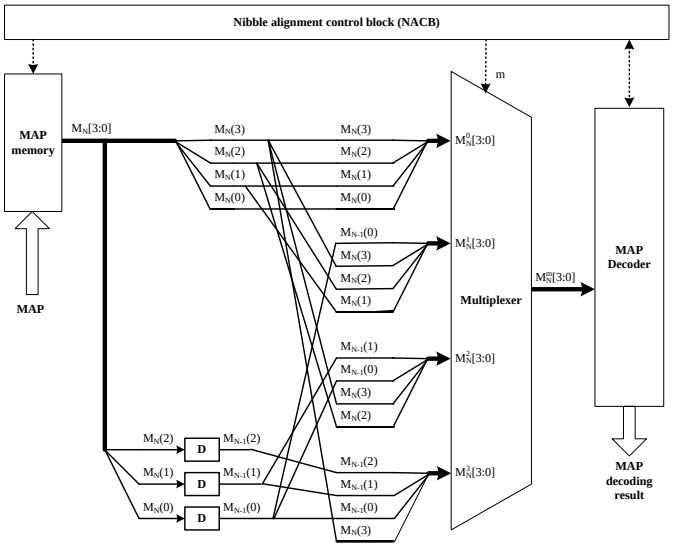


Figure 3. Proposed MAP decoder architecture for nibble alignment

control block, the MAP memory stores valid MAP data with no CRC error and outputs the MAP by the unit of nibble, $M_N[3:0]$, to both multiplexer and D flip-flop, where N denotes time index (N = 0, 1, 2, …). Let $M_0[3:0]$ be the arbitrary value before the MAP memory is filled with received MAP data. Thus $M_1[3:0]$ is the first nibble at the beginning of the MAP. D flip-flop takes a single bit $M_N(B)$ as an input and produces delayed version of it by one time index, $M_{N-1}(B)$, where B is bit index (B = 0, 1, 2, 3) and $M_N(3)$ corresponds to MSB of the nibble $M_N[3:0]$. The initial value of each D flip-flop is set to $M_0(2)$, $M_0(1)$, and $M_0(0)$, respectively, and they are allowed to be arbitrary value. The multiplexer is ordered to select one of the following four inputs according to the alignment mode m (m = 0, 1, 2, 3) specified by nibble alignment control block (NACB). Therefore, the output of the multiplexer is one of the following values:

$$M_N^0[3:0] = M_N[3:0]$$

$$M_N^1[3:0] = M_{N-1}(0) \,\&\, M_N[3:1]$$

$$M_N^2[3:0] = M_{N-1}[1:0] \,\&\, M_N[3:2]$$

$$M_N^3[3:0] = M_{N-1}[2:0] \,\&\, M_N(3)$$

where the operator & stands for bit concatenation. The multiplexer generates the nibbled-aligned MAP, $M_N^m[3:0]$, that is an input to MAP decoder. Finally, MAP decoder reads $M_N^m[3:0]$ and produces MAP decoding result.

More detail operation of nibble alignment operations are illustrated in Fig. 4. Assume that the MAP decoder is in state K and the current mode is m. Let the current output of MAP memory be $M_{n+1}[3:0]$ ($n \in N$). At this time, suppose that the MAP decoder has found that a loop in the MAP IE begins at the bit $M_{n+1}^m(b)$ where $b \in B$ and $b \neq 3$. Notice that if b=3, the current input of MAP decoder, $M_{n+1}^m[3:0]$, is nibble-aligned before the loop and we do not need any further operations for nibble alignment. Since the first bit corresponding to the beginning of a loop is not the MSB of $M_{n+1}^m[3:0]$, the MAP

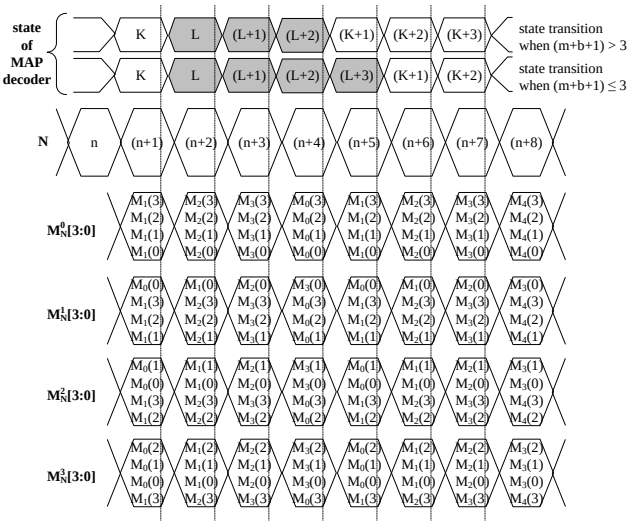


Figure 4. State transition of MAP decoder for nibble alignment

decoder makes a request for nibble alignment at the bit $M_{n+1}^{m}(b)$ to NACB. Thus only (3-*b*) bits of the current nibble, $M_{n+1}^{m}[3:(b+1)]$, is used for map decoding in state *K* and (*b*+1) bits are not used. The MAP decoder informs NACB of the number of bits that is not used for decoding in the current state and it transits to the overhead state *L*. The states from *L* to (*L*+3) are overhead states because no actual MAP decoding takes place in these states and they are used only for nibble alignment. In state *L*, NACB changes the read address for MAP memory so that the output of MAP memory can be $M_0[3:0]$ in state (*L*+2). At the same time, it carries out the addition of (*m*+*b*+1). In state (*L*+1), NACB finds a new mode, *m*', by performing following modulo operations with (*m*+*b*+1) obtained from the previous state:

$$m' = (m+b+1) \bmod 4.$$

where mod is a modulo operator. In state (*L*+2), NACB compares (*m*+*b*+1) with 3 and it transits to state (*K*+1) if (*m*+*b*+1) is greater than 3. Otherwise, it transits to state (*L*+3) which is simply a temporary state before state (*K*+1). The output of multiplexer in state (*K*+1) is now $M_{n+5}^{m'}[3:0]$ if (*m*+*b*+1) > 3 and $M_{n+6}^{m'}[3:0]$, otherwise. In either case, the MSB of the input to the MAP decoder is the start of the loop in state (*K*+1). The MAP decoder resumes MAP decoding process from the state (*K*+1) with the nibbled-aligned MAP data. Therefore, the presented alignment scheme makes it possible for the MAP decoder to operate on a nibble basis.

IV. IMPLEMENTATIONS

The feasibility of the proposed architecture is investigated by implementing the compressed MAP decoder using Xilinx Virtex-4 XC4VLX200 FPGA. The implemented MAP decoder is capable of decoding the MAP including 10 DL MAP IEs and 17 UL MAP IEs. For the commercial Mobile WiMAX system [6], only sub-MAP and its relevant MAP IEs are excluded from the implementations. The implementation reveals that the nibble alignment inside the loop and at the end of if-else clause is much more important than before a loop. If the if-else clause is not nibbled alignment, the logic size doubles from the end of if-else clause. Moreover, if the loop is not multiples of 4 bits, the logic size of the loop increases by two times when the remainder of the length of the loop modulo 4 equals to 2, or increases by four times when the remainder of the length of the loop modulo 4 equals to 1 or 3. Since the proposed architecture supports nibble alignment for any MAP IEs, the MAP decoder can operate on a nibble basis while avoiding the dramatic increase in the size of required logic when the MAP IE is not nibble-aligned. Table I indicates the implementation results in term of FPGA resource utilization. Only a small fraction of FPGA resources are utilized. According to the synthesis report, the implemented MAP Decoder can operate up to 140 MHz clock frequency.

TABLE I. FPGA IMPLEMENTATION RESULTS

Resources	Available (a)	Used (b)	Utilization (b/a)
Number of Slice Filp Flops	178,176	5,649	3%
Number of 4 input LUTs	178,176	6,492	3%
Number of occupied slices	89,088	5,090	5%
Number of FIFO16/RAMB16s	336	7	2%
Number of DSP48s	96	3	3%

V. PERFORMANCE

The performance of the implemented MAP decoder is investigated by register transfer level simulation with 80MHz clock speed. Performance evaluation is carried out in terms of MAP decoding time, which is defined as the time elapsed between the moment the MAP decoder reads the first nibble and the moment it reads the last nibble of the MAP. The MAP decoding time does not include the time required to verify CRC-32 appended at the end of MAP because the MAP memory stores valid MAP data with no CRC error. As we mentioned previously, most of subburst IEs are not nibble-aligned. Thus, in this section, we concentrate on investigating the MAP decoding time of a MAP comprised of subburst IEs.

Two different kinds of MAP constructions are considered to study the performance of the implemented MAP decoder. One is a MAP construction for DL 2x2 Spatial Multiplexing

TABLE II. MAP IES FOR a)DL 2X2 SM AND b)UL 2 LAYER CSM

	MAP IE
a)	Compressed DL MAP
	CID Switch IE
	Space-Time Coding (STC)/DL Zone switch IE
	HARQ DL MAP IE
	MIMO DL Chase HARQ subburst IE
	Dedicated MIMO DL Control IE
	Compressed UL MAP
b)	Compressed DL MAP
	Compressed UL MAP
	CDMA allocation IE
	CDMA allocation IE
	HARQ ACKCH Region Allocation IE
	FASTFEEDBACK allocation IE
	PAPR Reduction/Safety Zone/Sounding Zone Allocation IE
	UL Zone Switch IE
	HARQ UL MAP IE
	MIMO UL Chase HARQ subburst IE

 ISBN: 978-1-61208-164-9

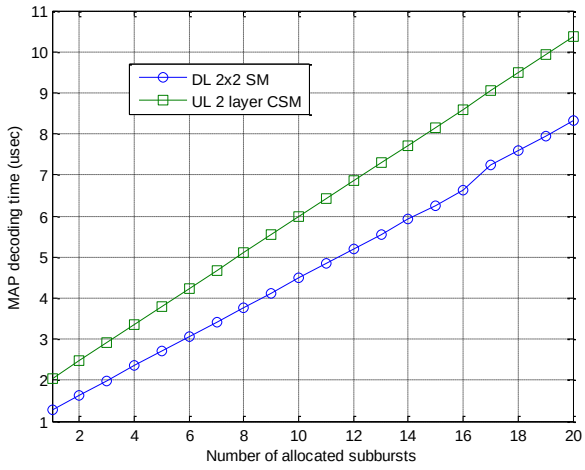


Figure 5. MAP decoding time versus the number of alloated subbursts

(SM) subburst allocations with no UL allocations. The other is a MAP construction for UL 2 Layer Collaborative Spatial Multiplexing (CSM) subburst allocations with no DL allocations. Table II shows the details of both MAP constructions. Each row of Table II is carefully indented in order to show MAP hierarchy. Notice that DL 2x2 SM subbursts and UL 2 Layer CSM subbursts are allocated by MIMO DL Chase HARQ subburst IE and MIMO UL Chase HARQ subburst IE, respectively. In case of DL 2x2 SM, the for() loop inside MIMO DL Chase HARQ subburst IE iterates as the number of subburst allocations increases. On the contrary, in case of UL 2 Layer CSM, the iteration is caused by for() loop surrounding MIMO UL Chase HARQ subburst IE. In either case, the size of MAP increases as the number of subburst allocations increases.

Fig. 5 shows the MAP decoding time of the implemented MAP decoder with respect to the number of allocated subbursts. Regardless of the number of subburst allocations, MAP decoding time of UL 2 layer CSM is larger than that of DL 2x2 SM . It is because the MAP size of UL 2 layer CSM is much bigger than that of DL 2x2 SM due to the overhead imposed by UL control region allocation. CDMA allocation IEs, HARQ ACKCH Region Allocation IE, FASTFEEDBACK allocation IE, and PAPR Reduction/Safety Zone/Sounding Zone Allocation IE is needed to allocate UL control regions such as ranging regions, HARQ ACK region, CQI region, and Sounding channel, respectively. Even though MAP coding time of each scheme is different, their MAP decoding time increase linearly for the observed number of allocated subbursts. However, in case of UL 2 layer CSM, MAP decoding time abruptly increases when the number of allocated subbursts changes from 16 to 17. Since MIMO UL Chase HARQ subburst IE can be used to allocate 16 subbursts at most, one more MIMO UL Chase HARQ subburst IE is needed to allocated 17 subbursts. Thus, the overhead arisen from one additional subburst IE makes a rapid increase at 17 subburst allocations.

Obviously, MAP decoding may be processed by software. In such a case, the initiation of MAP decoding will rely on an interrupt driven by channel decoder. It is interesting that the mean value of interrupt latency is approximately 10 μsec in a PC operating at 2GHz CPU core [2]. From Fig. 5, notice that the implemented MAP decoder operating at 80MHz can decode a MAP with 19 subburst allocations within 10 μsec. In other words, even before a software MAP decoder is invoked, the presented hardware MAP decoder can complete MAP decoding in most cases in Fig. 5. Thus, fast MAP decoding is the benefit of the proposed hardware acceleration architecture for map decoding that maintains nibble processing whether the map IE is nibble-aligned or not.

VI. CONCLUSIONS

MAP decoding by hardware is required for the fast MAP decoding that is needed to achieve the broadband access provided by the system based on IEEE 802.16-2009. The nibble alignment is adopted by the standard to support the fast MAP decoding while suppressing the increase in implementation complexity. However, some of the MAP IEs are not nibbled-aligned, which causes difficulties when the MAP decoder is implemented by hardware. In this paper, we propose a hardware architecture for MAP decoding that maintains nibble processing with the help of nibble alignment. From the synthesis result and the investigated MAP decoding time, we can conclude that the proposed architecture facilitates implementing a realistic fast MAP decoder with low hardware complexity.

ACKNOWLEDGMENT

This work was supported by the IT R&D program of MKE/KEIT. [KI002108, Research on Radio Transmission Technology for IEEE 802.11 VHT WLAN].

REFERENCES

[1] IEEE Standard for Local and Metropolitan Netoworks, Part 16: Air Interface for Broadband Wireless Access Systems, IEEE Std. 802.16-2009, May. 2009.

[2] Sunkyung Shin, Seungkwon Cho, Jaewoo Park, Yeong-Gon Lee, and Sok-Kyu Lee, “MAC HW/SW partitioning for aggregation in IEEE 802.11n based on a interrupt latency measurement”, ICCE 2010, pp. 291-292, Jan. 2010.

[3] Jaejoon Park, Seokheon Cho, Youngil Kim, Chulsik Yoon, Mihyun Lee, Inseok Hwang, and Panyuh Joo, “Corrections for Nibble Alignment in MAP_IEs”, IEEE 802.16’s 802.16 Task Group e, June. 2005.; http://www.ieee802.org/16/tge/contrib/C80216e-05_271r2.pdf

[4] Seungkwon Cho, Seokhun Cho, Jaesun Cha, and Young-il Kim, “Corrections for nibble alignment in MIMO HARQ subburst IEs”, IEEE 802.16’s 802.16 Task Group Maintence, Jun. 2008, http://www.ieee802.org/16/maint/contrib/C80216maint-08_251.doc.

[5] Fan Wang, A. Ghosh, C. Sankaran, P. Fleming, F. Hsieh, and S. Benes, “Mobile WiMAX Systems: Performance and Evolution”, IEEE Commun. Mag., vol. 46, pp. 41-49, Oct. 2008.

[6] WiMAX ForumTM Mobile Protocol Implementation Conformance Statement (PICS) Proforma, DRAFT-T24-001-R010v06-A, WiMAX Forum, Jun. 2009.

 ISBN: 978-1-61208-164-9

Digital Signature Platform on Mobile Devices

José Manuel Fornés Rumbao
Department of Telematic Engineering
Seville University
Seville, Spain
fornes@trajano.us.es

Francisco Rodriguez Rubio
Department of Systems and Automatic Engineering
Seville University
Seville, Spain
rubio@esi.us.es

***Abstract*— Since ancient times, the obsession with security and authenticity has been an issue that has produced the development of diverse technologies, to avoid the access of third persons to information or to private places, and to guarantee the identity of a person with regard to a certain fact. Undoubtedly, in a society like today, which comes naming like "of the information", these issues are a basic aspect given the large number of everyday situations that occur relating to the use of confidential information and purposes of authenticity. For it, it becomes necessary to investigate, legislate, and to develop applications and systems that help to preserve the security and authenticity of a user and so, to provide them with sufficient capacity in order that these aspects in telematic networks are exported to the world of the mobile devices. This article describes the mechanisms that can be used as digital signatures and certification to obtain the electronic security in mobile devices. In addition, we propose a real platform already realized for the implementation of the digital signature in mobile terminals.**

Keywords - digital signature; certificate; midlets; servlet; cryptography.

I. INTRODUCTION

In today's society mobile devices have become widely accepted. They have achieved a great popularity, thanks to its easy operation and low cost being widely used by people all around the world, increasing their power every day. This processing power, which grows more and more, will do possible the operations of calculating of summary data algorithms (hash functions) [1] and the digital signature, also; it will be possible to do on a mobile phone in a little period of time. And all this facts combined with the great improvement in the capacity of mobile networks makes it very interesting to research and it develops technologies that are suited to the terminals for conducting electronic signatures [2].

The mobile world should adapt itself to new trends in electronic signatures and digital certificates that it is concerned, both for electronic commerce and to the carry out administrative´s procedures online, as well as other possibilities offered by technology in this field [3]. Therefore, we find that, on the one hand, mobile phones are functional devices with a great potential. On the other hand, there is a way to get that the citizens interact with the government and other existing services in this field, through electronic means.

This article seeks two objectives, to report about the current existing technologies and develop an application that allows electronic signature capabilities through mobile phone with a digital certificate installed.

The paper is organized as follows. In Section II, we review the theory of cryptography, its objectives, its implementation on mobile devices and the different alternatives that we have. In Section III, we present the development of our platform divided in client and server. Section IV shows the results of our platform and in Section V we propose the possible improvements. Finally, Section VI concludes the paper.

II. STATE OF THE ART

A. Electronic Security

The basic pillars of security [4] in communications and secure exchange of electronic documents are:

- Privacy: Preventing that a third party may intercept (read) the information submitted in the case of sensitive data.
- Integrity: Preventing that another agent outside the issuer, get modify (insert, delete, mess, etc.) the information sent.
- Authenticity: Preventing that a third party can impersonate the other party.
- Non-repudiation: Preventing the other party could deny the participation in communication (either as a source or destination).

These are conditions that must be met in order to establish a secure environment in the digital world and we will see that this security can be ensured by the use of cryptography, digital signatures and certificates.

Current cryptography is mainly divided into two very distinct branches which are detailed below.

1) Symmetric Cryptography

Symmetric cryptography [5] refers to the set of methods that allow secure communication between the parties, because the key has been exchanged previously, which is called symmetric key. Symmetry means that parties have the same key to encrypt and decrypt. This type of cryptography is also known as private key cryptography.

Although there is no standard type of design, perhaps the most popular is the DES (Data Encryption Standard), which is essentially a cryptographic system that takes as input a

 ISBN: 978-1-61208-164-9

block of 64 bits of the message and is submitted to 16 interactions, with a key of 56-bit.

With a brute force method, it could break DES by making it unsafe for high security purposes. The option to replace DES has been a new encryption system which is now known as triple-DES or TDES which consists of applying DES three times.

A large number of symmetric cryptographic systems was designed in the past 20 years, including some of them which are: RC-5 [6], IDEA [7], FEAL [8], LOKI'91 [9], DESX [10], Blowfish [11], CAST [12], GOST [13], etc. However, they have not had the scope of DES, although some of them have better properties.

2) Asymmtric Cryptography

Asymmetric encryption algorithms [14] use a different key pair in communication, one to encrypt and another to decrypt. Both keys are mathematically related and it is virtually impossible to derive one from another. The key pair is generated based on asymmetric encryption algorithm used, being a secret (private) and the another is known for others (public key).

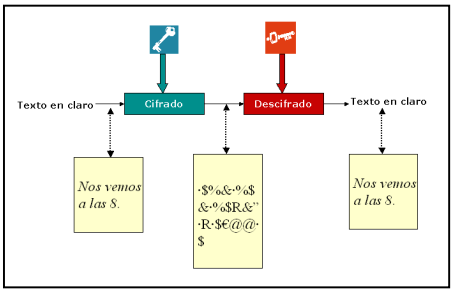


Figure 1. Process communication in an asymmetric key system [15].

The clear message is encrypted with the private / public key and only his partner (private / public) can decrypt it. The safety of this system is based on the impossibility of calculating a key from another, besides, of course, to keep the private key secret.

A public key cryptosystem to note is the RSA algorithm [16]. It is based on the difficulty of factoring large numbers. The messages sent using the RSA algorithm are represented by numbers and the operation is based on the product of two large prime numbers (greater than 10^{100}) chosen at random and the decryption key. His security is that there are no quick ways of factoring large numbers into prime factors using conventional computers. The size of his key is not fixed, meaning that if the factoring of RSA modules employees is committed; keys with greater lengths can be chosen to maintain the security of the cryptosystem.

Should be noted that one of the most important applications of public key cryptography is the digital signature. The origin signs a message with its private key.

B. Signature and Digital Certificate

The RSA algorithm is reversible; ie, in addition to allow the public key encryption and decrypt the message with the private, it also allows to encrypt with the private key and decrypt with the public. This latter mode of encryption does not provide confidentiality because anyone can decrypt the original message since they can always get the public component of the speaker, however, encrypt a message with the secret key of a user involves a clear identification (so get the authenticity and non-repudiation) and that only with the key associated with their identity can decipher, as does a handwritten signature, so this process is known as Digital Signature.

A Digital Signature [17] consists basically on three parts:

1. Key pair generation, private (with which it is signed) and public (with the one verified by a third party). This key generation is done according to a particular algorithm, as we have been seeing before: the RSA.
2. Signature of the document. With the private key signed the message.
3. Signature verification by a third party. Given the signature and public key, another user can validate the signature.

As the computational cost of public key algorithms is fairly high, to sign a large amount of data, when a message is large, the digital signature could be extremely slowly. For all this, is applied to the document a summary way function (hash function) to obtain a hash value, which is only a summary of the document. The digest or hash functions [18] are used to compress a text or document in a fixed length block. Hash functions should be public and irreversible, that is, from the abstract can not recover the original text. Not encrypt only compressed text or documents in a fixed length block. The hashing algorithm SHA-1 has been examined closely by the public cryptographic community, and has not found any effective attack.

The Digital Signature, thanks to the asymmetric cryptographic algorithms, can replace the traditional signature on paper, as it offers these features:

- Document integrity. If the document was modified during transmission, the signature verification will be missed.
- Identity and authenticity. Only a public key associated with the user who signed with his private key can correctly decrypt the digital signature.
- Non-repudiation. The user can not deny that he signed his authorship of the Digital Signature.

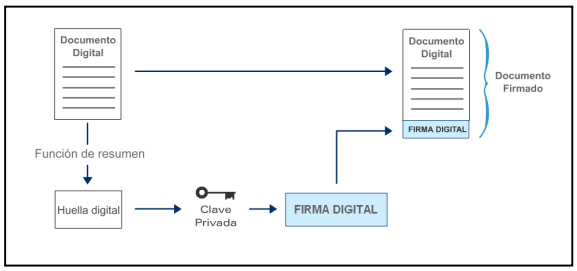


Figure 2. Generating Process of a Digital Signature [15].

The digital signature ensures data integrity and authenticity, so that errors caused during transmission, it

would become apparent during the process of signature verification:

- Separate content unencrypted (original) of the signature itself.
- The receiver proceeds to calculate the hash of the received document according to the algorithm used (in this case, the SHA-1). As a result you get 160 bits.
- Now we proceed to calculate the decryption of the signature received with the sender's public key, and by the same asymmetric algorithm that was encrypted (RSA in this case). This gives a string of 160 bits should match the digital signature calculated in the previous step.

If both matches, the verification is correct, the document was signed by the issuer and data were no corrupted.

The format of the digital signature depends on the way in which they perform. It is important note the signature with syntax ASN.1 referenced to the standard RFC 3852 and it is based on the set of standards PKCS # 7 / CMS.

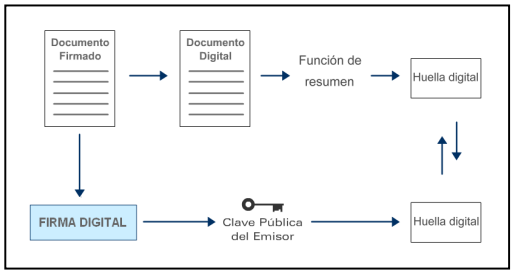


Figure 3. Verification Process of a Digital Signature [15].

A digital certificate [19] is a document issued and digitally signed by a Certification Authority that combines the distinctive name of a person or entity with its public key for a period of time. They are digital documents that serve to ensure the accuracy of the public key certificate belonging to the owner or the entity, which is digitally signed documents that can provide the most absolute security guarantees.

This mechanism provides the users to verify the authenticity through the network, controlling access to resources, etc. There are several formats for certificates and the most widespread is the X.509 version 3. This format is a standard of ITU and ISO / IEC.

C. Security on Mobile Devices

Different technologies are capable of offering digital signature capabilities and user authentication using a mobile phone. There are several articles about this [20], [21], [22]; however they are abstracts and with unrealistic solutions. Then, we explain those most important tools that have been used for the development of the platform.

We can attend to commercial and technical reasons [23] for the selection of the platform. In the first, we will consider the market penetration of the different alternatives. As for technical reasons, we will study the speed and simplicity of coding and prototyping. The following image shows an analysis of the status of the market related with installed based vs available apps disparity. Here we can see that the "old guard" (Symbian, Java ME and flash) has a larger installed base, but a smaller number of available apps than the newer platforms (Android, Iphone and BlackBerry).

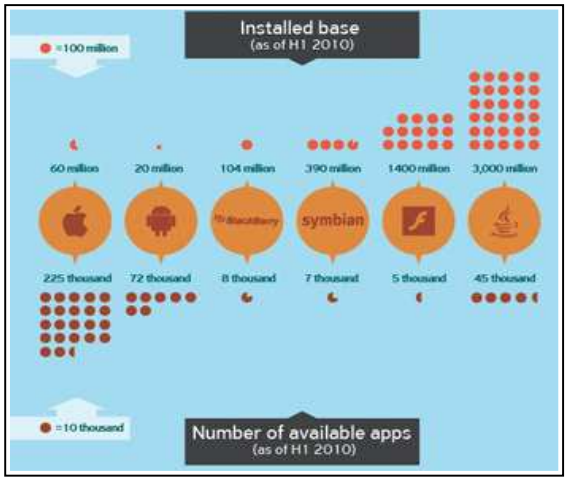


Figure 4. Installed base and number of available apps [24].

If now we attend to the quick coding and prototyping, we can consult another study reflected in this image. Here, we see that the easiest platform to master is Android (5 moths), while the hardest is Symbian (15 months).

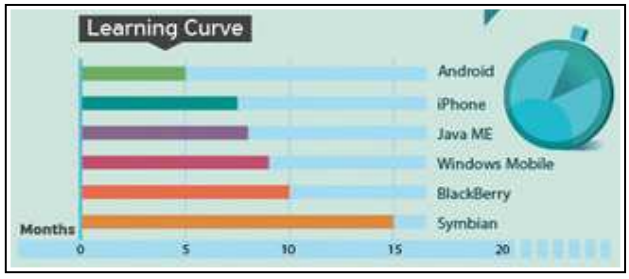


Figure 5. Platform´s learning curve [24].

In this article, we have chosen the Java ME platform for the rapid development and great penetration into the market. Moreover, this solution has low cost about developer tools.

1) Java Mobile Edition

Java ME, formerly known as Java 2 Platform Micro Edition or J2ME, is a technology for developing Java applications on mobile devices such as phones and PDAs. Java ME consists on programming specifications and a special virtual machine which allows that a Java ME program can be run on a mobile device. JavaME services are based on local programs, called MIDlets, which the user can download and install on his terminal. MIDlets can be run locally on the device and also provide client-server sessions. There are two settings for Java ME, configuration for devices with limited connection (CDLC) and other for connected devices (CDC).

 ISBN: 978-1-61208-164-9

The development of a Java ME application follows several steps, which we detail in four:

- Edit the source code.
- Compilation: The code is compiled and the class files are obtained.
- Pre-verification: To prevent malicious applications are installed.
- Packaging: It generates a .JAR.
- Deployment: Installing MIDlet on the mobile terminal.

Noted that the Java ME cryptographic operations are performed with the addition of cryptographic libraries according to the standardized set of APIs in the Java Cryptography Architecture (Java Cryptographic Architecture - JCA).

2) Security and Trust Services APIs

API Security and Trust Services (Security and Trust Services API) [25] is a specification for Java ME and give the possibility of opening a channel of communication between a Java MIDlet and a "security element." This security element can be a smart card with cryptographic module WIM, or the internal security element that provides the S.O. of a terminal.

Using SATSA, JSR-177 is also a security element which can perform all security processes such as electronic signature or authentication of users in Java ME applications. The API of Satsa has four optional packages for the different needs of communication with the security element. The communication mode depends on the type of application. These packages are APDU and JCRMI, intended to communicate with smart cards. The package PKI is used for credential management and digital signature. Finally, CRYPTO is used to perform cryptographic operations.

3) Symbian OS

This operating system [26] has been developed exclusively for mobile terminals and its code is provided to major phone manufacturers like Nokia. Provides important functions related to authentication, and confidentiality and integrity of data. Also, Symbian allows the management of certificates with a cryptographic module, and implementation of standard cryptographic algorithms, hash functions, key and random number generation.

These technologies have been used to create some examples of existing services such as Mobipay or some proprietary driven by large operators such as Vodafone and Telefónica.

III. DEVELOPMENT OF THE PLATFORM

The basic objective is to implement a digital signature mechanism in a mobile phone. Of all the existing technologies, Java is chosen to try to ensure more widely as possible among the existing terminals. For it, a Java ME application (MIDlet) for a Nokia N95, has been developed which uses the private key of a personal digital certificate from the Fabrica Nacional de Moneda y Timbre, issued in the conventional way, and so perform cryptographic tasks in the mobile digital signature.

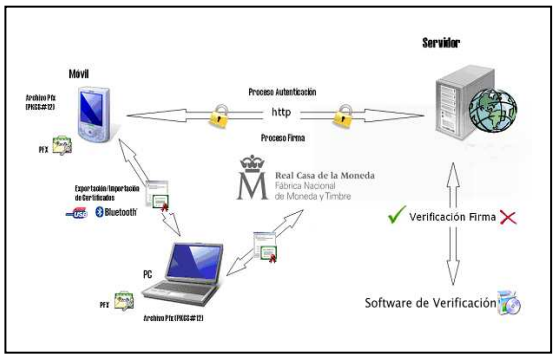


Figure 6. General scheme of the Platform [27].

The main elements involved in this platform are:

- Mobile phone N95. With Java and JSR-177 factory installed.
- Conventional certificate in PKCS # 12 format issued by the Fabrica Nacional de Moneda y Timbre.
- Server with Tomcat version 6 installed.
- Development environment NetBeans IDE 6.0.1.
- Software digital signature verification.

We will study the environment of the two most important components of this platform, the client and server.

A. Client

To give the client the functionality required for this platform is necessary, to incorporate into the mobile device (N95); two fundamental elements: the digital certificate and the MIDlet [28]. These will be discussed below, however, previously it would be useful to show the structure that contains the device protocols:

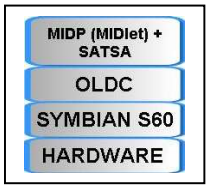


Figure 7. Structure of protocols of the mobile device [29].

1) Digital Certificate

The first thing is to have a personal digital certificate issued by the Fabrica Nacional de Moneda y Timbre. The steps to follow to obtain are the following:

- Request the certificate through the internet.
- Proof of identity in a registry office.
- Download digital certificate user.
- Export Certificate.

Once obtained the personal digital certificate in PFX / PKCS # 12 format, is exported to the mobile terminal. For it; the transfer file has been made through Bluetooth from the PC although it could be used for a USB cable. In the inbox is stored the message received by Bluetooth, which contains the file with the personal digital certificate and private key. When we try to open this message; a key is requested, which was inserted during the previous export stage. When you

 ISBN: 978-1-61208-164-9

enter the password correctly the phone detects that contains the file and will proceed to the installation of the certificate.

Thus, the private key is stored in the security element and the user certificate and the certificate authority in his default store.

2) Midlet

The application on the client side consists of a MIDlet called MozartPKI. This MIDlet has been developed for terminals with profile MIDP 2.0 and limited CLDC 1.0. connection settings. The MIDlet has been developed so that, using SATSA cryptographic services, a user can send to server a plain text message digitally signed by him, and that another person may recover the server and check it later. The MIDlet consists of 2 classes in the package es.minerva.mozart. A MIDlet class itself is called MozartPKI.java and the other is Infodata.java. The Infodata.java class is responsible for reading the IP address of an external file server, so if you want to change the IP delivery of the signature file (the server), we would not need to tweak the source code, just enough to change the file properties.

To install the MIDlet the first thing to do is compiling and to package the source code and resources of the client application. For it, NetBeans 6.0 is used for the development environment (compile and build the JAR file).

To install on the mobile phone it is enough to send the JAR file, which is the MIDlet itself. For this test, we will use the Bluetooth technology. Thus, once the mobile has received the JAR file, is able to install the MIDlet. With the MIDlet installed on the Nokia N95 can be checked by accessing the phone's application menu and noting that this application is available as the others:

Figure 8. Correct installation of the MIDlet "MozartPKI" [27].

B. Server

This server consists of a Java Servlet, ServerPKI with a web page, index.jsp. Java Servlets [30] are objects that are within the context of a servlet container (in this platform will be Tomcat 6). The Servlet works very similar to the MIDlet, when the client that runs on mobile phone, establishes a connection to the server and sends the POST request, the servlet responds by trying to keep the inflow as array of bytes in a file on his disk. If the process is successful, it opens an output stream to send the MIDlet acknowledgment message indicating that the process has finished successfully.

The Java Servlet is compiled by NetBeans and which has been developed to run on the server. As a result of the compilation and construction of the Web application; a file .WAR is created in NetBeans and must be installed on the Servlet container, which in this case is Tomcat 6.0. To do this you access the Tomcat management console and select the WAR file to deploy it for.

IV. RESULTS

A. Client

Accessing MozarPKI, we see the operation of this MIDlet. Once launched, the welcome screen appears like you can see in the image and the entrance to the right:

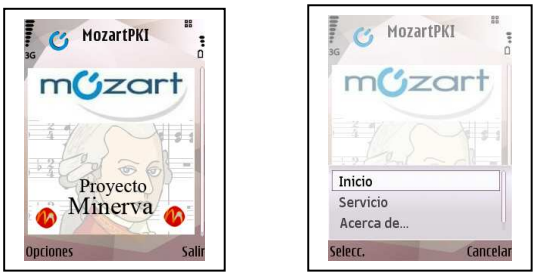


Figure 9. Screen of welcome and beginning of the N95 [27].

First, the MIDlet requests to the service user, the insertion of the message to sign, and once you have written is confirmed by pressing the corresponding button. The message to be signed by way of example is "I received your gift". The message is then introduced and offers the possibility to modify it if the client detects a fault in writing:

Figure 10. Text editor and confirmation [27].

If the user checks that the message entered is correct, click on the button "Sign" and the cryptographic process start. The security element of Symbian notice that an application (the MIDlet) is requesting access to it, and made a prompt the user to accept if you agree. Once the user has accepted the request from the MIDlet, to access the security element, SATSA does a "sweep" by the security element which, in the case of Nokia N95 is the internal operating system Symbian. The installation of the user's certificate is found and is shown on screen in order to be selected.

 ISBN: 978-1-61208-164-9

Figure 11. Certificate selection, access to elem. Sec And Signature [27].

When the user selects the certificate that he wants, SATSA tries to access the private key. For security reasons, it is under a password. If the key is entered correctly SATSA proceeds to perform the digital signature of the inserted data in the format CMS / PKCS # 7 SignedData containing the signer's certificate, the data that have been signed and the digital signature itself.

Once the signing process is complete, the file containing the digital signature should be sent to the server which is running the Servlet developed. When the user clicks on the button "Send" starts the process and then the MIDlet retrieves from the file properties; by the methods of the class InfoData.java, the URL of the server and displayed on screen for the user to confirm. The user must also to confirm the request of the MIDlet to connect to the Internet to send the file to the server, since this connection could lead to charges associated with the carried through the GPRS / UMTS. Once the HTTP connection is established the MIDlet looks forward to receiving the consent of the server.

Figure 12. URL's obtaining and sending to Server [27].

The MIDlet, as detailed above, sends the CMS file and looks forward to receiving the assent or the server-side error. If everything works properly, the message "OK" is got on screen as shown in the previous image. The process has been completed for the MIDlet as signed message has been stored correctly as indicated by the assent of Servlet.

B. *Server*

Looking index.jsp, the website of Servlet is presented: Server Status and message received.

The followings images shows the appearance of the website. The link server status is available to inform the user if the service is active or not. To download the signed file, click on the link and opens the download window.

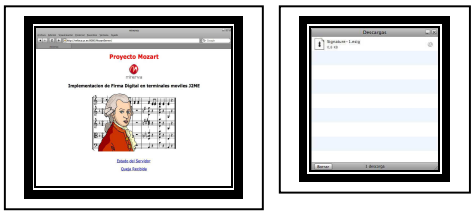


Figure 13. Server Web page and download window [27].

C. *Verification*

The PKCS # 7 / CMS file, that contains the signed data; needs to be checked to complete the process and check that the signature has been generated by the mobile correctly. To carry out the process of verifying of the digital signature; by the user on their mobile device, we have used the free tool eSign Viewer [31].

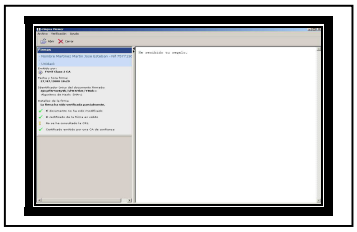


Figure 14. ESigna Interface Viewer [27].

This tool allows three things: view the contents of the signed, the signatures and the validity of the signatures applied to the same.

In the image, right, appears correctly separated the message sent from your digital signature. On the left, the program displays the report with the results of opening the file and checks the signature correction. This information shows how Esigna tool verifies if the signature made by the J2ME application in the mobile terminal is valid and if complies with the PKCS # 7 / CMS format. A third party who want to verify the authorship of the message can do it and be sure who sent it.

V. FUTURE WORKLINES

Three future worklines can be analyzed, taking into account the most important elements of the platform made:

We used certificates issued by the Fábrica Nacional de Moneda y Timbre (FNMT) since it is considered the major Certification Authority in Spain. In private business, the FNMT certificates are not used, because so the complicated problem of the validity check in LDAP directories is released, which have restricted access. This would be a possibility for a future attempt to open the application, and integrate developments, certificates issued by other Certification Authorities as Camerfirma.

This Project has used Java ME with libraries SATSA for signing and certificate management. It could mean a restriction on developments since the service provider may require other forms of signature or the use of certain cryptographic algorithms that do not implement this API. At

 ISBN: 978-1-61208-164-9

this point, it is worth mentioning the possible to use of other Java libraries like BouncyCastle or explore alternatives for the integration of cryptographic cards in mobile devices.

The fact that the digital signature has been implemented in the development of this project is conducted in CMS format, ensures interoperability of the developments made. In this respect, as a possible future work is the integration and improvement of the application developed in Java ME for interconnection with a public platform or other third parties wishing to provide service.

VI. CONCLUSIONS

The main goal was to develop a study of the state of the art in the field of digital signature applied to mobile telephony as well as investigate the possibility of integrating the use of digital certificates on mobile devices business. In this aspect, we have managed to integrate a digital certificate on a mobile phone and use it for services of authentication, digital signing of data and establish secure connections. All this in a closed and controlled environment, but easily extended to more realistic situations.

On the other hand, the field of digital signature and certification is currently at a low growth and maturity among the population, and less on mobile use, so job opportunities are numerous. The truth is that, as noted in this article, there is sufficient technological resources to address the field of cryptography and digital signatures using digital certificates. All that is needed is investment and effort to try to get what may be another big boom in mobile communications.

REFERENCES

[1] Kim Mooseop, Ryou Jaecheol, and Jun Sungik, "Compact Implementation of SHA-1 Hash Function for Mobile Trusted Module," Information Security Applicaions, Volume 5379/2009, pp: 292-304, 2009. http://www.springerlink.com/content/f2v64u64324w6q47/

[2] The National Electronic Commerce Coordinating Council (NECCC), "Impact of Electronic Signatures on Security Practices for Electronic Documents," 2001. http://www.azsos.gov/pa/ec3/Security_Practices_ED.pdf

[3] Trustgate Berhad. "The advent of an Interoperable Ecosystem for Secure Mobile Transactions," 2009. http://www.msctrustgate.com/pdf/Mobile_Signature.pdf

[4] Lex Nova Magacine Report, "*Firma electrónica: Seguridad a través de la red*," 2004. http://tecnojur.blogs.lexnova.es/2011/02/05/aspectos-basicos-de-la-firma-electronica/

[5] José Jesús Angel, "*Criptografía Simétrica*," 2007. http://elsitiodetelecomunicaciones.iespana.es/cbasica.pdf

[6] R. L. Rivest, "The RC5 Encryption Algorithm," Proceedings of the Second International Workshop on Fast Software Encryption (FSE) 1994e. pp. 86–96, 1994.

[7] Joan Daemen, Rene Govaerts, and Joos Vandewalle, "Weak Keys for IDEA," Advances in Cryptology, CRYPTO 93. pp: 224–231, 1993.

[8] Shoji Miyaguchi, "The FEAL Cipher Family," CRYPTO 1990, pp: 627–638, 1993.

[9] Lars R. Knudsen, "Cryptanalysis of LOKI," Advances in Cryptology - ASIACRYPT'91, LNCS 739, pp 22–35, H Imai et al. (eds), Springer-Verlag, 1993.

[10] Eli Biham and Adi Shamir, "Differential Cryptanalysis of the Data Encryption Standard," Springer Verlag. ISBN 0-387-97930-1, ISBN 3-540-97930-1, 1993.

[11] Bruce Schneier, " Description of a New Variable-Length Key, 64-Bit Block Cipher (Blowfish)," 1993. http://www.schneier.com/paper-blowfish-fse.html

[12] "RFC 5830: GOST 28147-89 encryption, decryption and MAC algorithms," IETF, 2010-03. http://tools.ietf.org/html/rfc5830

[13] C.M. Adams, "Constructing Symmetric Ciphers Using the CAST Design Procedure," Designs, Codes, and Cryptography, 12(3), pp. 283–316, 1997.

[14] L. M. Leyva, "Sistema criptográfico," 2008. http://investigacion.uagro.mx/3coloquio/exa/11.pdf

[15] Figure 1, 2 and 3, Source: gdp.globus.org, 2009.

[16] Real Academia de Ciencias, "*Criptografía de clave pública*," El sistema RSA, 2006. http://www.uam.es/personal_pdi/ciencias/ehernan/Talento/ViceenteMunoz/rsa_2006.pdf

[17] Mauricio Devoto, "*Comercio electrónico y la firma Digital*," 2008. 74.125.155.132/scholar?q=cache:sr3OWIhUeX0J:scholar.google.com/+firma+digital&hl=es&as_sdt=0,5

[18] Unizar, "*La seguridad en informática -Funciones Hash*," 2005. http://criptosec.unizar.es/doc/tema_c7_criptosec.pdf

[19] Sergio Talens – Oliag, "*Introducción a los Certificados Digitales*," 2003. http://www.uv.es/sto/articulos/BEI-200311/certificados_digitales.pdf

[20] Yu Lei, Deren Chen, and Zhongding Jiang, "Generating digital signaturas on Mobile devices," 18th International Conference on Advanced Inf. Networking and Applications, 2004. http://www.computer.org/portal/web/csdl/doi/10.1109/AINA.2004.1283860

[21] Santi Jarusombat and Surin Kittitornkun, "Digital Signature on Mobile Devices based on Location International," Symposium on Comm. and Information Technologies, 2006. ieeexplore.ieee.org/xpls/absall.jsp?arnumber=4141339&tag=1

[22] Scott Cambell, "Supporting digital in mobile environments," Twelfth IEEE International Workshops on Infrastructure for Collaborative Enterprises; 2003. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1231414

[23] Vision mobile, "The mobile developer journey," 2010. http://stackoverflow.com/questions/1414288/j2me-vs-android-vs-iphone-vs-symbian-vs-windows-ce

[24] Figure 4 and 5, Source: visionmobile.com, 2011.

[25] C. Enrique Ortiz, "*The Security and Trust Services API for J2ME," Part 1, March*, 2005. http://developers.sun.com/mobility/apis/articles/satsa1/

[26] Fundación Symbian**, "***Symbian," 2008*. http://www.xatakamovil.com/sistemas-operativos/fundacion-symbian

[27] Figure 6, 8, 9, 10, 11, 12, 13 and 14, Source: Own elaboration, 2011.

[28] Alessandro Distefano, A. Grillo, A. Lentini, Gianluigi Me, and Riccardo Galbani**, "***Communications in Computer and Information Science," 2009*. http://www.springerlink.com/content/l8n978414l371t88/

[29] Figure 7, Source: Universidad de Málaga – sicuma.uma.es, 2011.

[30] Javier García de Jalón, J. Ignacio Rodríguez, and Aitor Imaz **"***Los Java Servlet*," 2008. http://es.wikipedia.org/wiki/Java_Servlet

[31] Esigna Viewer, "*Visor gratuito,*" 2005. http://www.indenova.com/indenova.php?opc=5

 ISBN: 978-1-61208-164-9

SMARTPOS: Accurate and Precise Indoor Positioning on Mobile Phones

Moritz Kessel, Martin Werner
Mobile and Distributed Systems Group
Ludwig-Maximilians-University Munich
Munich, Germany
{*moritz.kessel,martin.werner*}*@ifi.lmu.de*

***Abstract*—Location-based services are possibly the most popular services with respect to mobility, since they allow for the automated filtering of information relevant to the user. This paper presents a detailed evaluation of SMARTPOS, an indoor positioning system based on deterministic 802.11 fingerprinting and a digital compass. SMARTPOS is accurate enough to supply location estimates for indoor location-based services and can be deployed standalone on a mobile phone. The system considers the user's orientation to avoid errors caused by the blocking effect of the human body. For location estimation it takes only that part of the fingerprint database into account that corresponds to the user's current orientation. SMARTPOS achieves a mean position error of 1.16 meters and a maximum position error of 2.74 meters in a 250 square meter environment.**

***Keywords*-Location Systems, 802.11 Fingerprinting, Mobile Phone Positioning, Location-Based Services.**

I. INTRODUCTION

In recent years, a trend towards mobility can be recognized. Smartphones, small devices with comparatively high processing power and mobile internet, make it possible to work while traveling, to stay connected to social networks, and to retrieve nearly any information anywhere at any time. One of the most popular mobile services are location-based services (LBS). These are value-added services, which utilize the location of the mobile to present the user with information about its surroundings. Navigation and information services, friend-finder, pet-tracker, and location-based games are only a small part of the number of services and applications filling the app-stores of the world.

The key enabler for LBS is the Global Positioning System (GPS) [1]. It enables accurate positioning in outdoor environments, the usage is free of charge, the system is globally available, and most of today's smartphones are equipped with a GPS-receiver. Unfortunately, GPS is not able to track people in indoor environments with acceptable accuracy. Signals might get lost due to attenuation effects of roofs and walls or lead to position fixes of very low accuracy due to multipath propagation.

Even worse, indoor location-based services require much higher precision guarantees than outdoor services. Errors should not exceed a few meters to allow for a differentiation between several floors or rooms. Otherwise, the service could provide information for places, which are quite far away from the actual position of the target. Despite these challenges many users would appreciate indoor location-based services, especially in large and complex buildings such as museums, shopping malls, airports, hospitals, or university buildings.

Existing indoor positioning techniques can be grouped by their level of precision and the expenses for additional infrastructure. Dedicated indoor positioning systems such as ultra wide band or ultrasonic systems consist of several components with the sole purpose of determining the positions of possibly multiple targets in indoor environments. The precision is often high, but an expensive infrastructure is needed and hence the space where positioning is possible is usually limited to a small area, where higher accuracy compensates the high cost. Another class of systems is built on existing infrastructure such as WLAN, Bluetooth or inertial sensors for positioning. The precision of such systems is limited, but the system can be deployed with few additional expenses.

In this paper, we present SMARTPOS, an indoor positioning system for smartphones based on deterministic WLAN fingerprinting and a digital compass. The system is self-positioning, meaning that the whole positioning process (including all measurements) is carried out on the phone. It achieves a high accuracy within few meters and therefore is able to provide indoor location-based services with high quality location estimates at no additional expenses. SMARTPOS makes use of the user's orientation to avoid errors caused by the blocking effect of the human body. Only those fingerprints are considered for location estimation that were measured while viewing in a similar direction like the user.

The remainder of this paper is structured as follows: In the next section, a short overview of existing indoor positioning systems is given. In Section III, SMARTPOS is presented in detail while in Section IV, the impact of several parameters is analyzed and discussed. Weighted and non-weighted kNN (k-nearest neighbors) in signal space, the influence of missing values on the algorithm and the performance gain of including the orientation on SMARTPOS and a Naive Bayesian Estimator are evaluated. Section V concludes the paper and gives hints on future work.

 ISBN: 978-1-61208-164-9

II. Related Work

In the past 15 years, a variety of technologies for indoor positioning have been proposed. A good overview of existing indoor positioning systems using radio frequency (RF) technologies such as radio frequency identification (RFID), ultra wide band (UWB), ultra high frequency (UHF), WLAN and Bluetooth is given in [2]. However, they do not describe up-to-date systems, which have been developed since 2007. We therefore focus in this section on the recent development and work closely related to our research.

Many state-of-the art systems rely on fingerprinting algorithms [3], [4], [5], [6]. These algorithms work in two phases: The first phase, the offline phase, is used to collect signal strength measurements (the fingerprints) from access points throughout the building at predefined reference positions. In the second phase, the online phase, the signal strength information is continually measured and compared to a database of all fingerprints from the offline phase. Different algorithms calculate the position as the nearest fingerprint in signal space, the average of the k-nearest neighbors with or without the distance in signal space as additional weight or utilize probabilistic methods. Localization techniques based on fingerprinting can be divided into two classes: deterministic and probabilistic techniques.

Deterministic systems compute the location estimate as a function of the received signal stregth (RSS) values measured using a physical model incorporating the values stored in the fingerprint database. One of the first systems working with WLAN fingerprints to retrieve a position estimate is the RADAR system [3]. RADAR is a deterministic system that utilizes kNN for position estimation and offers an optional signal propagation model for the automated creation of the fingerprint database. The authors noticed already the impact of the user's orientation and proposed obtaining empirical data for multiple orientations. Kaemarungsi et al. analyze the effects of the user's presence and orientation on RSS values in [7]. The results show that the attenuation effects of the human body can lower the RSS by more then 9dBm.

Probabilistic techniques [4] on the other hand compute a distribution based on the measurements from the offline phase and use probabilistic techniques to estimate the user's position. COMPASS [5] is one of the first probabilistic indoor positioning systems that adresses the problem of attenuation effects caused by the human body by adding a compass to the system. In the offline phase, fingerprints for several selected orientations (typically each $45°$ or $90°$) are collected at reference positions. In the online phase, the user's orientation is calculated by a digital compass and only the fingerprints with a similar orientation are used for the positioning algorithm. COMPASS presents the most similar approach to our system. However, we additionally analyze the impact of orientation information for deterministic techniques as well as for a bayesian approach. We also compare our results with a system not filtering the orientation information and thus benefiting from a much larger database. Chan et al. also present a system running on a mobile phone considering the orientation of the user in [8], but apply a technique called Newton Trust Region for further position refinement. Martin et al. present one of the first WLAN positioning systems, which integrates both offline and online phase on a mobile phone [9].

Most up-to-date systems combine WLAN fingerprinting with additional technologies such as inertial sensors to offer more accurate position estimates and continuous tracking functionality [10]. The authors utilize a particle filter for fusing WLAN fingerprint location estimates with an accelerometer.

III. SMARTPOS: A System for Self-Contained Mobile Positioning

In this section, we describe SMARTPOS, a system for an accurate and self-contained indoor positioning based on deterministic 802.11 fingerprinting and a digital compass. The system runs stand-alone on a mobile phone and consists of a management module for the creation and maintenance of the fingerprint database and a module for location determination. The latter offers the possibility of modifying several parameters concerning the deterministic location estimation or allows a change of the positioning method to a room-based bayesian approach.

A. Database Creation on a Mobile Phone

During the offline phase, active scans for WLAN signals from surrounding access points (APs) are executed with a mobile phone at several reference positions. The measured signal strength values are enhanced with the viewing direction and the pixel coordinates of the reference position on a bitmap of the floor. The viewing direction is obtained by the digital compass of the smartphone, the position is assigned by tapping on a zoomable and scrollable map displayed on the screen of the mobile. Finally these values (in the following referred to as fingerprints) are stored in a database. At each reference position, four fingerprints are created, one in the direction of each axis of the specific building. The alignment along the axes of the building instead of the geographic directions is carried out to improve the accuracy of the application in tracking scenarios since most users move along the main axes of a building, e.g., when walking down a corridor. For each fingerprint, five scans are executed and the average of the received signal strengths is stored in the database to reduce the impact of short-time fluctuations. Furthermore, the orientation of the phone, which is derived from the mobile phone's compass, is averaged throughout the sampling time and also stored in the database. This is done to remedy the disturbances of the magnetic field inside of buildings, especially near electronic sources or large amounts of metal.

 ISBN: 978-1-61208-164-9

B. Deterministic Location Estimation

During the online phase, SMARTPOS utilizes a deterministic positioning algorithm based on weighted kNN to estimate the approximate position of the user. WLAN signal strength measurements are carried out in a continuous fashion and for each measurement m the current orientation o of the phone is measured by its digital compass.

The orientation is considered to represent the approximate viewing direction of the user and hence implicitly yields the information about the attenuation of his body. The online RSS values should therefore not be compared to all fingerprints in the database due to possible influence of the human body, but only to those fingerprints that correspond to a similar viewing direction to o during the offline phase. Since the viewing direction is retrieved from the noisy readings of the compass, the orientation is averaged over the duration of each scan. This mechanism could also be replaced by advanced filtering algorithms to reduce the impact of outliers. SMARTPOS considers only a subset S of all fingerprints in the database containing those with a maximal deviation of $50°$ from o and is therefore able to reduce the number of fingerprints matched in the online phase to an extent of 25% of the database size.

On the remaining subset S of filtered fingerprints, the nearest neighbours in signal space with respect to m are computed. SMARTPOS uses a sophisticated distance metric for the comparison of two RSS measurements (i.e., the online measurement m and a fingerprint $f \in S$): Each measurement contains the information about all RSS values with the mac adress of the AP, which sent the signal. Since at a given position only signals of a subset of all access points in the building can be received, the question arises how to treat missing signal strength imformation in one of two compared measurements. One possibility would be to assign a fixed value MIN to the RSS of all access points missing in one measurement. This mechanism favors combinations of measurements, where signals by an AP are of very small strength in one measurement and missing in the other instead of combinations, where a high RSS value in one measurement is missing a counterpiece in the other. The value of MIN should be below the minimal RSS value measureable by the device. The other possibility is to ignore all signal strength information missing at least in one of the compared measurements. Based on the results of a detailed evaluation (see Section IV) SMARTPOS utilizes the second approach, which is expected to be more robust in the case a new AP is turned on or an existing AP is turned off.

Based on the Euclidean distance $d_i = dist(m, f_i)$ in signal space the subset $N \subset S$ of the k nearest neighbours is computed. In addition SMARTPOS assigns a weight w_i to each fingerprint $f_i \in N, i \in \{1, \ldots, k\}$ according to the following formula:

$$w_i = \left(d_i \sum_{j=1}^{k} \frac{1}{d_j} \right)^{-1} \quad (1)$$

It is easy to see that the w_i are normalized since $\sum_{i=1}^{k} w_i = 1$. For the computation of the user's position l, SMARTPOS calculates the weighted average of $l_i, i \in \{1, \ldots, k\}$, l_i being the reference position of the fingerprint f_i:

$$l = \sum_{i=1}^{k} l_i w_i \quad (2)$$

C. A Naive Bayesian Location Estimator

A Naive Bayesian Estimator is a simple and still very powerful classification scheme. The main ingredient is Bayes theorem, which is used to infer the probability $P(I|M)$ of an event I conditional on a measurement M. Using Bayes rule, we can turn over I and M and calculate from the probability of measuring M in the case that we are inside a given room I.

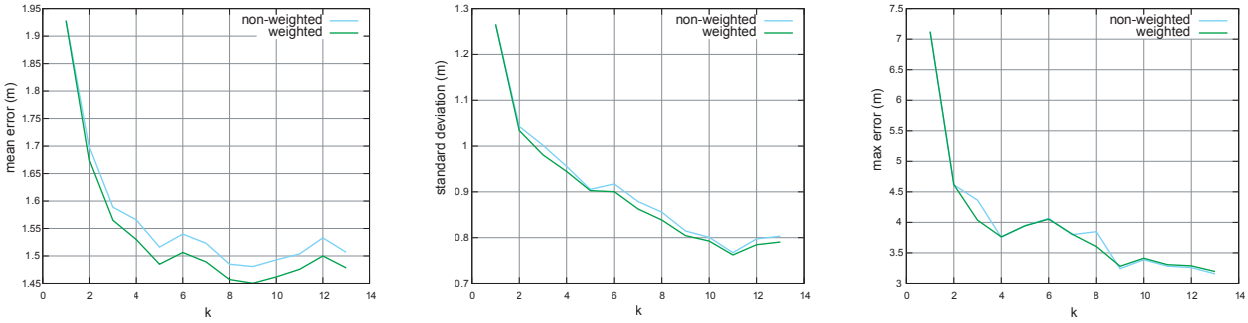
$$P(I|M) = \frac{P(M|I)P(I)}{P(M)} \quad (3)$$

The probabilities on the right hand side are estimated from a labelled set of instances simply by counting or calculating the mean and variance of each one-dimensional parameter and assuming a normal distribution. This trick assumes that the parameters are statistically independent. As this is usually not true, the performance of a classifier on a given problem has to be carefully estimated. It is common to use a method called cross-validation for measuring the quality of a classifier. This is done by splitting the training data and using a majority for training and holding back a minority to calculate a success rate on this test set. The details of how to do this and the basic caveats can be found in many textbooks on data mining.

IV. Evaluation

For the evaluation of our system, we created two sets of fingerprints in a part of our university building. All RSS information was gathered with a HTC Desire. The first set is arranged in an approximate grid of 79 reference positions with fingerprints measured in the direction of all four main axes of the building, which results in 316 fingerprints in total (the grey dots in Figure 1). The second set is a much smaller set of 64 fingerprints at 16 pseudo-randomly distributed reference positions (again measured in the direction of all four axes) within the coverage of the database and is used as substitution for online measurements (the black dots in Figure 1). This ensures that our results originate from an identical setting for all the different location estimators. The estimators are evaluated in respect to four criteria

 ISBN: 978-1-61208-164-9



(a) The mean error in meter (b) The standard deviation in meter (c) The maximum error in meter

Figure 2: Comparison of weighted and non-weighted kNN

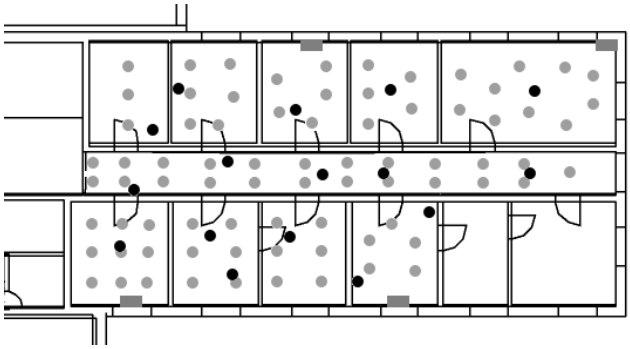


Figure 1: Reference database (gray dots) and online testset (black dots). APs are displayed as grey rectangles.

according to [2]: the accuracy as the mean position error, the precision as the maximal and the standard deviation, and the complexity as the number of compared fingerprints. The question of scalability, cost and robustness is not considered, since the scalability and the cost are the same in all systems and the robustness is hard to measure. In the following the results from a detailed evaluation of SMARTPOS in the described setting are presented and discussed. SMARTPOS is evaluated as follows: First the deterministic kNN approach is analyzed and the settings of several parameters compared to each other. The questions of assigning a weight to the nearest neighbors and whether missing signal strenght imformation should be considered or ignored are discussed and the impact of the user's orientation on accuracy and precision presented. In a consecutive step an optimal value for k is determined for SMARTPOS. Finally, the usage of orientation information in a Naive Bayesian Estimator is analyzed.

A. Weighted or Non-Weighted kNN

When using a kNN approach together with WLAN fingerprinting one has to decide whether just to compute the center of the nearest neighbors or to add a weight to each of the k-nearest neighbors according to the distance in signal space and then calculate the center of mass. With SMARTPOS, we evaluated both approaches for variable k. Figure 2 shows the results. The weighted approach behaves similarly, but performs better for each $k > 1$. The same applies for the deviation while the maximum error shows no significant difference except for two outliers ($k = 3$ and $k = 8$), for which the weighted approach also performs better. SMARTPOS therefore utilizes a weighted kNN as described in Section III-B.

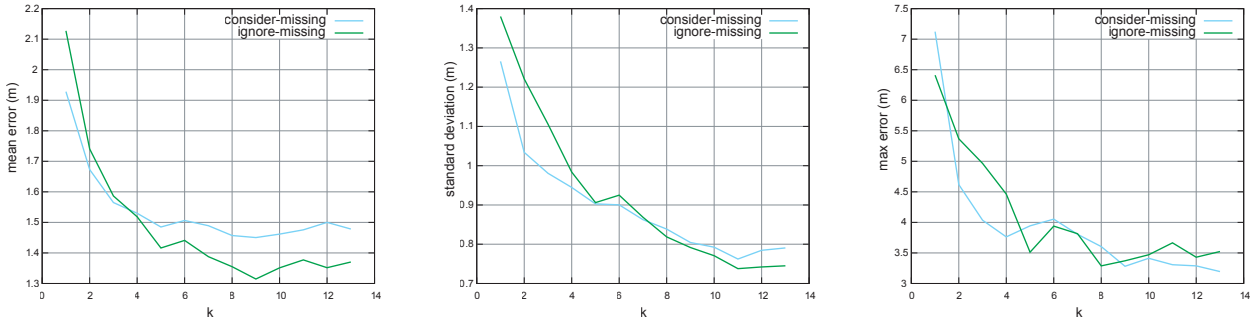
B. Treatment of Missing RSS

In Section III-B, two approaches for the treatment of missing signal strength information when comparing two RSS measurements are described. One considers the information by assigning a minimal value of -100dBm for the missing RSS information, the other ignores all RSS values from APs measured only in one of the two compared measurements. Both approaches were tested for a variable k and the results are presented in Figure 3. The accuracy of a system ignoring missing values is higher than the accuracy of a system considering the information for each $k > 3$ and also offers a minimum mean error for $k = 9$. The deviation only becomes smaller for each $k > 7$ with the minimum for $k = 11$, while the maximum error oscillates and therefore adds little information. Hence, SMARTPOS ignores missing RSS values.

C. Impact of Orientation Information

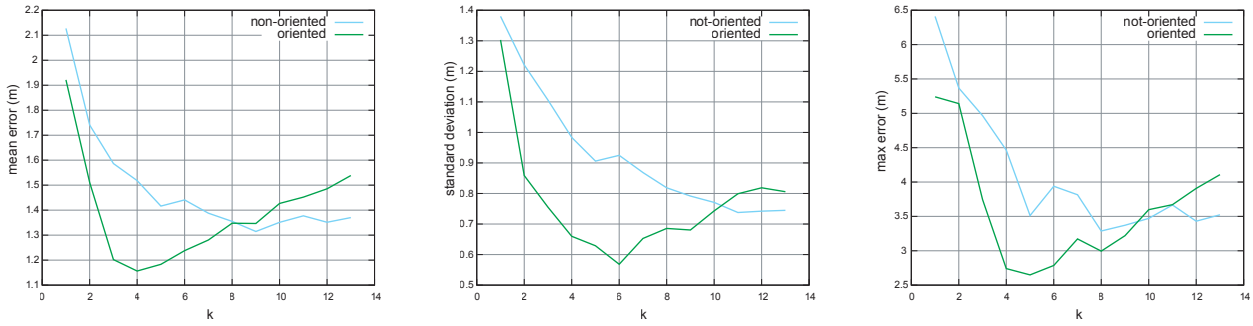
The most profound innovation of SMARTPOS is the usage of orientation information in a deterministic location estimation system on a smartphone. With the filtering of the fingerprints in the offline database with respect to the orientation information of the user, the compexity of the online matching can be quartered (when using the state of the art four directions for each reference position) and the accuracy and precision increased by a considerable amount. Figure 4 shows the results of the tests. The mean error is much smaller when using the orientation information and also reaches its minimum of 1.16m for $k = 4$, while the approach without orientation information reaches its minimum of 1.31m for $k = 9$. The minimal deviation of 0.57m for

 ISBN: 978-1-61208-164-9



(a) The mean error in meter (b) The standard deviation in meter (c) The maximum error in meter

Figure 3: Comparison of considering and ignoring missing RSS values



(a) The mean error in meter (b) The standard deviation in meter (c) The maximum error in meter

Figure 4: Comparison of considering and ignoring the user's orientation

$k = 6$ is also much smaller than the minimal deviation of 0.74m for $k = 11$ without considering the orientation. The same is true for the maximum error, which is minimal for $k = 5$ with a value of 2.65m when considering the user's orientation, whereas without the orientation information the minimum is 3.29m for $k = 8$. The much smaller number of k when using the orientation approach can be explained by the fact that the number of fingerprints for comparision is quartered and each online measurement has at most 4 neighbors in the grid, while without the filtering of the user's orientation the number of neighbors can increase to a total of 16 neighbors, because 4 fingerprints are stored for each reference position. In conclusion SMARTPOS utilizes the orientation information of the user to improve accuracy and precision of the location determination, while reducing the complexity at the same time.

D. Determination of k

Based on our experiments with SMARTPOS, we recommend utilizing an orientation-based weighted kNN approach with $k = 4$. For the comparison of measurements one should ignore all signal strength information of each AP missing at least in one of the measurements. With these parameters, the system offers the lowest mean error of 1.16m of all possible combinations with an acceptable deviation of 0.66m and a small maximum error of 2.74m.

E. Orientation and the Naive Bayesian Esitmator

The influence of filtering fingerprints according to their orientation on deterministic kNN positioning has been described. To get a deeper understanding of what influence the reduction of the search space according to the viewing direction has on indoor positioning, we chose to evaluate on the most simple (and often most effective) way of inducing a position from given measurements: Assuming that the variance in measurements is normally distributed, we estimate the mean and variance of a set of measurements taken in the same room and reuse this information for identification.

In order to do so, we assigned a label with each fingerprint specifying the room that it lies in. The long corridor has been cut into three rooms to reduce the variance of measurements in this long area as depicted in figure 5. Using this labeled data, we constructed a Bayesian Estimator, which calculates for each pair of access point and label the mean, standard deviation, weight sum and precision and reuses them for classification. We tested the classification performance with 10-fold stratified cross-validation training on 90% and evaluation on the remaining 10% of the data.

We used this technique on five different datasets: A dataset for each quadrant and a dataset where a random subset of 25% of all measurements in all directions were taken. In this

 ISBN: 978-1-61208-164-9

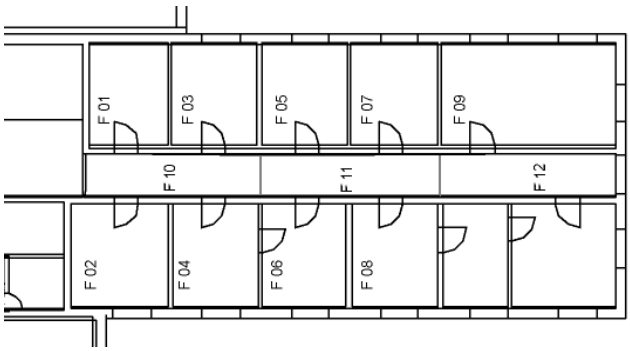


Figure 5: Labeled rooms for the Naive Bayesian Estimatior.

Table I: Evaluation results

Dataset	Number of Fingerprints	Success Rate
All directions	78	79%
North	72	62.5%
West	77	70.13%
East	82	65.85%
South	82	71.95%

way we achieve comparable training set sizes.

The results from this experiment are negative: A Bayesian classification of room-labels performs better on the total set of measurements than on the direction-dependent subsets. The results are given in Table I. Hence, for a system based on Bayesian estimation theory, we propose not to use the direction as a filter.

V. Conclusion and Future Work

In this paper, we presented SMARTPOS, a positioning system on a smartphone based on deterministic WLAN fingerprinting and a digital compass. SMARTPOS utilizes a weighted kNN approach with $k = 4$ and with a distance metric in signal strength space, which ignores RSS values from access points visible only at one fingerprint. We analyzed the impact of several parameters and conclude that a weighted approach results in more accurate and precise results than a non-weighted approach. Ignoring missing RSS values provides better results than assigning a minimal value, at least for higher values of k. In our setting this was the case for $k > 3$ in the oriented approach and for $k > 7$ in the approach without the user's orientation. With adding the user's orientation, SMARTPOS is able to reduce the mean positioning error to 1.16m and the variance to 0.66m. The maximal error in this case is 2.74m, which is 55cm smaller and therefore much better than the minimal maximum error of 3.29m in all experiments without the orientation information. We conclude that the user's orientation should be considered in deterministic 802.11 fingerprinting. However, we also discovered that the orientation information should not be used as a filter in a Naive Bayesian Estimator, since the percentage of correctly recognized rooms was smaller than that of the same algorithm trained with a similar large set of data containing fingerprints of all viewing directions.

In the near future, we want to expand the mechanism for filtering the database for faster access by including an accelerometer to the system. We hope that after an initial position fix we are able to further reduce the candidate set and can therefore support even large databases (e.g., at airports) standalone on the phone. Furthermore, we are currently working on mechanisms for a self calibrating system to replace the cumbersome process of keeping the fingerprint database up-to-date.

References

[1] E. Kaplan, *Understanding GPS: Principles and Applications*, ser. Artech House Mobile Communications. Artech House Publishers, 2006.

[2] H. Liu, H. Darabi, P. Banerjee, and J. Liu, "Survey of wireless indoor positioning techniques and systems," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 37, no. 6, pp. 1067–1080, 2007.

[3] P. Bahl and V. N. Padmanabhan, "RADAR: an in-building RF-based user location and tracking system," in *19th Annual Joint Conference of the IEEE Computer and Communications Societies*, ser. INFOCOM 2000, vol. 2, pp. 775–784.

[4] M. Youssef and A. Agrawala, "The horus wlan location determination system," in *3rd International Conference on Mobile Systems, Applications, and Services*, ser. MobiSys 2005, pp. 205–218.

[5] T. King, S. Kopf, T. Haenselmann, C. Lubberger, and W. Effelsberg, "Compass: A probabilistic indoor positioning system based on 802.11 and digital compasses," in *1st International Workshop on Wireless Network Testbeds, Experimental Evaluation and Characterization*, ser. WiNTECH 2006, pp. 34–40.

[6] A. Teuber and B. Eissfeller, "Wlan indoor positioning based on euclidean distances and fuzzy logic," in *3th Workshop on Positioning, Navigation and Communication*, ser. WPNC 2006, pp. 159–168.

[7] K. Kaemarungsi and P. Krishnamurthy, "Properties of indoor received signal strength for wlan location fingerprinting," in *1st Annual International Conference on Mobile and Ubiquitous Systems*, ser. MobiQuitous 2004, pp. 14–23.

[8] E. C. L. Chan, G. Baciu, and S. C. Mak, "Orientation-based wi-fi positioning on the google nexus one," in *6th International Conference on Wireless and Mobile Computing, Networking and Communications*, ser. WiMob 2010, pp. 392–397.

[9] E. Martin, O. Vinyals, G. Friedland, and R. Bajcsy, "Precise indoor localization using smart phones," in *International Conference on Multimedia*, ser. MM 2010, pp. 787–790.

[10] F. Evennou and F. Marx, "Advanced integration of wifi and inertial navigation systems for indoor mobile positioning," *EURASIP Journal on Advances in Signal Processing*, vol. 2006, pp. 1–11, 2006.

 ISBN: 978-1-61208-164-9

Balancing High-Load Scenarios with Next Cell Predictions and Mobility Pattern Recognition

Stefan Michaelis
Artificial Intelligence Group
Faculty of Computer Science
Dortmund University of Technology
Dortmund, Germany
Email: stefan.michaelis@tu-dortmund.de

***Abstract*—Knowing where a mobile user will be next can deliver a tremendous increase in network performance under high load, as this knowledge enables pro-active load balancing. To derive this information, sequences of traversed cells are fed into pattern detection algorithms. After the training phase the learned model predicts each user's next cell. Even for complex scenarios, the prediction accuracy can exceed 90%. Predictions are used to rearrange mobile connections in a simulated high-load scenario centered around an event at a soccer stadium. To prevent call drops for mobile users targeting the stadium, apropriate resources in the predicted next cell are reserved. The results exceed 20% in improvements for throughput and call drop rates, enabling the network to bear a much higher load before stalling.**

***Keywords*-Handoff Optimization; Mobility Prediction; Load Balancing**

I. The problem of users changing cells

Seamless handoff from basestation to basestation is essential for preserving mobility in cellular networks. Here we provide an additional indicator for handoff, which complements existing decision algorithms and can be used to manage overall mobile network load. The major advantage of this approach is the early availability of the handoff indicator, being in the range of several seconds compared to short-term measurements of signal strength and quality.

The idea is, that moving users are bound to the geographical topology, i.e., street and rail networks, and therefore are forced to partial deterministic behavior. Each movement provides a trail of traversed cells, which deliver a coverage fingerprint for the mobile network. Using knowledge discovery or data mining algorithms to learn the historical sequences of cells lead to a prediction of the most likely next cell each time a user enters a new cell.

This document consists of two main parts: In Section II, the overall achievable next-cell prediction accuracy is calculated for a sample geographical topology. The scenario demonstrates, how the artificial intelligence algorithm performs for varying road and railway networks, depending on the available input data for training the algorithm. The way the mobile data is handled, the privacy of the user's is respected and it is unnecessary to trace complete profiles on a per-user basis.

In Section III, the same methods are applied to a high load scenario: Mobile users moving to and from a soccer stadium.

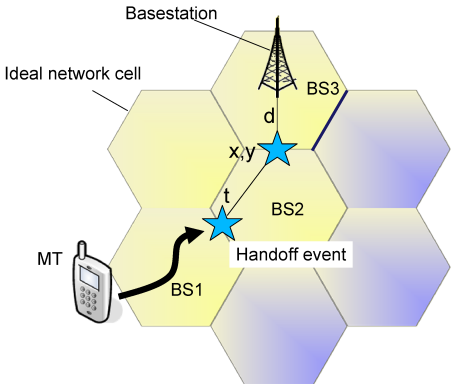


Fig. 1. Available features during user movements

The knowledge of the predicted next cell users are moving into is used to balance the load in the mobile network and enhance the user's quality of experience. The scenario is built using network coverage measurements, the underlying road and railway network, typical network traffic, and finally, the numbers of visitors for each means of transport. The results show the potential for gaining benefits by actively rearranging connections with knowledge about expected handoffs.

II. Space-domain prediction of next cells

Space-domain prediction of cells estimates *where* a mobile user will move next and identify the basestation candidate for the new association. In this section, the complete process from building mobility traces up to prediction of expected next cells with pattern detection algorithms is examined.

A. Mobility trace generation for pattern recognition

Every mobile cellular network needs a component, which is informed about the current location of the user based on the associated cell. Where this information is available depends on the type of network, e.g., mobile switching centers and location registers for mobile networks or remote authentication servers for wireless local area networks (WLAN). The known location is typically rather coarse and one reason why smartphones deploy GPS receivers for location based services.

 ISBN: 978-1-61208-164-9

Here we rely on the most common denominator independent of the specific network type, the basestation identifier. In addition, the duration t each user is connected to each basestation can be easily derived from the sequence of association/deassociation events. Further parameters may include position or distance to the basestation (see Figure 1), but are proved not to be mandatory for a good next cell prediction accuracy. An example sequence generated by a mobile user consisting of information tuples *BS-ID, Residence Time* may look like *BS1, 20s, BS2, 35s, BS4, 32s, ...* The length of these sequences is limited by the overall call duration, as mobile nodes in many networks can only be tracked during active connections or otherwise in large location areas bundling several cells. Very long sequences of traversed cells may occur for example when driving on vacation and the kids playing mobile online games during the trip on the backseat of the car. As very long sequences in most cases do not provide more information, all generated sequences are limited in length and split into shorter sequences. The optimal upper bound of length is also part of the analysis and results for sequence lengths between 3 and 6 cells are compared.

The set of all generated traces are used as training data for the pattern detection algorithm. The goal is to correctly predict the last basestation in each sequence.

B. Related research in cellular predictions

Predicting the next cell for moving mobile users has been in focus of mobile positioning research for several years. Macroscopic mobility prediction as discussed here sets the focus on the cellular level, which is useful in network load balancing.

In [3], a fundamental approach has been described for macromobility predictions: A variant of the ZIP-compression algorithm called LeZi is used to build a tree per user from the cell sequences. This algorithm delivers a good prediction accuracy for complete sequences, i.e., without missing values or changes in the cell sizes due to radio effects, and different variants are still popular today (see [7]) due to its simplicity and low consumption of computing resources.

The work on algorithms for mobility prediction can be classified into several categories as defined in [4]: Domain-specific, user dependent and usage of time.

In our previous work in [1], we demonstrated that the selection of the specific algorithm used for predicting the position is of secondary importance. While of course some algorithms may deliver higher accuracy compared to others, in most cases the question whether the mobility sequences contain learnable patterns of movements at all is more critical. Typically, general purpose data mining algorithms as the Support Vector Machine used in this publication are able to extract a minimum of patterns in the data if existent. Therefore we only investigate *domain-independent algorithms* without the need for mobile network specific parameters and keep the pattern detection algorithm replaceable

The feasibility of predictions per user profile has been demonstrated in several recent publications as [11], [12], in [6] for WLAN or in [8] with a prediction accuracy up to 93%. Nevertheless, learning individual movement patterns comes at the price of impacting privacy. The training data used for the predictions in these scenarios has every user identification removed, resulting in pattern detection *independent of specific user behavior*. Of course, approaches like this can only work in case the geographical topology restricts the users in their movement (e.g., on highways or in trains), so that meaningful patterns are generated.

Beside the spatial prediction, predicting the time of the handoff to the next cell is also necessary to reserve resources promptly. This timing can be integrated into the prediction model itself as demonstrated in [5]. This approach is reasonable if different points of time lead to different user behavior (e.g., weekend/weekdays, morning/evening). For short term changes, e.g., during traffic jam, incorporating time into the model increases the complexity of the training process. Our previous work in [2] presented an approach to deal with short term behavioral changes. For the work presented here we are *independent of absolute timing* and simply use cell residence time as a learning attribute.

C. Dynamic user-agent based mobility models

The feasibility of next-cell predictions strongly correlates with constraints moving users have to face due to the geographical topology, network coverage and most important the degree of determinism in the movements itself.

For this work several mobility models are combined to include different behavior. Essentially, these models are *Path Follower*, *Gravity* and *Random Walk* models. The path follower model can closely resemble commuting behavior: Following a preset path, staying at the target area for a certain amount of time and following a similar path back to the origin. This mobility model presents the highest level of determinism in the traces, introducing uncertainty only in variance of speed or residence times at target areas.

The gravity model assigns for different areas a so-called gravity value. This parameter sets a level of attractiveness to the areas, defining the probability for selecting this area as the target for movements.

Finally, the random walk model provides no determinism, but is still useful to generate a certain amount of *background noise* for the pattern recognition algorithm. Nevertheless, the random movement of course is still constrained by the road network, leading partly to the same traces as the other mobility model, e.g., on highways without a chance to leave at will. Random mobility is valuable to generate traces for areas, where the road density is high compared to the diameter of network cells, for example for GSM cells covering dense urban areas.

All mobile users are modelled as *Agents* without a fixed mobility model. This enables user traveling by car and switching to walking at the destination.

Figure 2 presents an exemplary scenario combining the road network, mobile network coverage and mobile user agents for simulation. The focus is put on situations where

 ISBN: 978-1-61208-164-9

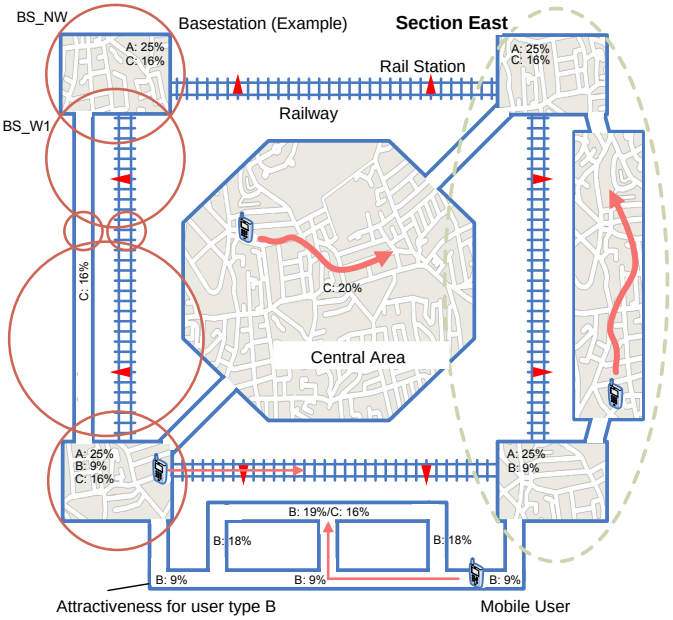


Fig. 2. Simulation scenario with different mobility models

mobile networks may easily receive high load: Highways and especially railways. Here a large amount of potential mobile application users switch cells nearly simultaneously and stay in a cell only for a short amount of time. In contrast, the rail network disables freedom of movement, forcing the users to certain sequences of cell transitions. The next section delivers results for the predictability of these sequences.

The geographical topology consists of several different areas: In the center one large area like an urban center, enabling random movements. This area is adjacent to the four areas to each side, introducing noisy patterns to the more regular streets and rails in these outer areas. Each of the four areas provides a different combination of possible user mobility: Rails only in the north, rails parallel to a simple street network in the south, to an area to the east and parallel to a highway in the west. Each line of rails also incorporates stations, where the simulated train stops for several seconds.

All mentioned mobility models are integrated into the scenario. For the gravity model the attractiveness distributions of user types $A - C$ is given. The topology is covered by overlapping cells, most of them are not show in Figure 2 for sake of simplicity. The western area as an exception shows some cells, as this part of the scenario is especially difficult for pattern recognition. Highway and railway users generate identical sequences of traversed cells except for two tiny cells individual to each path. A correct prediction for these cells is only possible in case the pattern detection algorithm can distinguish users on the parallel tracks.

D. Predicting next cells with pattern detection

The generated sequences are used to train pattern detection algorithms and predict the next cell (the target class) for new sequences. As classification of examples is a well-known task for pattern detection, several algorithms are available for this classification task. The more expressive the algorithm is, the better it can be adapted to complex traces, but the more difficult it is to find the optimal set of parameters for the algorithms. In parallel, the input data has to be selected carefully: What is the optimal maximum length for mobility sequences? Which features beside the basestation identifier enhance the pattern detection process?

For the results presented here the well-known *Support Vector Machine (SVM)* machine learner has been used for the prediction process, see [9]. SVMs try to separate the data samples by optimal hyperplanes and new examples are classified depending on which side of the hyperplane they are positioned. The hyperplane's location is defined by the so-called support vectors, which consist of a selected subset of all provided examples. The plane is considered optimal, if it minimizes the number of samples on the wrong side and maximizes the distance to the support vectors.

As a simple plane can not always capture the nature of data distributions, kernel functions allow to transform the input data into a modified space. A popular kernel is for example the polynomial kernel with the degree as a parameter. Selection of kernel and parameters like degree has to be done consistently for every example set.

To extend the SVM's ability to predict more than two possible classes (due to only comparing the side of the hyperplane of the example in question), the problem of multiple classes, as necessary for predicting the next cell, can be covered by pairwise predictions between each class.

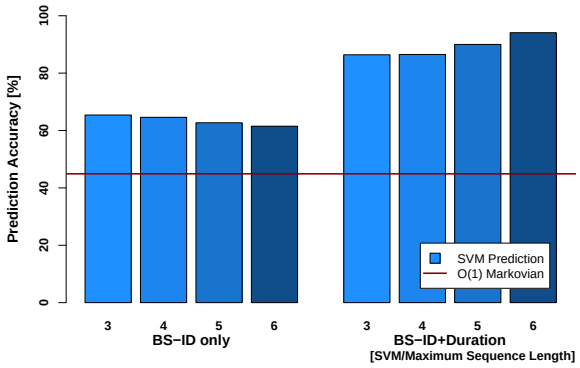


Fig. 3. Cell prediction accuracy for different features

Figure 3 presents the prediction accuracy as relation of correct to all predictions. Four different maximum sequence lengths 3-6 have been evaluated as well as using only the basestation identifier (on the left side) and identifier in combination with cell residence time (right side of the figure). The SVM can handle a combination of nominal data (BS-Id) and continuous data (time) without changing the algorithm. All results are generated using a ten-fold stratified cross-validation, delivering a 90% confidence interval in the range of ± 1.65.

The reference prediction (horizontal line) is calculated using a Markovian O(1)-predictor. This simple classifier uses only

 ISBN: 978-1-61208-164-9

the currently associated basestation as input and predicts the next which occurred most frequently in the training examples. The O(1)-predictor therefore delivers an estimation of the learning complexity, with more neighboring basestations and a uniform transition probability resulting in a lower accuracy.

The results in Figure 3 show for the id-only case an accuracy of around 65%, which is 20% higher compared to the O(1)-predictor, but still not sufficient for reliable enhancements of handoff and network management. For longer sequences of up to six cells the accuracy even slightly decreases. Effects like this appear in cases, where the added data masks the valuable bits of information provided by the rest of the features (here the higher information value of the latest basestation compared to previous ones.

A great boost in prediction accuracy can be seen for the second evaluation, BS-Id with residence times. Users traveling by car provide different patterns compared to users traveling by rail. Using the duration in each cell these users become separable, increasing the prediction accuracy up to 94%. Here the predictions even benefit from longer sequences, as the likelihood of identifying a user's means of travel increases with more durations available.

		Predicted cell				
SW	W3	W1	NW	W2b	W2a	Real
0.81	0.00	0.00	0.00	0.00	0.00	SW
0.02	**0.79**	0.01	0.00	0.00	0.00	W3
0.00	0.00	**0.93**	0.01	0.02	0.01	W1
0.00	0.04	0.02	**0.94**	0.00	0.00	NW
0.01	0.00	0.00	0.01	**0.56**	0.56	W2b
0.01	0.00	0.00	0.01	0.42	**0.43**	W2a

TABLE I
CONFUSION MATRIX, WEST SIDE OF SCENARIO

		Predicted cell				
SW	W3	W1	NW	W2b	W2a	Real
0.99	0.01	0.00	0.00	0.00	0.00	SW
0.00	**0.99**	0.01	0.00	0.00	0.00	W3
0.00	0.00	**0.99**	0.00	0.00	0.00	W1
0.00	0.00	0.00	**1.00**	0.00	0.00	NW
0.00	0.00	0.00	0.00	**1.00**	0.00	W2b
0.00	0.00	0.00	0.00	0.00	**1.00**	W2a

TABLE II
CONFUSION MATRIX, WEST SIDE, INCLUDING RESIDENCE TIME

Tables I and II enable a detailed comparison of this effect per basestation for the western part of the scenario. Cells with Ids *W2a* and *W2b* are the small cells distinct for highway and railway. Table I presents an overall good accuracy for most cells with the exception of these two cells (56% and 43%). Including the duration needed to cross the cells into the training data increases the accuracy for all cells and enables perfect predictions of W2a and W2b. The duration enables to distinguish users without any further information like GPS positions, knowing in advance which of the cells will be next.

III. BALANCING HIGH-LOAD SCENARIOS

This section applies the next cell predictions of the former section to balance network load in the mobile network itself. The early knowledge about users entering a new cell delivers a convenient time frame for reservation of resources.

A. Scenario description: Soccer stadium

The scenario used for evaluating the effect of predictions is based on the same principles as the scenario presented before and incorporates a real geographical topology, network coverage measurements and user movement profiles.

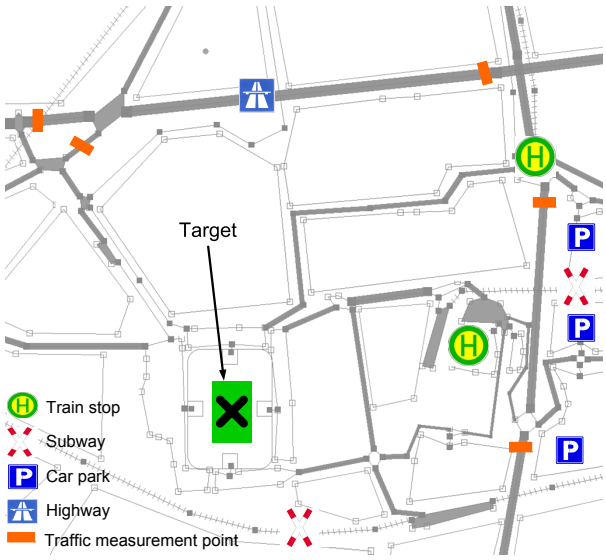


Fig. 4. Topology for mobility simulation of the stadium scenario

Figure 4 illustrates the scenario: The central point of interest is the soccer stadium in Dortmund, a German city with more than 500,000 residents. During an event at the stadium more than 60,000 people are arriving and leaving the stadium; 20,000 people unrelated to the event are expected to move in the region. Data provided by the local Department for Traffic, the regional transport and the stadium operator enables detailed modelling of the movement behaviors. The floating car data is measured using sensors in the streets and have been provided for several days, with and without events at the stadium to calculate the difference in paths and car density. The distribution of visitors arriving by train, car and foot determine the parameterization of the simulated agents, which are again able to switch mobility models. Visitors arriving by car change to a walk model after arriving at the parking sites etc.

For later evaluation two main paths have been selected: At the northern top the urban high way *B1* crosses the scenario from east to west. Secondly, a railway track from north-west to south-east provides one main access route to the stadium.

Together with user movements, the traversed basestations need to be captured. To gain a realistic view of the coverage, the basestations in range have been measured. Figure 5 displays results of measurements by car and foot. Each measurement has been associated with GPS positions and

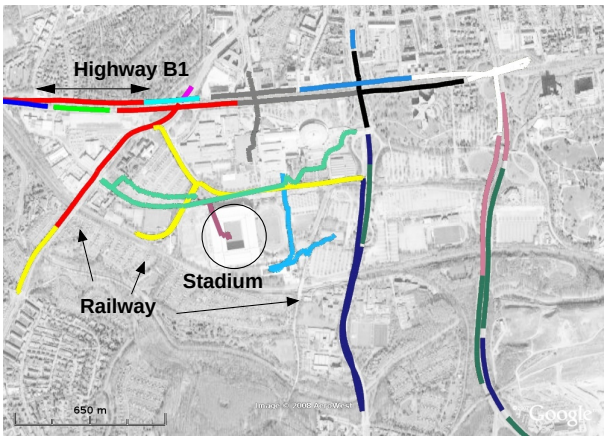


Fig. 5. Coverage measurements for stadium scenario

shows the associated primary UMTS-HSPA basestation. At most positions an active set of 4 was available, showing a high overlap of adjacent cells. This is a necessary precondition to enable rearrangement of connections into neighbor cells.

Interestingly, the measurements highlight a classical handoff parameter, the handover margin. The position of handover is shifted due to this margin depending on the direction of travel, as can be seen for the B1 at the north of the figure.

B. Mobile network management

This section concentrates on applying the next-cell predictions for different dynamic network management techniques like reserving radio resources for expected users or rearranging existing connections to maximize data throughput.

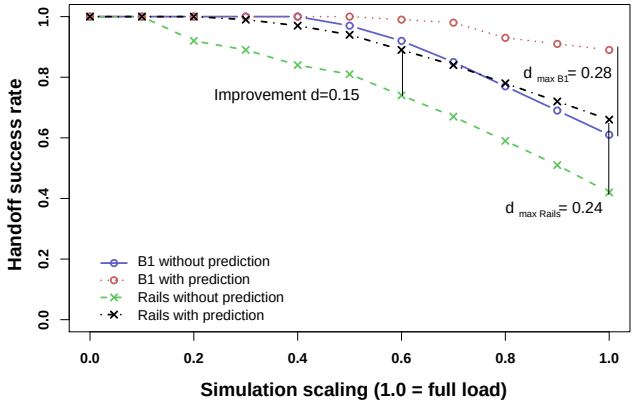


Fig. 6. Handoff success rate based on scenario load

Figure 6 presents the enhancements for handoff success rate. According to data provided by the German mobile network operators, the aggregated data traffic exceeds 1,000 Erlang voice calls equivalent during the event in and in direct neighborhood to the stadium. In case a user with an active connection gets into a cell without free resources, the connection has to be dropped. When a user successfully enters a new cell, the next cell is predicted and the resources for this user are blocked in the expected cell.

To examine the effects for different load situations, the simulation has been scaled for different percentages up to the full load simulating all 80,000 mobile users. Please be aware, that a scaling factor of 0 includes still one user for each mobility model.

As to expect, for a small load scale, the handoff is equal to or nearly 100%, as no cell is completely filled with connections. The success rate starts to decline with increasing load. Figure 6 displays the success rate for two paths, B1 and railway, and for two modes: With and without using the predictions. The success rate declines faster for users arriving by train, as these users travel faster and in higher numbers, increasing the probability for arriving at resource depleted cells.

Reserving resources can not completely avoid this effect, but significantly improve the success rate. The decline is slowed and the improvement of handoff success can go up by 28% for the fully loaded scenario.

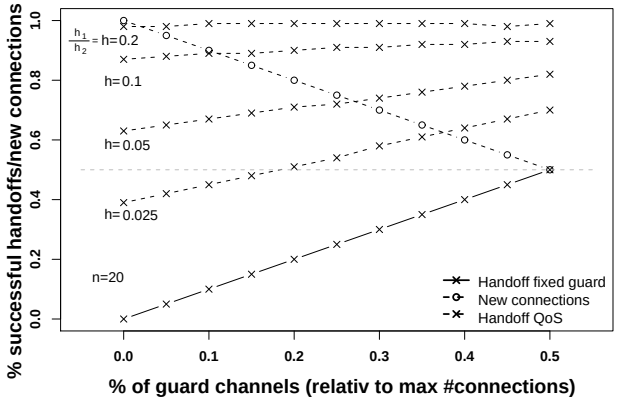


Fig. 7. Handoff success rate based on call holding time

Whether a reservation based on the prediction can be successfully executed, depends on the holding time of connections. Mobile operators report an ever increasing utilization, starting with one minute mean call duration for voice calls and two minutes for video calls in pre-iPhone times. Figure 7 presents the handoff-success rate based on mean holding times and a pre-reserved guard channel. Most network operators prioritize active over new connections and set a fixed amount of cell capacity for users arriving in the cell, decreasing the amount of capacity for new connections accordingly.

Users targeting a fully loaded cell get a reservation in case for a prediction (Handoff QoS), when an existing connection is closed or the guard channels are not completely used. The success rate therefore depends on ratio of the mean cell transient time h_1 of the moving users and the mean holding time h_2 of the resident users in the target cell, $h = \frac{h_1}{h_2}$.

Figure 7 presents the results for a fixed h_1 and varying h_2 of $n = 20$ users resident to the cell. For large ratios $h \geq 0.2$ nearly all handoffs can be handled perfectly without the need for fixed guard channels. For increased holding times and smaller ratios $h = 0.025$, the probability of ending connections in the target cell drops and the success rate is below 40%.

 ISBN: 978-1-61208-164-9

This can be compensated with the classical guard channels. Nevertheless, using the predictions for channel reservation, a smaller fixed guard channel is needed to achieve the same rate of successful handoffs.

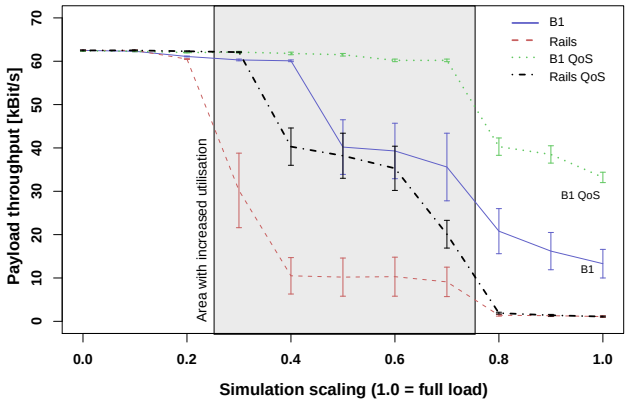


Fig. 8. Network throughput with rearrangement of connections

Lastly, the results presented here further enhance network management by rearranging active connections to neighboring cells for predicted incoming connections to the cell. This handles situations, where throughput can be maximized by distributing the traffic more evenly. Figure 8 illustrates the effect again for two sample users with a reference bitrate of 64 kBit/s TCP traffic. Moving into high-load cells, especially with mixed traffic types and the need to compete with UDP, degrades the mean throughput to 0 kBit/s for the highest load in the scenario.

As for the handoff success rate, the throughput starts to decline with increased load. The error bars for each point present the median absolute deviation, increasing with higher variance and bias in the data transmission. Again, a higher amount of Quality of Service can be guaranteed by using the predictions and moving existing connections from predicted next cells. Nevertheless, in the end with full load in the scenario, the TCP connection looses against other traffic source like UDP. This leaves an area between the two extremes of underused and exhausted network, where the rearrangement balances network load and improves QoS for all mobile users.

IV. The Benefit of Users Changing Cells

Instead of compensating for the effects of moving users on the network routing, we actively use knowledge about previously visited cells to predict the next location and balance traffic load accordingly.The results demonstrate, that even with simple features like identifiers and cell residence times the geographical constraints mobile users face can be detected.

Predicting user's next cell with a high accuracy of more than 90% provides mobile network operators with a powerful tool to rearrange traffic. This enhances quality of service for the users as well as saving costs for operators due to more efficient utilization of infrastructure. The approach is non-intrusive and intended to co-exist with mandatory network management for handoff, call admission control and routing. User's privacy is preserved as no individual patterns need to be learned. The only time where the user's id can be associated with the sequence of basestations is when preparing for predicting the next cell. In the future, further methods to make users anonymous like proposed in [10] may enable to provide the data to external location based service providers without breaching privacy.

The final step, before the methods proposed here are considered ready for production use, relates to the selection of subsets of cells for model training. As an example, for the region of Dortmund for the combined network types from GSM to 3G, including sectorization, more than 500 cell ids can be measured. For the whole country this will result in an amount of cells too large for most data mining algorithms. Future research concentrates on distributed data mining for automatically generated clusters of cells.

Acknowledgments

Part of the work in this paper is supported by Deutsche Forschungsgemeinschaft (DFG) within the Collaborative Research Center SFB 876 "Providing Information by Resource-Constrained Analysis", A1: http://sfb876.tu-dortmund.de

References

[1] Ms. ichaelis, A. Lewandowski, K. Daniel, and C. Wietfeld: *Macromobility Prognosis for high-priority Resource Reservation in Wireless Networks*, 5th IEEE International Symposium on Wireless Communication Systems (ISWCS), Reykjavik 2008

[2] S. Michaelis, A. Lewandowski, K. Daniel, F.Z. Yousaf, and C. Wietfeld: *A comprehensive mobility management solution for handling peak load in cellular network scenarios*, 6th ACM International Symposium on Mobility Management and Wireless Access (MobiWac), Vancouver 2008

[3] A. Bhattacharya, and S.K. Das: *LeZi-Update: An Information-Theoretic Approach to Track Mobile Users in PCS Networks*, Proceedings of ACM/IEEE International Conference on Mobile Computing and Networking, MobiCom '99, Seattle 1999

[4] C. Cheng, R. Jain, and E.v.d. Berg: *Location Prediction for Mobile Wireless Systems*, in: Furht, B. (Editor): *Wireless Internet Handbook*, pp. 245-264, CRC Press, Boca Raton 2003

[5] J.-M. François, G. Leduc, and S. Martin: *Learning movement patterns in mobile networks: a generic method*, Proceedings of European Wireless, Barcelona 2004

[6] L. Song, D. Kotz, and R. Jain: *Evaluating Next-Cell Predictors with Extensive Wi-Fi Mobility Data*, Transactions on Mobile Computing, Vol. 5, No. 12, pp. 1633-1650, IEEE 2006

[7] A. Rodriguez-Carrion, C. Garcia-Rubio, and C. Campo: *Performance Evaluation of LZ-Based Location Prediction Algorithms in Cellular Networks*, in IEEE Communications Letters, Volume 14 Issue:8, pp. 707-709, 2010

[8] C. Song, Z. Qu, N. Blumm, and A.-L. Barab'asi: *Limits of Predictability in Human Mobility*, Science Vol. 327 no. 5968 pp. 1018-1021, 2010

[9] J. Platt: *Fast Training of Support Vector Machines using Sequential Minimal Optimization*, in Advances in Kernel Methods - Support Vector Learning, B. Schoelkopf, C. Burges, A. Smola (Editoren), MIT Press, Cambridge 1998

[10] A. Monreal, R. Trasarti, C. Renso, D. Predreschi, and V. Bogorny: *Preserving privacy in semantic-rich trajectories of human mobility*, Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Security and Privacy in GIS and LBS, San Jose 2010

[11] P. Salvador, and A. Nogueira: *Markov Modulated Bi-variate Gaussian Processes for Mobility Modeling and Location Prediction*, in Proceedings of NETWORKING 2011, LNCS, Volume 6640/2011, pp. 227-240, Springer Berlin 2011

[12] N. Zhao, W. Huang, G. Song, and K. Xie: *Discrete Trajectory Prediction on Mobile Data*, in Proceedings of Web Technologies and Applications 2001, LNCS, Volume 6612/2011, pp. 77-88, Springer Berlin 2011

 ISBN: 978-1-61208-164-9

Real-time Cognitive-Capacity-Sensitive Multimodal Information Exchange for the Cockpit Environment

Atta Badii, Ali Khan
Intelligent Systems Research Laboratory,
School of Systems Engineering, University of Reading,
Reading RG6 6AY United Kingdom
atta.badii@reading.ac.uk, a.a.khan@reading.ac.uk

Abstract – **Deployment of multimodal interfaces in an environment of high cognitive-load, e.g., the cockpit environment of a police helicopter or motorcyclist in-pursuit, using audio communication with the need to frequently change radio channel or mode/focus of attention under time-constraints, whilst multi-tasking, can impose an extra load on the driver in addition to that due to the complexity of vehicle manoeuvring. This can raise concerns regarding driver safety. There is a need for a multi-sensor, multimodal information exchange interface that is minimally distracting by design, with consideration for the cognitive capacity of the driver at any given point in time. This paper presents the MoveON Jacket prototype with a Cognitive Capacity Respecting Multimodal Information Exchange framework that offers a minimal-distraction interface to motorcyclists enabling access to information online in real-time while supporting enhanced compliance with road safety constraints. This project has delivered the first validated implementation of an architecture for: a) minimal-distraction, b) cognitive capacity awareness, c) multimodal communication and control, d) graceful in-line man-in-the-loop cognitive control integration, e) performance improvement context caching to support an off-line learning capability for incrementally optimised man-machine mixed-initiative taking, f) performance enhancement and personalised ambient driver's pal adaptation.**

Keywords – multimodal command and control; cognitive-load; minimal distraction; cognitive control; mixed initiative taking; man-in-the-loop; machine learning

I. INTRODUCTION

While the level of sophistication of systems on-board vehicles has steadily progressed in recent years, provision of similar functions has not relatively matured for motorcycles and other two-wheeled vehicles. The main reason underlying this lack in provision can be related to safety concerns, as the manoeuvre context of a motorcycle is more complex than that of four-wheeled vehicle. Safety of an individual moving on the road on a two-wheeled vehicle is of paramount importance. This manoeuvre complexity is further increased in the context of on-board features usage, and communication with entities external to the immediate environment of the individual, e.g., a central command and control station as in the case of police motorcyclists. Therefore, any provision of added-value system features on-board two-wheel vehicles needs to consider potential compromises on safety. Thus, a major requirement of any such development is a minimally distracting unobtrusive interface that is sensitive to, and keeps track of the cognitive load of the user at any given point in time. A multimodal, multi-sensor, minimally distracting interaction system is required for motorcyclists as at any arbitrary point of time they may not able to interact readily and safely through a visual or tactile interface. For such a system to be sensibly deployed it must be capable of gracefully receiving training on the job (in-line) without this imposing too high a cognitive load on the trainer (the user, the driver/rider in the vehicular control environment). Such an ambient Communications Command and Control assistant (C^3 Pal) will have to take experientially -based best-estimate control actions on the driver's behalf and be ready to stand corrected by the driver if mistaken and learn from the experience to get it right next time. Such man-machine mixed initiative systems require situated context caching at man-machine control handover points so as to support offline learning for enhanced adaptation to the user's requirements during man-machine team-working real-time.

In the following sections, background domain knowledge is presented in Section II. Section III details the MoveON project, whereas Section IV focuses on the cognitive capacity sensitive multimodal information exchange framework. In Section V, evaluation of the prototype is discussed, and Section VI concludes this paper.

II. MINIMISING DISTRACTION

Cognitive load is defined as the total of ongoing mental activity, at any given point in time, in the working memory – a major contributor to which is the number of attention seeking factors [10]. Cognitive load theory [11][12] states that the quality of the design of instruction is accentuated if the limitations of working memory are given careful consideration. Long term memory stores knowledge and skills while working memory performs computational tasks

 ISBN: 978-1-61208-164-9

associated with consciousness. Information is first processed by working memory, which is very limited in terms of capacity and duration, and then may be stored in long term memory. Cognitive load theory is employed in the development of learning structures or instructional strategies to circumvent any learning impedance imposed by the limitations of working memory in conventional strategies [10]. Of the three components of the cognitive load, intrinsic load represents the inherent difficulty of the content; extraneous load may be introduced by the designer of the instructional materials; germane load represents the effort involved in the processing of content and is often associated with motivation and interest. In the context of minimising distraction for motorcyclists, cognitive load theory can inform the design of the communication interface for audio-visual and tactile information that may be relayed to the rider in real-time on the road.

Research in the measurement of cognitive load presents several constructs; for instance, the relative condition efficiency [13]. Availability of usable cognitive load measures is crucial for providing support for minimal distraction interfaces. Relatively simple subjective measures are in most frequent use; however more sophisticated methods also exist for multimodal environments to gauge the relative complexity of cognitive tasks, such as empirical approaches to cognitive load measurement, which can be divided into *i)* direct object measures, e.g., eye tracking techniques, brain activity measures, dual-task methodology; *ii)* indirect objective measures, e.g., physiological correlations (cardiovascular, EEG etc.); *iii)* direct subjective measures, e.g., self-reporting of stress level; *iv)* indirect subjective measures, e.g., self-reporting of perceived mental effort [6].

Niculescu et al. [1] studied the impact of stress and cognitive load on the perceived quality of interaction in the context of a multimodal dialogue system for crisis management, using physiological sensors and subjective measures including an evaluation questionnaire regarding the quality of system interaction, an interview, and video recordings of trials to perform behaviour analysis. Niculescu et al. report that both stress and cognitive load impact the subjects' perception of the quality of interaction with the system.

A. MoveON Project

The MoveON project [3] has developed an innovative multimodal multi-sensor system to support a distraction minimising communication control architecture using multiple modalities (tactile visual and auditory) in full duplex mode. The MoveON system can provide motorcyclists access to services and information online in real-time while attempting to protect the driver from experiencing excessive levels of cognitive load and therefore, unsafe levels of distraction arising from the perceptual-cognitive task environment including the manoeuvring of the motorcycle plus handling multiple communication channels; some in full duplex. MoveON employs a multimodal interface with the capacity for presentation of visual information via a sleeve touch screen, text-to-vibration functionality and speech recognition.

Figure 1. MoveON Jacket Interface

Information gathered by multiple sensors is used to train the Cognitive Capacity Management component that continuously assesses the user's level of distraction so as to channel information in a maximally safe and timely manner. Usability of the demonstrated prototype has been evaluated by applying a socio-technical Usability Relationship Evaluation Methodology [8].

A specialised user group of UK police motorcyclists participated in a user-centred system co-design and evaluation to assess user's distraction while interacting with the prototype system, and, identifying driving conditions in which the motorcyclist's attention is disrupted, thus, causing unsafe levels of distraction and endangering safety.

1) The MoveON Jacket

The MoveON system is fully wearable with all its components on a body area network and not on the bike. The MoveON environment comprises of a jacket interface, helmet interface and the motorcycle. The jacket is a classic motorcycle, jacket augmented with assistive electronic components. These include a microprocessor (VIA Pico-ITX housed in a Travla AnkerPC casing) affixed to the left posterior side of the jacket. A GPS device resides on the inside of the jacket near the right arm side-deltoid. This device connects to GPS satellite to obtain the geo-position of the motorcyclist.

2) Sleeve Screen

The sleeve screen (LCD touch screen) is one of the major input and output modalities of the MoveON system. Situated on the left forearm, it provides an interface between the system and the motorcyclist to call the various commands of the system or to acknowledge receipt of information from a command and control station.

The sleeve screen GUI is designed with a touch-screen smart phone / PDA interface in mind, where the user is able to interact with the system without a keypad / keyboard, using gestures (with either a stylus or finger). The buttons on the GUI are designed so that the officer can use the interface whilst wearing motorcycle gloves. The menu, as shown in Figure 2, presents various options to the user at any given time. The list of options is scrollable with either a

 ISBN: 978-1-61208-164-9

flick gesture, whereby the list scrolls automatically with a velocity calculated from the flick gesture, or a press-down hold-and-drag gesture whereby the scrolling is controlled by the user. At any point during a flick scroll, the user may touch the screen to stop scrolling at a particular point to highlight and select an option by just tapping on it once.

The GUI is a generic container with the functionality to add/remove various listed options. This functionality is used by the Olympus dialog management system [9] to present options to the user depending on what has been selected by the user. A menu-driven command mode takes the user through three steps for device selection (e.g., radio), command selection (e.g., change channel to …) followed by parameter selection (e.g., 9).

The touch driver circuit and the main LCD display have been integrated so as to make one single module. The VGA and power supply board are fitted at the back of the jacket but housed inside the lining of the jacket so as to preserve the normal way that the jacket is worn. The main battery powers the sleeve screen as well as the CPU. The power supply cable comes from the main battery through the left arm / sleeve (inside the lining). An additional screen (dot-matrix) functions purely as a debugging facility hidden inside the jacket, attached to a micro-controller that supports the functionality of the vibration motor array.

3) Vibration Motor Array

The vibration feedback motors are asymmetric load rotary motors that give a powerful vibration output. The choice of vibration sensors was improved after the first user evaluation process whereupon it was highlighted that the vibration from the motorcycle engine muffled the vibration output from Precision 10mm shaft-less sensors i.e., submerging their vibration and making it difficult to be sensed by the rider. Two motors were attached to the driver's shoulders and the third just above the right elbow on the sleeve of the jacket. These motors provide for the MoveON text-to-vibration (ttv) functionality, controlled by an Olympus agent that supports four simple commands for the motorcyclist in a policing scenario. These commands correspond to a simple tactile vocabulary requiring only minor user-training. For this, initially, seven motors were strapped to the right arm of the motorcyclist, but user evaluation highlighted two salient issues: *a)* tactile vocabulary was too large, requiring a wider pattern of vibrations to be comfortably and reliably recognised by the motorcyclist rider under various conditions, *b)* reliable sensing of the orchestrated vibration pattern for each word was made difficult by ambient conditions such as swamping of the messaging vibration by continuous high amplitude motorcycle vibration that was also transmitted to the rider's arms through the handle bar, and other ambient detractors such as noise, wind and rain..

Figure 2. MoveON Sleeve Screen Interface

User evaluation informed optimisation of the physical design, vocabulary size limits, and vibration frequency for maximum user comfort.

The motors were powered by a 9V battery that was placed within the inside pocket of the jacket. A voltage regulator IC maintained the 5V required by each rotary motor. Owing to the level of power consumption of the motors, a decision to use a separate power supply for the batteries was taken. The main battery powered the CPU and other devices on the jacket. To further ensure circumvention of a single-point-of-failure, a back-up energy source was also installed in the jacket.

The jacket was connected to the helmet with a single cable that incorporated the required connection points for all devices such as a helmet camera (USB), helmet LEDs array (USB), microphones and headphones. The energy source was in the jacket so the electronic devices in the helmet were powered by the wired connection between the jacket and the helmet.

III. COGNITIVE CAPACITY SENSITIVE MULTIMODAL INFORMATION EXCHANGE

It is proposed that the user's vulnerability to stress, in a given context of stress load, affects their perceived stress levels. These, in turn, affect the perceived cognitive load and thus, the available safely 'grab-able' cognitive headroom that could be deployed to *minimally* (i.e., safely) distract but *maximally* inform an individual (*just-in-time*) whilst engaged on the move. The motorcyclist's operational context is one of the most significant determinants of the relevant situation assessment as to the relative significance and timing of various messages that the rider may have to be updated on with mission-critical timing. Typically, the police motorcyclist's residual stress level, at any instant, during a shift is expected, mainly, to have six correlations: *i)* update value of rider's baseline stress; *ii)* stress index of events already experienced during the shift; *iii)* time elapsed since the start of the shift; *iv)* the stress levels induced by instantaneous ambient conditions experienced by the rider (e.g., current task, traffic conditions, weather, road surface, communications messaging intensity and noise); *v)* stress due to unknown risks that may be waiting for the police motorcyclist en-route to an incident to which he is called (unknown risk anticipation stress); *vi)* the rider's stress vulnerability profile in the context of *i–v* above; i.e., the individual's stress management capability, which is their learned or genetic endowed ability to cope with stress. Such a context-specific stress vulnerability profile of the

 ISBN: 978-1-61208-164-9

individual is in turn influenced by the interplay of a spectrum of associated factors, e.g., idiosyncratic, and other deterministic influences relating to the Life-Course Perspective [4] and a whole host of psycho-physiological, experiential and operational factors (historical, static, dynamic and current) spanning over time from birth to the present point in the spectral context (i.e., now, which are the day, time, place, space and point-of-decision/execution in the individual's task flow). The safely distracting informative and timely messaging of the rider has to be optimised to allow for the uncertainty in assessment of the instantaneous residual stress level and thus, the instantaneously disposable level of cognitive capacity headroom that the rider can, safely, be expected to make available for the next messaging input/output as may be optimally allowed to occur as decided by the multimodal interaction controller. The available cognitive headroom is inversely proportional to the perceived cognitive load, which is directly influenced by the perceived (residual) stress level as instantaneously experienced by the individual; which, in turn, is influenced by the individual's stress vulnerability profile. Minimising distraction can be viewed as maximising safe distraction, in an approach whereby multimodality control is based on seeking safe distraction or opportunities for grabbing safely available disposable cognitive capacity i.e., grabbing the attention of the driver for passing messages that wait in the queue to be conveyed to the driver. Paradigmatically this depends on the level of available and disposable cognitive capacity (stress levels), and requires the following constraining factors to be resolved and satisfied as part of the cognitive load resolution and optimisation.

A. Operational Context

In the context of the MoveON application domain, i.e., in the given practice context, stress level is a temporary psycho-physiological variant that can be assessed by combining the effects of

- the attention required for driving the motorcycle
- road / traffic conditions, level of activity in surroundings, GPS information indicating the present location of the driver and thus, the traffic context, e.g., approaching or passing through a junction, joining a motorway or leaving it etc.
- pre-existing communication level of the user with the MoveON system (thus, indicating the cognitive load arising from the user communication channels that are currently active).
- elapsed time on current shift (and indirectly cognitive load attendant with the Shift Envelope described as the range of experiences, from extreme to mild incidents, typically encountered on a given shift, up-to-the-decision-moment, as a variable mix of stress drivers).
- time of the day (Morning, Afternoon, Evening, Night).
- experience level and age of the driver.
- motion dynamics of the motorcycle driver (indirectly related to the motorcycle), as an indicator of the user-motorcycle neuro-motor interaction load.

In MoveON, the operational context includes, notably, the pattern of traffic conditions experienced within the particular officer's shift envelope up-to-the-moment, and critically, the present traffic conditions, and, the responsive coping capacity of the rider, their attendant stress/cognitive-headroom cost and thus, their remaining currently disposable cognitive headroom. This is because context-sensitivity in stress assessment implies consideration of the stress inducers in the task flow context. Therefore, as context-specifically identified stress inducers are the most relevant determinants of the remaining cognitive headroom, for traffic police as in the MoveON application domain, the most important stress inducers arise from the operational environment parameters; chiefly the traffic conditions. However, the cumulative amount of residual stress incurred by the driver and thus, the respective depletion of the available cognitive headroom has a highly personal pattern; itself dependent on a number of correlations of stress whose cumulative influence on the individual would be modulated by the individual's personal coping capability, which is their *stress vulnerability profile*. The correlations of stress include relatively static as well as other relatively dynamic personal and operational profile patterns. These mainly involve the psycho-physiological correlations of stress such as life-course, coping-style, cognitive-style, work/life-style as well as the operationally significant influences such as the historical and current Shift Envelope (the range of variable operational stress inducers experienced as a mix of incidents that can typically occur within a shift as time elapses. For example, Saturday night to Sunday morning as a shift having its stressors considered cumulatively up to the present instantaneous context of task challenges having to be coped with instantaneously by the rider at their current position (both geographical and task flow-specifically).

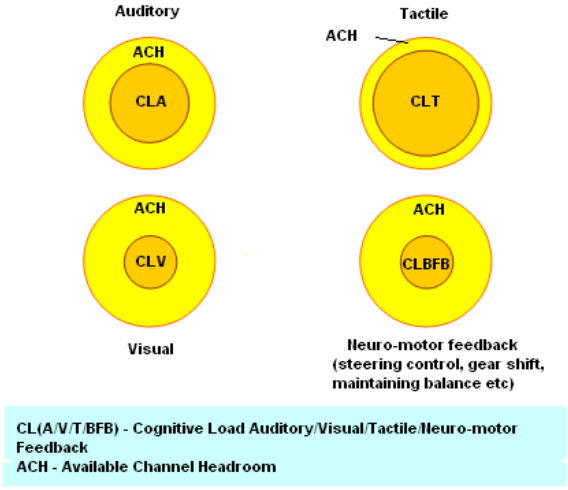


Figure 3. Attention and available cognitive headroom

B. Cognitive Load vs. Available Headroom

The visual, auditory, tactile and neuro-motor channels can be deployed, as illustrated in Figure 3. The deployed cognitive space is denoted by CL* where CL is Cognitive Load and * denotes the channel {Auditory, Visual, Tactile, Neuro-motor (bio)feedback} while the complementary area is denoted by ACH i.e., Available Channel Headroom. The available cognitive headroom perceived for a channel by an individual can be variably influenced not only by the actual cognitive load on that channel but also by cross-channel loading effects that exist in certain contexts of co-occurring neuro-motor and perceptual tasks to be discharged by the brain and such cross-channel loading effects have an irreducible person-invariant baseline plus a person-specific variation depending on a person's gender/genetic context and the task context. Essentially there are genetically variable influences as well as universally applicable influences and environmentally/ emotionally amplifying overlay influences that are associated with the extent of perceived sensory (cross)-channel loading in humans.

Within our minimal distraction C^3 paradigm, we model the above-named four sensory modality channels under the assumption of bounded cognitive headroom whereby the Available Cognitive Headroom (ACH) ranges from 0% to 100%. Incoming events from different input channels during the runtime of MoveON, affect the ACH differently for each modality in respect of a given context (e.g., the person-specific and all other relevant profiles etc as discussed previously).

As explained above, for calculating the variability of ACH per channel, we also have to take into account all possible cross-channel effects, i.e., increased levels of cognitive load when two incoming events require the simultaneous use of the same output channel or co-occurring channels and neuro-motor/bio-feedback activity as normally occurs in riding/driving; for example particularly when taking corners as a motor-cycle rider has to do in wet weather and/or poor visibility conditions when this requires more exacting orchestration of bio-feedback control mediated through coordinative structures output by the brain based on the sensory-perceptual-cognitive input.

C. Traffic Analytics

As for the rider's geographical position, the respective GPS sub-system provides information regarding the location of the officer at all times in terms of Latitude, Longitude, speed and direction with time-stamps. This information is used by the traffic analytics module in conjunction with the traffic congestion information relevant to the officer's current location as well as their stress-response profile. Such analytics are carried out to provide the requisite salient pattern discovery to help inform of the rider's currently disposable cognitive headroom level based on the de facto cumulative cognitive-load/stress-effects resulting from the interplay of a mix of idiosyncratic and deterministic influences. The analytics serve to precompile these into principal indicators of disposable cognitive headroom in so far as this is modulated by the traffic context as the main stress inducer. Thus, the rider's response to such stress inducers is influenced by the precompiled idiosyncratic personally-specific underlying factors that are thus taken into account by the analytics to inform the context-aware minimal distraction information interaction process.

This constitutes a knowledge basis that underpins the data intelligence pattern discovery for our multimodal communication dynamic optimisation architecture, including analytics, to enable the necessary predictive modelling process that can support the modality control decision engine in the context of the current situation assessment. This is to help the MoveON modality controller ultimately to derive, based on the situation assessment, the values of the currently ACH for each of the mutually cross-loading channels (auditory, visual, tactile) and hence, to decide on the channel to be the next selected in attempting to safely grab some of the user's attention within a target time window (e.g., now). This amounts effectively to deciding on the highest priority data to be transmitted, the transmission modality (the specific user's sensory channel to be the next selected conduit for conveying information), timing, place, surface and space in which to convey the data so as to safely grab some of the attention of the user. This would require the expectation values of the current perceived stress load to be estimated for the particular officer at a particular spatio-temporal point in the shift-envelope to help decide the optimal control of multimodal information input/output to best support the user in the execution of their task.

D. Predictive Learning

A Predictive Learning Model of the road traffic conditions was deployed to inform the modality control system with the congestion profile update regarding a road segment. This constituted one of the inputs to the controller to help deduce the associated stress levels to be estimated in respect of a particular rider/driver. Thus, the Predictive Traffic Learning Model exploits the static information coming from the officers' tacit knowledge as well as the historical and current dynamically evolving traffic information, e.g., from Rich Site Summary (RSS) feeds. Long-term patterns reflect the average traffic conditions within a certain area over a long time period and thus, the congestion expectation levels for a given spatio-temporal envelope. Short-term patterns contain the current traffic conditions, which are expected to point to a more accurate assessment of congestion expectation at a given time and place on the road. In cases, where short-term patterns have a traffic record for the requested position, this will take precedence over long-term patterns.

E. Biological Data

Instantaneously monitored psycho-physiological parametrics (voice pitch, heart rate, blood pressure, breathing rate, perspiration) compared to normal baselines are used as supplementary metrics, to enable the effective measurement of the overall instantaneous stress level as being proportional to the safely available disposable cognitive load or attention.

A stress vulnerability model component exists as part of the Cognitive Capacity Management in our architecture that handles measurement and evaluation of a user's biological

 ISBN: 978-1-61208-164-9

context. After evaluation of general measurable biological data and consultation with medical personnel, a pool of biological data can be selected to adequately serve as supplementary optional knowledge to be captured to support the MoveON application domain. This includes heart rate, respiration rate, skin temperature, and blood pressure. An evaluation process has shown that each of the above parameters represents a well-established indicator for higher body activity and that all are as such directly related to psycho-physiological stress, the concomitant stress and cognition load and thus, the ACH. Heart rate, respiration rate, and blood pressure are for example likely to rise under a heavier physical load and psychological stress that manifests itself in the individual physical condition.

The evaluation process had to account for additional hardware requirements in terms of garments that can measure the identified biological data. Such garments must not disrupt or degrade a task force member's ability to perform his tasks, but must seamlessly integrate with his task performance style in strict accordance with the overall purpose of this system, which is to support officers on duty. While the vast number of suitable garments can measure heart and respiration rate, those garments able to measure skin temperature and blood pressure had to be rejected for reasons of impracticality in the context of an officer's daily routine. For this reason, skin temperature and blood pressure were discarded from the data acquisition agenda. Heart rate and respiration rate were used to estimate the measure of stress vulnerability. The fact that heart rate and respiration rate baseline patterns are highly person-specific was taken into consideration by using moving averages on available individual biological data to derive a reasonable rest-state heart rate (i.e., a personal cardiovascular baseline as an operational reference datum) and by using fuzzy parameters estimated stress factors in a given situation.

IV. EVALUATION

A prototypical user-group of UK police motorcyclists comprising a number of officers of a range of ages and ranks participated in our user-centred system co-design and evaluation process to assess the user's distraction whilst interacting with the prototype system, and identifying driving conditions in which the motorcyclist's attention is disrupted thus, potentially exposing the driver to safety risks. Additional feedback from an expert user group including Human-Computer Interaction (HCI) experts as lifelong avid motorcyclists themselves shaped the research, design and development of the minimal distraction real-time information exchange system. Biometric data on physiological factors and visual recordings collected by the officers, on-the-road in real scenarios, was used to help the quantitative and qualitative evaluation to inform our evolutionary interactive design and development of the system. Questionnaires and in-depth interviews were conducted with the user group regarding the usage of the system *i)* in a laboratory, *ii)* while on a parked motorcycle, and *iii)* whilst riding the motorcycle "in-pursuit mode" on the road. Although, ideally, further extensive tests would be needed to assess the scalability of the system fully; it is possible to confidently expect that this architecture is scalable within a range accommodating the aforementioned six variables in the context high level classification of ambient condition variations to the extent that the size of the controller optimisation search space remains below 300 nodes (decisional state space) consistent with real-time performance and controller in-line responsive mode. For offline learning mode, there will be no such limitation to scalability as long as the context caching parametric space is adjusted to include any additional variables of interest. However such scalability boundary condition as might be considered are in any case significantly mitigated by this architecture through *a)* a more conservative incremental approach to disposable cognitive load recruitment, *b)* risk-minimising instantaneous and graceful handover of control to the rider through rider simply signalling an "over-ride" command, *c)* hardware acceleration, *d)* context-caching all decision instances for offline training, thus, enhancing the learning case base of the system so as to continuously evolve a more efficient and reliable and safe multimodal C^3 assistance capability for all mobile task forces and their field commanders and the Central Taskforce Command.

V. CONCLUSION

This paper has set out the relevant state-of-the-art and thus motivated, as well as described, the architecture of the system arising from our user-centred design and development of a real-time multimodal communications controller interface system for (hostile) high cognitive-load environments as exemplified by our police motorcyclist test-case. The resulting system has been evaluated and shown to support the key objective of maximising real-time and time-critical mission information exchange between each taskforce member and other command and control units whilst minimising driver distraction that may become unsafe. The overall architecture of the MoveON Jacket prototype has been described. This has been shown, under real-road traffic conditions, to enable motorcyclists to carry out various technologically assisted functions. A detailed description of the minimal distraction design concepts has been presented with reference to cognitive-load and perceived stress levels considering physiological factors, manoeuvre-context of the vehicle, and the interaction between the user and the system on the road. This architecture as powerfully supported by the Cognitive Capacity Management paradigm represents a pioneering dynamic taskforce stress load management architecture that has delivered the first validated implementation of a mobile taskforce multimodal C^3 optimisation architecture Ambient Assistive Partner capable of graceful in-line man-in-the-loop cognitive control supported by man-machine mixed-initiative taking and messaging prioritisation focused on timing critical mission situation assessment updates in full duplex.

ACKNOWLEDGEMENT

The research and development as reported in this paper was undertaken as part of the MoveON project (EC-funded IST FP6). The authors would like to acknowledge ex-colleagues Patrick Seidler and Chaoxin Wu for their

 ISBN: 978-1-61208-164-9

contribution to the final stages of module coding and integration of the system, and, the MoveON Consortium Partners, in particular A&E Solutions and officers from the West Midlands Police, UK as users, for helping with system evaluation and/or acting as our user group to support our test-cases for real-life performance evaluation of the system.

REFERENCES

[1] Niculescu, A.I., van Dijk, E.M.A.G., Cao, Y., and Nijholt, A. "Measuring stress and cognitive load effects on the perceived quality of a multimodal dialogue system". Proc. 7th Int Conference on Methods and Techniques in Behavioral Research: Measuring Behavior 2010, 24-27 August 2010, Eindhoven. pp. 453-455. ISBN 978-90-74821-86-5. Noldus Information Technology. 2010.

[2] Kun, A. L., Miller, W.T., and Lenharth, W.H. "Evaluating the user interfaces of an integrated system of in-car electronic devices". Proc. IEEE Intelligent Transportation Systems Conference, Vienna, Austria, September 13-16. pp. 953-958. 2005.

[3] MoveOn. [online]. Available on the WWW: <http://showcase.m0ve0n.net/> [Accessed 14 October 2011]. 2009.

[4] Elder, G.H., Pavalko, E.K., and Clipp, E.C. "Working with archival data: studying lives". Sage University Paper Series on Quantitative Applications in the Social Sciences, No. 07-088. Newbury Park, CA: Sage Publications 1993.

[5] Schmorrow, D.D., and Reeves, L.M. (Eds.): Augmented Cognition, HCII 2007, LNAI 4565, pp. 147–156, 2007.

[6] Brünken, R., Plass, J.L., and Leutner, D. "Direct measurement of cognitive load in multimedia learning". Educational Psychologist, 38, pp. 53-61. 2003.

[7] Haapalainen, E., Kim, S., Forlizzi, J.F., and Dey, A.K. "Psycho-physiological measures for assessing cognitive load". Proc. 12th ACM Int conference on Ubiquitous computing (Ubicomp '10). ACM, New York, NY, USA, pp. 301-310. 2010.

[8] Badii A, "User-intimate requirements hierarchy resolution framework (UI-REF): methodology for capturing ambient assisted living needs", Proc. Int. Ambient Intelligence Systems Conference (AmI'08), Nuremberg, Germany. pp. 91. 2008.

[9] Antoine, R. and Eskenazi, M. "A Multi-Layer Architecture for Semi-Synchronous Event-Driven Dialogue Management", IEEE Automatic Speech Recognition and Understanding Workshop, ASRU2007. Kyoto, Japan. pp. 514-519. 2007.

[10] Cooper, G. "Research into cognitive load theory and instructional design at UNSW". [online] School of Education Studies, University of New South Wales, Australia. Available on WWW: < http://dwb.unl.edu/Diss/Cooper/UNSW.htm> [Accessed 14 October 2011]. 1998.

[11] Sweller, J. "Cognitive load during problem solving: effects on learning". Cognitive Science, 12, pp. 257-285. 1988.

[12] Sweller, J. "Cognitive load theory, learning difficulty and instructional design". Learning and Instruction, 4, pp. 295-312. 1994.

[13] Paas, F.G.W.C., and van Merriënboer, J.J.G. "The efficiency of instructional conditions: An approach to combine mental-effort and performance measures". Human Factors 35 (4): pp. 737–743. 1993.

 ISBN: 978-1-61208-164-9

Using Vision-Based Driver Assistance to Augment Vehicular Ad-Hoc Network Communication

Kyle Charbonneau, Michael Bauer and Steven Beauchemin
Department of Computer Science
University of Western Ontario
London, Ontario, Canada
{*kcharbo,bauer,beau*}*@csd.uwo.ca*

***Abstract*—Using Vehicular Ad-Hoc Network (VANET) communication for a Cooperative Collision Warning System (CCWS) has been explored and has shown promise in improving vehicle safety. However, the performance of such a system under different adoption rates has not been examined in depth. We first examine what effects varying adoption rates will have on a CCWS protocol with a variable broadcast scheme. We then examine the implementation of a VANET alongside a Vision-Based Driver Assistance (VBDA) system that monitors the environment surrounding the vehicle using cameras. We propose an Enhanced CCWS protocol where information from VBDA is included with CCWS related VANET communication to significantly increase its effectiveness under low adoption rates.**

Vehicular-Ad Hoc Network, Vision-Based Driver Assistance, Cooperative Collision Warning System

I. Introduction

In a Vehicular Ad-Hoc Network (VANET)-based Cooperative Collision Warning System (CCWS) each vehicle periodically shares information about itself, primarily its current location and trajectory, with surrounding vehicles. Through these location updates each vehicle can build a model of neighbouring vehicles in the surrounding environment. The concept of a CCWS has been introduced, studied and validated by a number of researchers [1] [2] [3]. However, these studies typically look at the operation of a CCWS with a 100% adoption rate, or in other words 100% of vehicles are equipped for VANET communication. Unfortunately upon adoption of the Wireless Access in Vehicular Environments (WAVE) set of standards in production vehicles there will be a long gap between the initial introduction and nearing 100% adoption.

In addition to VANET communication a Collision Warning System (CWS) utilizing on-board vehicle sensors is another technology of interest for improving safety. We see it in the form of Adaptive Cruise Control (ACC) and Forward Collision Warning Systems (FCWS), for example the Active Cruise Control system found on BMW vehicles. These types of sensors have been extended for use in autonomous vehicles in the DARPA challenges and for Advanced Driver Assistance Systems (ADAS) in the RoadLab project at the University of Western Ontario [4]. In this paper, we describe the use of a Vision-Based Driver Assistance (VBDA) system, which uses cameras and computer vision algorithms, designed as part of the RoadLab project.

In a vehicle capable of both VANET communication and VBDA the information attained from each technology can be merged into a unified model to increase accuracy. This can be taken one step further and information gained from VBDA can be used to enhance VANET communication. In this paper we assume vehicles are either equipped with both VANET and VBDA technologies or have neither. An Enhanced CCWS (ECCWS) protocol is proposed where equipped vehicles append information attained about unequipped vehicles to CCWS location updates. This allows a more complete model of the environment to be built even under low adoption rates.

This paper describes research that has explored the potential for such a system. In a simulation environment both VANET and VBDA technologies are tested alongside one another. Under varying adoption rates between 10% and 100% a CCWS and an ECCWS protocol are tested. The effect of varying adoption rates and the potential for improvement with an ECCWS protocol are examined. To study this we have created a robust simulation environment for realistic vehicular traffic, wireless network communication and computer vision. Multiple open source projects are combined with custom modules to achieve this.

The remainder of the paper is structured as follows. In Section II, we present related work that this research was based on. In Section III, the unified model built from both sources is explained. Following in Section IV, we explain the specifics of the ECCWS protocol. In Section V, we explain the simulation environment. Then, in Section VI, we examine the results of our simulations. Finally, in Section VII, are concluding remarks and future directions for this research.

II. Related Work

The feasibility of a CCWS is analyzed by H. Tan and J. Huang where they examine the technologies necessary to implement such a system effectively [1]. They find implementing a CCWS based on current technologies is

 ISBN: 978-1-61208-164-9

feasible and proceed to test out a theoretical CCWS system using two vehicles to produce promising real world results.

Expanding on this work, the frequency of location update broadcasts is examined further by S. Rezaei et al. [2]. In their paper they examine a number of different broadcast schemes under simulation. One such broadcast scheme is periodic communication intervals where location updates are generated on a set interval, from 25ms to 500ms. A second broadcast scheme is variable communication intervals where an error threshold between actual vehicle location and the estimated vehicle location, based on the last location update, must be exceeded before a new location update is broadcast. The paper also introduces a model for Differential Global Position System (DGPS) error which we use in our simulations.

The best broadcast scheme is found to be a variable communication interval with repetition within 50ms. A similar broadcast scheme is again selected by C. Huang et al. for further testing [3]. Our simulations confirm that this is an excellent broadcast scheme for CCWS communication and as such our CCWS protocol is based on it.

The VBDA system is based on the RoadLab project[4]. Vehicles are instrumented with 10 cameras arranged in stereo pairs monitoring the world surrounding the vehicle. To improve vision performance looking forwards there are two pairs of cameras monitoring that direction. The layout of the cameras and range they provide useful information for is shown in Figure 1. The images provided by these cameras are analyzed in real time to identify vehicles and objects surrounding the instrumented vehicle at a rate of 30Hz or higher. The results produced for each object identified include a distance to the object and 2D bounding box drawn over the object in 3D space.

Information from both VANET communication and VBDA is integrated into a unified model [5]. While RoadLab relies on VBDA the results from a RADAR or LIDAR based driver assistance system could be used instead as all three are fundamentally based on line of sight.

Finally, our simulation environment is based on work done by C. Sommer et al. in linking the discrete event simulator OMNeT++ and traffic simulator Simulation of Urban Mobility (SUMO) [6]. Both simulators are linked together for realistic wireless network and node mobility simulation.

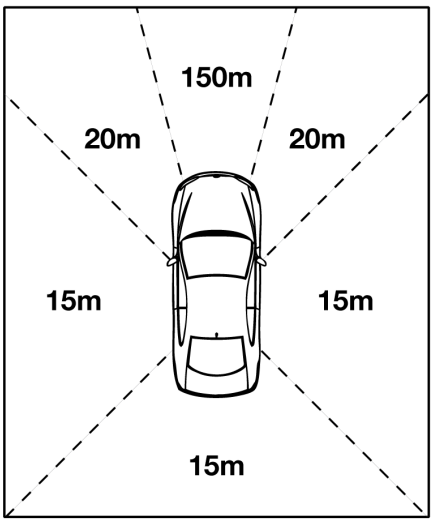


Figure 1. Range and layout of RoadLab cameras

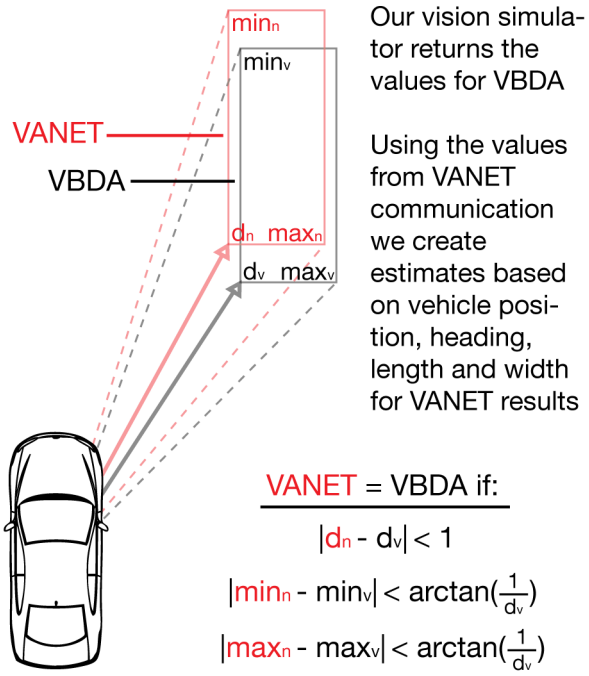


Figure 2. Unified model demonstrated

III. Unified Model

In order to use information from both VANET communication and VBDA, we first create a unified model. We will often have location estimates for neighbouring vehicles from both sources with varying amounts of error. These position estimates should be linked when they are both in reference to the same vehicle.

Through the CCWS, location estimates consist of vehicle position, heading and size. This provides us with a good estimate of the space occupied by the vehicle however the actual vehicle location may be inaccurate due to communication errors. Through VBDA we only receive a distance to the vehicle along with a 2D bounding box. This does not provide us with the physical space occupied by the vehicle however the information we attain should be very accurate

 ISBN: 978-1-61208-164-9

due to the frequency of VBDA updates. In our simulations we assume no measurable error in VBDA results.

In order to combine the two estimates we take the information attained from our CCWS and draw a 2D bounding box where we believe the vehicle is and calculate the distance to the vehicle. We compare this result for each CCWS estimate with each of our VBDA results in order to determine which VANET estimates and VBDA results are closely matched. The simulations we will describe in Section V are performed in 2D so instead of calculating a 2D bounding box in 3D space we calculate the minimum and maximum angle the object would occupy in the cameras' field of view as described in Figure 2.

For each potential match of a VBDA result and VANET estimate we compare the distance to both. A maximum difference in distance of 1m is allowed for a match to be made. If less than 1m of difference in distance exists we compare both angles. A difference in angle equivalent to a maximum of 1m at the distance the vehicle is away from us or $\arctan(\frac{1}{d})$ is allowed. This is to ensure the difference in angle scales along with distance to the vehicle.

If more than one vehicle exists that matches these parameters the one with smallest combined difference in distance and angles is chosen as a match. We record all matches along with the number of errors made. We also record any vehicles that go unmatched but could potentially be matched.

IV. ECCWS

Our regular CCWS protocol involves broadcasting the vehicle location, trajectory, current time and information on the vehicle, such as vehicle dimensions, with a variable communication interval. The variable communication interval works as follows. When a location update packet is broadcast the vehicle position and trajectory are saved. Every 10ms the difference between the current vehicle position, from DGPS, and the estimated vehicle location, based on the last vehicle position update, are compared. If the error between the actual vehicle location and estimated vehicle location is greater than our error threshold of 0.5m then a new location update packet is broadcast. In addition we repeat each broadcast a second time within 50ms. These location updates allow other vehicles to estimate the space our vehicle is currently occupying along with where it will be in the next few seconds. The CCWS protocol only broadcasts information for the vehicle itself.

Our ECCWS protocol appends information to each location update for unequipped nearby vehicles. Since a large portion of our CCWS packets are made up of physical, MAC and network layer headers along with message security features, adding information on nearby vehicles will be more efficient than broadcasting additional packets.

If we have identified a vehicle using VBDA and have not received any VANET communication from it, based on our unified model identifying the vehicle, we will mark the vehicle as being unequipped. We will append the information of up to the four closest unequipped vehicles to our own location updates. By doing this we share information that vehicles outside of visual range could not possibly receive with VBDA alone and give them a more complete picture of the environment.

The size of our CCWS application layers packets is 242 bytes including a 54 byte signature and 128 byte certificate [7]. For each appended vehicle the packet size is increased by 40 bytes to include all relevant information. If extra information for four other vehicles is included, the application layer packet increases in size by only 66% providing us with an efficient way to increase the amount of information shared between vehicles. For our ECCWS protocol location updates are broadcast at the same time as the regular CCWS protocol, only the extra information is appended and the packet size is increased accordingly.

V. SIMULATION

Using our simulation environment we test both the CCWS and ECCWS protocols. Network simulation is done using OMNeT++ using the MiXiM framework. Our CCWS application layer is implemented as a custom module. The WAVE Short Message Protocol (WSMP) is implemented for the network layer. An existing 802.11b MAC layer is adapted with appropriate timing parameters for 802.11p. Finally, a Packet Error Rate (PER) model developed by S. Cocorada for Orthogonal Frequency Division Multiplexing (OFDM) broadcasts is used to decide if incoming packets are accepted or rejected [8].

We transmit our messages with a bitrate of 6Mbps and transmission power of 35.4dBm on IEEE 802.11p channel 178 or the Control Channel (CCH). We model path loss with a path loss coefficient of 3.0 and shadowing with a mean signal attenuation of 0dB and standard deviation of 4dB [9].

The Vehicles in Network Simulation (VEINS) project is used to link OMNeT++ with SUMO. This controls node movement inside a provided road network. We test our CCWS and ECCWS protocols on three different road networks, a Manhattan grid type network with roads running in a grid pattern, a city network based on downtown London, Canada and a highway network based on Highway 401, Canada.

Finally, our vision simulation is implemented as a custom module in OMNeT++. Each vehicle is modeled as a 2D rectangle. Every 100ms we update our vision algorithm and for each vehicle create a list of visible neighbouring vehicles. We determine if a vehicle is visible by calculating what percentage of it is occluded. If less than 50% of the vehicle is occluded it is determined to be visible. It is assumed that we cannot see through any vehicles and anything behind them is occluded.

The simulations are each 120 seconds in length and statistics are recorded throughout the entire simulation runtime.

 ISBN: 978-1-61208-164-9

The average number of vehicles in the Manhattan grid, London and highway road networks is approximately 640, 720 and 1100 vehicles respectively. The adoption rates of 10%, 25%, 50%, 75%, 90% and 100% are tested on each road network once with our regular CCWS protocol and once with our ECCWS protocol. The results are recorded and analyzed afterwards.

VI. Results

We execute our simulation once for all three road networks, under six different adoption levels with both CCWS schemes for a total of 36 executions. Every 100ms during the simulation we record statistics for each vehicle on VBDA, VANET communication and the unified model. Additionally, for each CCWS position estimate we record the error between the estimated position and the actual vehicle position. Finally, we record statistics on packets sent, received and error rates. For each statistic collected the mean and standard deviations are calculated both on a per vehicle and overall basis.

Our unified model, despite being quite simplistic, performs well. There are two types of unified model errors. First, matches-missed, which is a vehicle tracked by both CCWS and VBDA but incorrectly assumed to be two separate vehicles. Second, match-errors, which are matches made between two separate vehicles, one tracked by CCWS and and one tracked by VBDA, that are incorrectly assumed to be the same vehicle. In general both matches-missed and match-errors are below 0.5% of all possible matches or all matches made respectively. In the highway road network, matches-missed is slightly higher at approximately 1%. With the higher speeds present on a highway compared to city driving there is the potential for a larger error between actual vehicle location and estimate vehicle location. This would explain the higher matches-missed on the highway network. Using our unified model we implement a ECCWS.

From Figure 3, we can see the number of vehicles tracked by the CCWS and VBDA in our unified model increases in a linear fashion as the adoption rate rises. This is expected since the number of vehicles within communication range will increase linearly. The number of vehicles tracked by the ECCWS and VBDA is very promising though. This initially increases quite rapidly until we reach approximately 50% adoption. At this point the number of vehicles tracked is approximately the same as the number tracked at 100% adoption. The result levels off and is stable from 50% to 100%.

This result shows that by 50% adoption our ECCWS protocol can track essentially all vehicles that the CCWS protocol would be able to at 100% adoption. Additionally, by 25% adoption, the ECCWS protocol can track the same number of vehicles as the CCWS protocol at 75% adoption. This presents a strong case for an ECCWS in extending the reach of VANET communication during its initial stages.

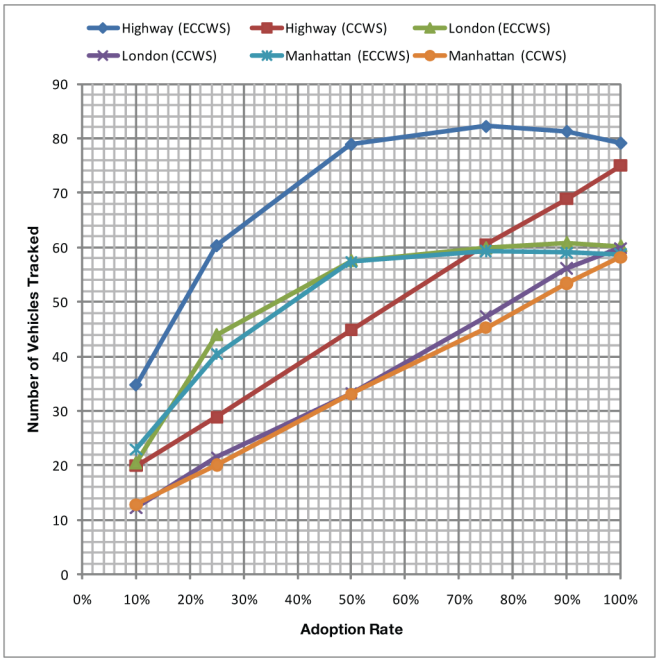


Figure 3. Number of vehicles tracked at various adoption rates

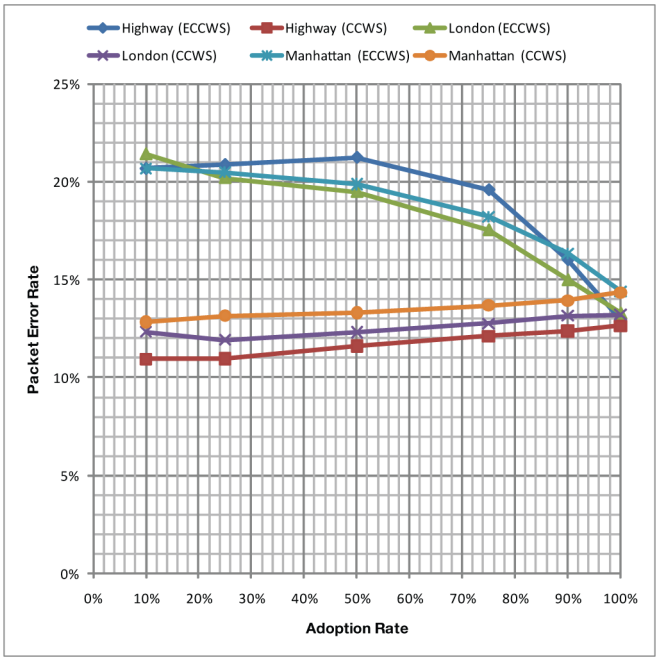


Figure 4. PER at various adoption rates

Furthermore, we can look at the PER for these simulations under different conditions in Figure 4. As expected the PER for the CCWS increases slightly as adoption rate increases. This is the result of increased number of transmissions and related interference causing lost packets. The packet error

 ISBN: 978-1-61208-164-9

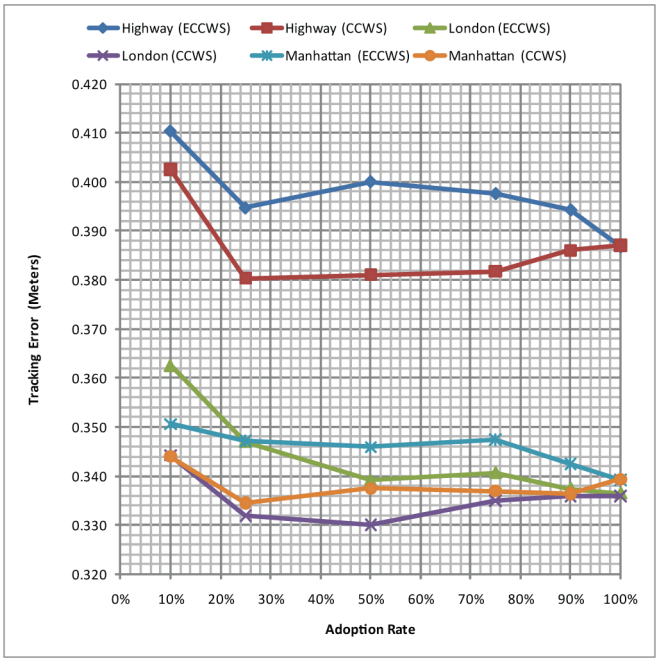


Figure 5. VANET CCWS tracking error at various adoption rates

rate for the ECCWS is higher for low adoption rates and levels off under 100% adoption at approximately equal to the CCWS protocol. Since under lower adoption rates we have larger packets, leading to more bit errors and therefore more packet errors, this result is to be expected. The PER for both protocols at 100% adoption converges as there are few vehicles visible but not tracked by our CCWSS. Therefore no additional information is appended to our location updates and the packet size remains unchanged between CCWS and ECCWS protocols.

However, PER does not necessarily give us an indication of CCWS performance. We calculate the mean tracking error for each system for each simulation in Figure 5. Under low adoption rates, the ECCWS does have a slightly higher mean tracking error than the CCWS, however it only 2-3cm higher at most. This difference is smaller than the mean DGPS error [2] and well under the 0.5m accuracy requirement for accurate position of a vehicle within a lane [1]. Despite the increase in PER the ECCWS performs well under all adoption rates tested.

VII. Conclusion and Future Works

Overall, these results show that an ECCWS protocol with additional information from VBDA shows great potential for improving system performance under low adoption rates. Of course, VBDA or any similar sensor based driver assistance system also presents additional benefits in terms of accuracy, latency and security under all adoption rates. As such, the implementation of both VBDA and VANET communication together and the use of an ECCWS protocol shows great potential.

In future work we plan to improve the simulation environment by extending it into 3D space and adding obstruction information for vision and radio shadowing. Furthermore, by examining how to use our unified model to improve vehicle safety and what information is necessary we can better quantify the benefits of an ECCWS.

References

[1] H. Tan and J. Huang, "DGPS-based vehicle-to-vehicle cooperative collision warning: Engineering feasibility viewpoints," *IEEE Transactions on Intelligent Transportation Systems*, vol. 7, no. 4, pp. 415–428, 2006.

[2] S. Rezaei, R. Sengupta, H. Krishnan, X. Guan, and R. Bhatia, "Tracking the position of neighboring vehicles using wireless communications," *Transportation Research Part C*, 2009.

[3] H. Ching-Ling, Y. Fallah, R. Sengupta, and H. Krishnan, "Adaptive intervehicle communication control for cooperative safety systems," *Network, IEEE*, vol. 24, no. 1, pp. 6 –13, 2010.

[4] B. Steven, M. Bauer, D. Laurendeau, T. Kowsari, J. Cho, M. Hunter, and O. McCarthy, "Roadlab: An in-vehicle laboratory for developing cognitive cars," 2010.

[5] M. Bauer, K. Chabonneau, and S. Beauchemin, "V2eye: Enhancement of automated visual perception from v2v communication," 2011.

[6] C. Sommer, R. German, and F. Dressler, "Bidirectionally Coupled Network and Road Traffic Simulation for Improved IVC Analysis," *IEEE Transactions on Mobile Computing*, vol. 10, pp. 3–15, January 2011.

[7] T. Chen, W. Jin, and A. Regan, "Multi-Hop Broadcasting in Vehicular Ad Hoc Networks with Shockwave Traffic," in *IEEE CCNC, 2010. Proceedings*, 2010.

[8] S. Cocorada, "An IEEE 802.11g simulation model with extended debug capabilities," 2008.

[9] K. Wehrle, M. Gnes, and J. Gross, *Modeling and Tools for Network Simulation*. Springer, 1st ed., 2010.

 ISBN: 978-1-61208-164-9